TITLE

Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis

AUTHORS

The COVID-19 Host Genetics Initiative
[*The complete list of authors can be found online at*: COVID19-HGI_authorship_list]

Address for correspondence:
Dr Andrea Ganna, Institute for Molecular Medicine Finland, Tukholmankatu 8, Helsinki
email: andrea.ganna@helsinki.fi

ABSTRACT

The genetic makeup of an individual contributes to susceptibility and response to viral infection. While environmental, clinical and social factors play a role in exposure to SARS-CoV-2 and COVID-19 disease severity, host genetics may also be important. Identifying host-specific genetic factors indicate biological mechanisms of therapeutic relevance and clarify causal relationships of modifiable environmental risk factors for SARS-CoV-2 infection and outcomes. We formed a global network of researchers to investigate the role of human genetics in SARS-COV-2 infection and COVID-19 severity. We describe the results of three genome-wide association meta-analyses comprising 49,562 COVID-19 patients from 46 studies across 19 countries worldwide. We reported 15 genome-wide significant loci that are associated with SARS-CoV-2 infection or severe manifestations of COVID-19. Several of these loci correspond to previously documented associations to lung or autoimmune and inflammatory diseases. They also represent potentially actionable mechanisms in response to infection. We further identified smoking and body mass index as causal risk factors for severe COVID-19. The identification of novel host genetic factors associated with COVID-19, with unprecedented speed, was enabled by prioritization of shared resources and analytical frameworks. This working model of international collaboration a blue-print for future genetic discoveries in the event of pandemics or for any complex human disease.

INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic, caused by infections with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has resulted in enormous health and economic burden worldwide. One of the most remarkable features of SARS-CoV-2 infection is that a large proportion of individuals [1] are asymptomatic while others experience progressive, even life-threatening, viral pneumonia and acute respiratory distress syndrome. While established host factors contribute to disease severity (e.g., increasing age, male gender, and higher body mass index [2]), these risk factors alone do not explain all variability in disease severity observed among individuals.

The contribution of host genetics to susceptibility and severity of infectious disease is well-documented, and encompasses rare inborn errors of immunity [3,4] as well as common genetic variation [5–10]. Characterizing which genetic factors contribute to COVID-19 susceptibility and severity may uncover novel biological insights into disease pathogenesis and identify mechanistic targets for therapeutic development or drug repurposing, as treating the disease remains a highly important goal despite the recent development of vaccines. For example, rare loss-of-function variants in genes involved in type I interferon (*IFN*) response may be involved in severe forms of COVID-19 [11–14]. At the same time, several genome-wide association studies (GWAS) that investigate the contribution of common genetic variation [15–18] to COVID-19 have provided support for the involvement of several genomic loci associated with COVID-19 severity and susceptibility, with the strongest and most robust finding at locus 3p21.31. However, much remains unknown about the genetic basis of susceptibility to SARS-CoV-2 and severity of COVID-19.

The COVID-19 Host Genetics Initiative (COVID-19 HGI) (https://www.covid19hg.org/) [19] is an international, open-science collaboration to share scientific methods and resources with research groups across the world with the goal to robustly map the host genetic determinants of SARS-CoV-2 infection and severity of the resulting COVID-19 disease. We have carefully aligned phenotype definitions and incorporated variable ascertainment strategies to achieve greater statistical confidence in our results. We openly and continuously share updated results to the research community. Here, we report the latest results of meta-analyses of 46 studies from 19 countries (**Fig. 1**) for COVID-19 host genetic effects.

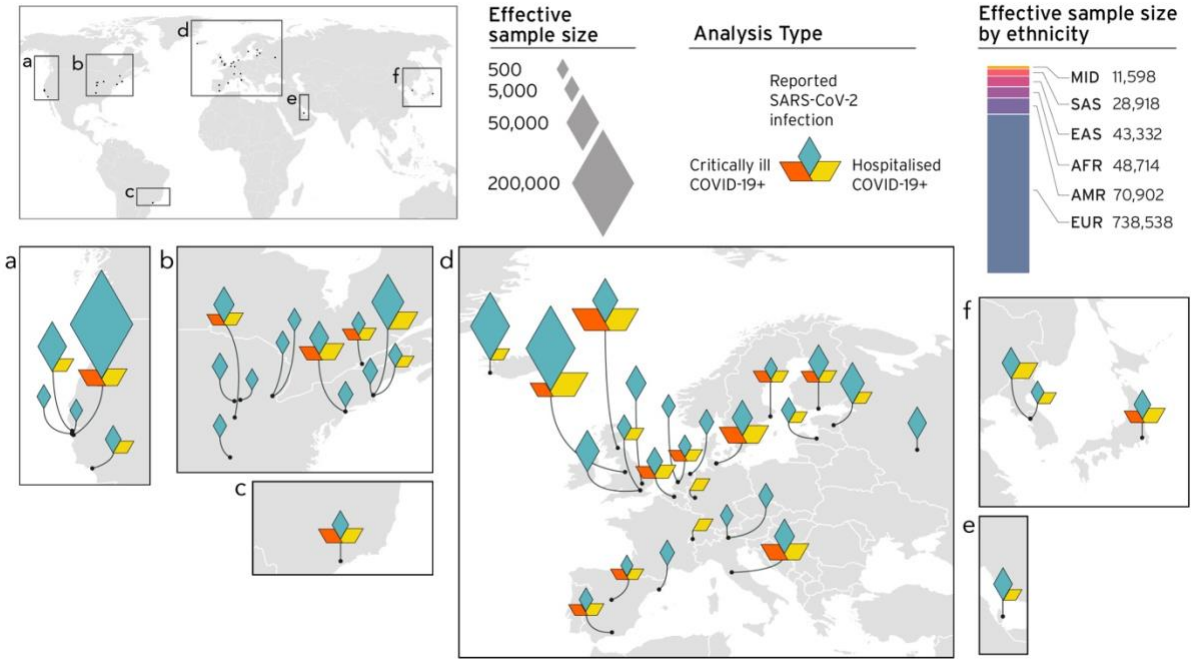


Figure 1. *Geographical overview of the contributing studies to the COVID-19 HGI and composition by major ancestry groups. Middle Eastern (MID), South Asian (SAS), East Asian (EAS), African (AFR), Admixed American (AMR), European (EUR).*

RESULTS

Worldwide meta-analyses of COVID-19

Overall, the COVID-19 Host Genetics Initiative combined genetic data for up to 49,562 cases and two million controls across 46 distinct studies (**Fig. 1**). The data included studies from populations of different genetic ancestries, including European, Admixed American, African, Middle Eastern, South Asian and East Asian individuals (**Supplementary Table 1**). We performed case-control meta-analyses in three main categories of COVID-19 disease according to predefined and partially overlapping phenotypic criteria. These were (1) critically ill COVID-19 cases defined as those who required respiratory support in hospital or who were deceased due to the disease, (2) cases with moderate or severe COVID-19 defined as those hospitalized due to symptoms associated with the infection, and (3) all cases with reported SARS-CoV-2 infection with or without symptoms of any severity (**Methods**). Controls for all three analyses were selected as genetically ancestry-matched samples without known SARS-CoV-2 infection, if that information was available (**Methods**). Each individual study that contributed data to a particular analysis met a minimum threshold of 50 cases, as defined by the aforementioned phenotypic criteria, for statistical robustness. Where more detailed demographic data was available, the average age of COVID-19 cases was 55.3 years (**Supplementary Table 1**).

Across our three analyses, we reported a total of 15 independent genome-wide significant loci associated with COVID-19 (**Table 1, Supplementary Table 2**), most of which were shared between two or more COVID-19 phenotypes. Specifically, we reported six genome-wide significant ($P < 5 \times 10^{-8}$) associations for critical illness due to COVID-19, using data for 6,179 cases and 1,483,780 controls from 16 studies. Eleven genome-wide significant loci were detected for moderate to severe hospitalized COVID-19, from an analysis of 13,641 COVID-19 cases and 2,070,709 controls, across 29 studies (**Fig. 2 top panel**). Finally, eight loci reached genome-wide significance in the analysis using data for all available 49,562 reported cases of SARS-CoV-2 infection and 1,770,206 controls, using data from a total of 44 studies (**Fig. 2 bottom panel**).

Comparison of effect for genome-wide significant results across studies and phenotype definitions

We found no genome-wide significant sex-specific effects at the 15 loci. However, we did identify significant heterogeneous effects (P <0.003) across studies for 3 out of the 15 loci, likely reflecting heterogeneous ascertainment of cases across studies contributing data to these analyses (**Table 1**). Two additional loci reached genome-wide significance but showed extreme heterogeneity across contributing studies (**Fig 2**); these loci were removed from downstream analyses and are not reported among the genome-wide significant results in **Table 1**. There was minor sample overlap (n = 8,380 EUR; n = 745 EAS) between controls from the genOMICC and the UK Biobank studies, but leave-one-out sensitivity analyses did not reveal any bias in the corresponding effect sizes or P-values (**Extended Data Fig. 1**).

We next wanted to better understand whether the 15 loci were acting through mechanisms increasing susceptibility to infection or by affecting the progression of symptoms towards more severe disease. For all

15 loci, we compared the lead variant (strongest association *P*-value) odds ratios (ORs) for the risk-increasing allele across our different COVID-19 phenotype definitions.

We first noted that four loci had consistent ORs between the two larger and better powered analyses; all cases with reported infection and all cases hospitalized due to COVID-19 (**Methods**) (**Table 1, Supplementary Table 2**). Such consistency implied that these four loci were likely associated with overall susceptibility to SARS-CoV-2 infection, but not with the progression to more severe COVID-19 phenotypes. Notably, these susceptibility loci included the previously reported *ABO* locus [15,16,18,20]. The lead variant rs912805253 at this locus reached genome-wide significance in both our reported infection (OR [95%CI] = 0.90 [0.89, 0.92]; $P = 1.4 \times 10^{-39}$) and hospitalized COVID-19 (OR [95%CI] = 0.90 [0.87, 0.93]; $P = 5.4 \times 10^{-10}$) analyses, but the odds of becoming hospitalized were no different from the odds of becoming infected when carrying this allele.

In contrast, 11 out of the 15 loci were associated with increased risk of severe symptoms with significantly larger ORs for hospitalized COVID-19 compared to the mildest phenotype of reported infection ($P < 0.003$ (0.05/15) test for effect size difference) (**Table 1**, **Supplementary Table 2)**. We further compared the ORs for these 11 loci for critical illness due to COVID-19 *vs.* hospitalized due to COVID-19, and found that these loci exhibited a general increase in effect risk for critical illness (**Methods**) (**Extended Data Fig. 2A, Supplementary Table 3)**. These results indicated that these eleven loci were more likely associated with progression of the disease and worse outcome from SARS-CoV-2 infection compared to being associated with susceptibility to SARS-CoV-2 infection. We noted that two loci, tagged by lead variants rs1886814 and rs72711165, were identified primarily from East Asian genetic ancestry samples (n = 1,414 cases hospitalized due to COVID-19) with minor allele frequencies in European populations being < 3%. This highlights the value of including data from diverse populations for genetic discovery. Another locus at 3p21.31, which is the strongest, most replicated signal for COVID-19 severity [15–18,20], showed substantial differences in allele frequency across ancestry groups, probably explained by its recent introgression [21]. We explored the effect of this locus in the Bangladeshi population, which carries the highest frequency for this haplotype in 1000 Genomes. Using data from the East London Genes & Health study [22] for a proxy variant rs34288077 in the locus (r^2=0.99 to our lead variant rs10490770), we found that in British-Bangladeshi individuals, the variant frequency was 34.6% of the hospitalized COVID-19 positive patients (n = 76) compared to 23.8% in non-hospitalized population (n = 22,215) (OR [95%CI] = 2.11 [1.39, 3.21]; $P = 4.7 \times 10^{-4}$).

Our phenotype definitions include population controls without known SARS-CoV-2 infection. This is not an optimal control group because some individuals, if exposed to SARS-CoV-2 could develop a severe form of COVID-19 disease and should be classified as cases. To better understand the effect of such potential misclassification, we conducted a new meta-analysis, including only the studies that compared hospitalized COVID-19 cases with controls with laboratory-confirmed SARS-CoV-2 infection but who had mild symptoms or were asymptomatic (n = 5,773 cases and n = 15,497 controls). We then compared the effect sizes obtained from this analysis with those from the main phenotype definition (hospitalized cases *vs.* controls without known SARS-CoV-2 infection, if that information was available) using only studies that reported results for both analyses (**Methods**). We found that across the 11 loci that had reached genome-wide significance in our main hospitalized COVID-19 analysis, the ORs were not significantly different in the analysis with better refined controls (**Extended Data Fig. 2B, Supplementary Table 3**).

These results indicate that using population controls can be a valid and powerful strategy for host genetic discovery of infectious disease.

rsid	Chr:pos (b38)	Ref allele	Effect allele	Effect allele frequency	COVID-19 phenotype	OR	*P*-value (association)	*P*-value (het)	Suggested phenotypic impact	Genes in LD region (closest gene in bold)	Genes with coding variants	eGenes
rs67579710	1:155203736	G	A	0.098	Critical illness	0.82	8.31E-08	0.890	disease severity	*KRTCAP2, TRIM46, MUC1,* ***THBS3****, MTX1*	*THBS3*	
				0.102	Hospitalized	0.87	3.38E-08	0.504				
				0.105	Reported infection	0.98	6.38E-02	0.449				
rs1381109	2:166061783	G	T	0.376	Critical illness	0.90	2.32E-04	0.358	disease severity	***SCN1A***		
				0.385	Hospitalized	0.91	4.21E-08	0.813				
				0.390	Reported infection	0.98	7.63E-03	0.174				
rs10490770	3:45823240	T	C	0.075	Critical illness	1.89	2.20E-61	0.051	disease severity	***LZTFL1***		
				0.081	Hospitalized	1.65	1.44E-73	2.09E-03				
				0.085	Reported infection	1.16	9.72E-30	7.24E-25				
rs11919389	3:101705614	T	C	0.344	Critical illness	0.98	3.51E-01	0.082	infection susceptibility	*ZBTB11,* ***RPL24****, CEP97, NXPE3*		
				0.348	Hospitalized	0.95	8.13E-04	0.304				
				0.352	Reported infection	0.94	3.46E-15	0.714				
rs10070196	5:13939721	A	C	0.681	Critical illness	1.02	5.39E-01	0.836	infection susceptibility	***DNAH5***		
				0.678	Hospitalized	1.05	4.18E-03	0.642				
				0.680	reported infection	1.05	2.32E-08	0.191				
rs1886814	6:41534945	A	C	0.038	Critical illness	1.32	1.79E-05	0.815	disease severity	***FOXP4***		*FOXP4*
				0.042	Hospitalized	1.26	1.11E-09	0.348				
				0.047	Reported infection	1.11	2.41E-08	0.216				
rs72711165	8:124324323	T	C	0.011	Critical illness	1.23	2.14E-02	0.267	disease severity	***TMEM65***		
				0.013	Hospitalized	1.37	2.13E-09	0.551				
				0.018	Reported infection	1.08	9.43E-03	0.177				
rs912805253	9:133274084	C	T	0.653	Critical illness	0.88	2.16E-06	0.008	infection susceptibility	***ABO***	*ABO*	*ABO*
				0.655	Hospitalized	0.90	5.37E-10	0.007				
				0.651	Reported infection	0.91	1.45E-39	0.014				
rs10774671	12:112919388	G	A	0.652	Critical illness	1.20	4.08E-13	0.923	disease severity	***OAS1****, OAS3, OAS2*	*OAS1,OAS3*	
				0.664	Hospitalized	1.11	6.14E-10	0.550				
				0.669	Reported infection	1.06	1.61E-11	0.200				
rs1819040	17:46142465	T	A	0.197	Critical illness	0.91	5.02E-04	0.076	disease severity	*ARHGAP27, PLEKHM1, LINC02210-CRHR1, CRHR1, SPPL2C, MAPT, STH,* ***KANSL1****, LRRC37A, ARL17B, LRRC37A2, ARL17A, NSF, WNT3*	*ARHGAP27, KANSL1, MAPT, SPPL2C, STH*	*ARL17B, KANSL1, MAPT, WNT3*
				0.186	Hospitalized	0.88	1.83E-10	0.327				
				0.184	Reported infection	0.96	5.05E-06	0.828				
rs77534576	17:49863303	C	T	0.033	Critical illness	1.45	4.37E-09	0.508	disease severity	*KAT7,* ***TAC4***		*DLX3*
				0.033	Hospitalized	1.26	2.26E-07	0.009				
				0.037	Reported infection	1.08	1.08E-04	0.016				
rs2109069	19:4719431	G	A	0.316	Critical illness	1.26	9.68E-22	0.318	disease severity	***DPP9***		
				0.312	Hospitalized	1.15	2.76E-17	0.157				
				0.315	Reported infection	1.05	4.08E-09	6.71E-05				
rs74956615	19:10317045	T	A	0.048	Critical illness	1.43	9.71E-12	0.223	disease severity	*ICAM1, ICAM4, ICAM5, ZGLP1, FDX2,* ***RAVER1****, ICAM3, TYK2*	*TYK2*	
				0.048	Hospitalized	1.27	5.05E-10	1.94E-04				
				0.052	Reported infection	1.06	3.68E-03	0.002				
rs4801778	19:48867352	G	T	0.176	Critical illness	0.91	3.12E-03	0.579	infection susceptibility	***PLEKHA4****, PPP1R15A, TULP2, NUCB1*	*PPP1R15A*	
				0.176	Hospitalized	0.96	2.90E-02	0.908				
				0.180	Reported infection	0.95	1.18E-08	0.901				
rs13050728	21:33242905	T	C	0.662	Critical illness	0.82	1.05E-16	0.642	disease severity	***IFNAR2***		*IFNAR2*
				0.651	Hospitalized	0.86	2.72E-20	0.194				
				0.651	Reported infection	0.97	1.28E-05	0.001				

Table 1: Genome-wide significant results for each worldwide meta-analysis.
Meta-analysis results that reached genome-wide significance are coloured in black; associations not reaching significance threshold of $P < 5 \times 10^{-8}$ are coloured in grey. Effect allele frequency is the sample size weighted frequency across studies included in each meta-analysis. P-value is reported for meta-analysis variant association with trait (P-value association) and heterogeneity (P-value het) between studies included in each meta-analysis. Suggested phenotypic impact of the locus was inferred using a test comparing variant effects across analyses (see ***Methods****). Closest gene: A gene with a minimum distance from each lead variant to gene body. Genes in linkage disequilibrium (LD) region: Genes that overlap with a genomic range that contains any variants in LD ($r^2 > 0.6$) with each lead variant. Genes with coding variants: Genes with a loss-of-function or missense variant in LD with a lead variant ($r^2 > 0.6$). eGenes: Genes with a fine-mapped cis-eQTL variant (PIP > 0.1) in GTEx Lung that is in LD with a lead variant ($r^2 > 0.6$). V2G: Highest gene prioritized by OpenTargetGenetics' V2G score.*

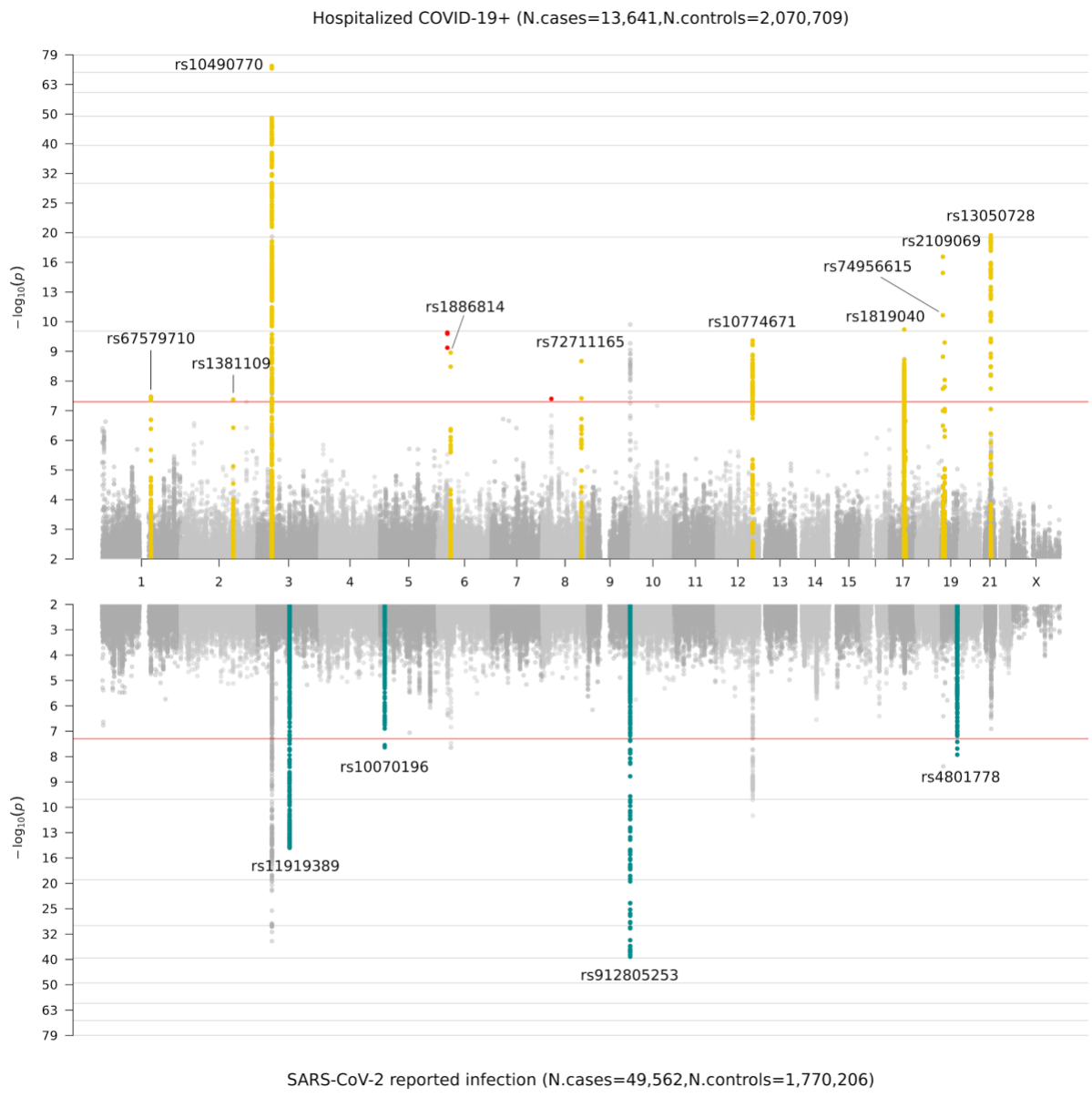


Figure 2. Genome-wide association results for COVID-19. *Top panel shows results of genome-wide association study of hospitalized COVID-19 and controls, and bottom panel the results of reported SARS-CoV-2 infection and controls. Eleven loci highlighted in yellow (top panel) represent regions associated with severity of COVID-19 manifestation i.e. increasing odds for more severe COVID-19 phenotypes, where loci highlighted in green (bottom panel) are regions associated with reported SARS-CoV-2 infection, i.e. the effect is the same across mild and severe COVID-19 phenotypes. A window of +/- 500kb region from the lead variant (**Table 1**) was used to highlight these loci. Variants highlighted in red represent genome-wide significant variants that had high heterogeneity across studies that contributed data, and which were therefore not considered in the final analyses.*

Gene prioritization and association with other diseases or traits

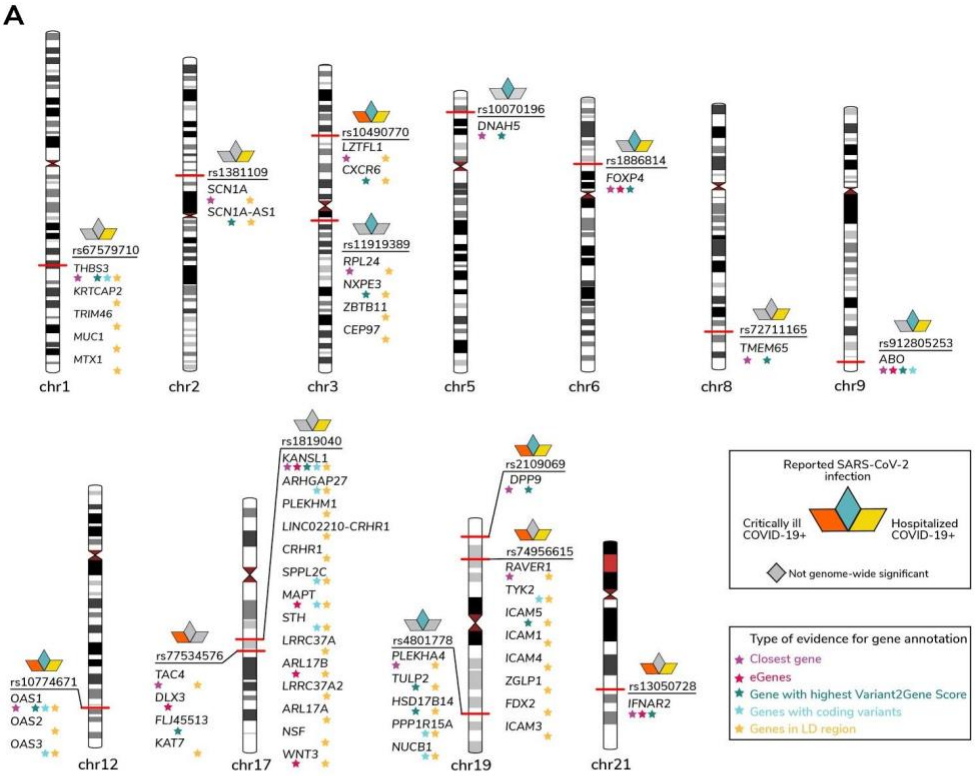
To better understand the potential biological mechanism of each locus, we applied several approaches to prioritize candidate causal genes and explore additional associations with other complex diseases and traits. For gene prioritization, we first identified genes within each COVID-19 associated region by distance or linkage disequilibrium (LD) to a lead variant, and then prioritized those with protein-alternating variants, lung eQTLs, or having the highest prioritization score in the OpenTargets V2G (Variant-to-Gene) algorithm [23] (see **Methods, Supplementary Tables 2 and 4**). For reporting PheWAS associations (**Supplementary Table 5**), we only considered phenotypes for which the lead variants were in high LD ($r^2 > 0.8$) with the 15 genome-wide significant lead variants from our main COVID-19 meta-analysis. This conservative approach allowed spurious signals primarily driven by proximity rather than actual colocalization to be removed (see **Methods**).

Of the 15 genome-wide significant loci, we found nine loci to have a distinct candidate gene(s), including biologically plausible genes (**Table 1**). Protein-altering variants in LD with lead variants implicated genes at six loci, including *TYK2* (19p13.2) and *PPP1R15A* (19q13.33). The COVID-19 lead variant rs74956615:T>A in *TYK2* , which confers risk for critical illness (O[95%] = 1.43 [1.29, 1.59]; $P = 9.71 \times 10^{-12}$) and hospitalization (OR [95%CI] = 1.27 [1.18, 1.36]; $P = 5.05 \times 10^{-10}$) due to COVID-19, is correlated with the missense variant rs34536443:G>C (p.Pro1104Ala; $r^2 = 0.82$) . This is consistent with the primary immunodeficiency described with complete *TYK2* loss of function [24]. In contrast, this missense variant was previously reported to be protective against autoimmune diseases, including rheumatoid arthritis (OR = 0.74; $P = 3.0 \times 10^{-8}$; UKB SAIGE), and hypothyroidism (OR = 0.84; $P = 1.8 \times 10^{-10}$; UK Biobank) (**Fig. 3**). An additional independent missense variant rs2304256:C>A (p.Val362Phe; $r^2 = 0.08$ with rs34536443) in *TYK2* was also associated with critical illness (OR [95%] = 1.17 [1.11, 1.23]; $P = 2.4 \times 10^{-10}$). At the 19q13.33 locus, the lead variant rs4801778, that was significantly associated with reported infection (OR [95%CI] = 0.95 [0.93, 0.96]; $P = 2.1 \times 10^{-8}$), is in LD ($r^2 = 0.93$) with a missense variant rs11541192:G>A (p.Gly312Ser) in *PPP1R15A*. In an additional lookup, this missense variant was significantly associated with reticulocyte count and strongly correlated with a reticulocyte lead variant rs56104184:C>T (beta= 0.033; $P = 4.1 \times 10^{-13}$)[25].

Lung-specific *cis*-eQTL from GTEx v8 [26] ($n = 515$) and the Lung eQTL Consortium [27] ($n = 1,103$) provided further support for a subset of loci, including *FOXP4* (6p21.1) and *ABO* (9q34.2), *OAS1*/*OAS3*/*OAS2* (12q24.13), and *IFNAR2/IL10RB* (21q22.11), where the COVID-19 associated variants modifies gene expression in lung. Furthermore, our PheWAS analysis implicated three additional loci related to lung function, with modest lung eQTL evidence, i.e. the lead variant was not fine-mapped but significantly associated. An intronic variant rs2109069:G>A in *DPP9* (19p13.3), positively associated with critical illness, was previously reported to be risk-increasing for interstitial lung disease (tag lead variant rs12610495:A>G [p.Leu8Pro], OR = 1.29, $P = 2.0 \times 10^{-12}$) [28]. The COVID-19 lead variant rs1886814:A>C in *FOXP4* locus is modestly LD-linked ($r^2 = 0.64$) with a lead variant of lung adenocarcinoma (tag variant=rs7741164; OR=1.2, $P=6.0 \times 10^{-13}$) [29]. We also found that intronic variants rs67579710:A>T in *THBS3* (1q22) and rs1819040:T>A in *KANSL1* (17q21.31), associated protectively against hospitalization due to COVID-19, were previously reported for reduced lung function (e.g. tag lead variant rs141942982:G>T, beta = -3.6×10^{-2}, $P = 1.00 \times 10^{-20}$) [30]. Notably, the 17q21.31 locus is a well-known locus for structural variants containing a megabase inversion polymorphism (H1 and inverted H2 forms)

and complex copy-number variations, where the inverted H2 forms were shown to be positively selected in Europeans [31,32].

Lastly, there are remaining six loci with varying evidence for candidate causal genes. For example, the 3p21.31 locus has a complex structure with varying genes prioritized by different methods, where we prioritized *CXCR6* with the Variant2Gene (V2G) algorithm [23], while *LZTFL1* is the closest gene. The *CXCR6* plays a role in chemokine signaling [33], and *LZTFL1* has been implicated in lung cancer [34]. Nonetheless, these results provide supporting *in-silico* evidence for candidate causal gene prioritization, while we strongly need further functional characterization. Detailed locus descriptions and LocusZoom plots are provided in **Extended Data Fig. 3**.


A
rs67579710
THBS3
KRTCAP2
TRIM46
MUC1
MTX1
chr1
rs1381109
SCN1A
SCN1A-AS1
chr2
rs10490770
LZTFL1
CXCR6
rs11919389
RPL24
NXPE3
ZBTB11
CEP97
chr3
rs10070196
DNAH5
chr5
rs1886814
FOXP4
chr6
rs72711165
TMEM65
chr8
rs912805253
ABO
chr9
rs10774671
OAS1
OAS2
OAS3
chr12
rs77534576
TAC4
DLX3
FLJ45513
KAT7
rs1819040
KANSL1
ARHGAPP27
PLEKHM1
LINC02210-CRHR1
CRHR1
SPPL2C
MAPT
STH
LRRC37A
ARL17B
LRRC37A2
ARL17A
NSF
WNT3
chr17
rs4801778
PLEKHA4
TULP2
HSD17B14
PPP1R15A
NUCB1
rs2109069
DPP9
rs74956615
RAVER1
TYK2
ICAM5
ICAM1
ICAM4
ZGLP1
FDX2
ICAM3
chr19
rs13050728
IFNAR2
chr21
Reported SARS-CoV-2 infection
Critically ill COVID-19+
Hospitalized COVID-19+
Not genome-wide significant
Type of evidence for gene annotation
Closest gene
eGenes
Gene with highest Variant2Gene Score
Genes with coding variants
Genes in LD region

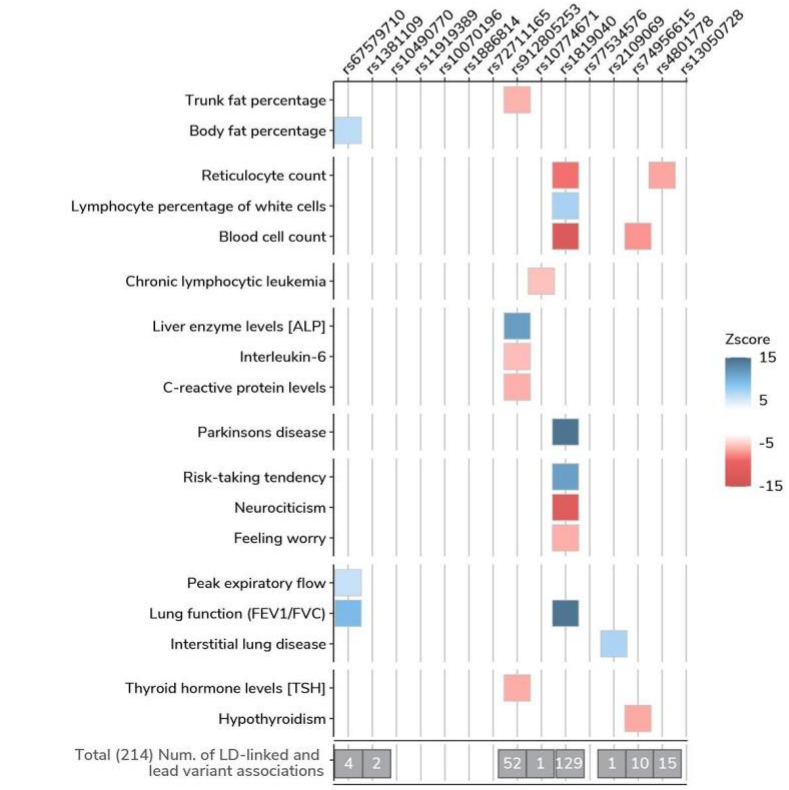

B
rs67579710
rs1381109
rs10490770
rs11919389
rs10070196
rs1886814
rs72711165
rs912805253
rs10774671
rs1819040
rs77534576
rs2109069
rs74956615
rs4801778
rs13050728
Trunk fat percentage
Body fat percentage
Reticulocyte count
Lymphocyte percentage of white cells
Blood cell count
Chronic lymphocytic leukemia
Liver enzyme levels [ALP]
Interleukin-6
C-reactive protein levels
Parkinsons disease
Risk-taking tendency
Neuroticism
Feeling worry
Peak expiratory flow
Lung function (FEV1/FVC)
Interstitial lung disease
Thyroid hormone levels [TSH]
Hypothyroidism
Total (214) Num. of LD-linked and lead variant associations
4
2
52
1
129
1
10
15
Zscore
15
5
-5
-15

Figure 3. **Gene prioritization and PheWas.** ***A)*** *Gene prioritization using different evidence measures of gene annotation. The ideogram shows the position of the COVID-19 index SNPs and associated genes with type of gene annotation highlighted as stars. The squares above the SNP rsid represent COVID-19 sub-phenotypes for which the SNP is genome-wide significant. Closest gene: A gene with a minimum distance from each lead variant to gene body. Genes in linkage disequilibrium (LD) region: Genes that overlap with a genomic range that contains any variants in LD ($r^2 > 0.6$) with each lead variant. Genes with coding variants: Genes with a loss-of-function or missense variant in LD with a lead variant ($r^2 > 0.6$). eGenes: Genes with a fine-mapped cis-eQTL variant (PIP > 0.1) in GTEx Lung that is in LD with a lead variant ($r^2 > 0.6$) (see* ***Supplementary Table 4*** *for a complete list). V2G: Highest gene prioritized by OpenTargetGenetics' V2G score.* ***B)*** *Selected phenotypes associated with genome-wide significant COVID-19 variants (see* ***Supplementary Table 5*** *for a complete list). We report those associations for which a lead variant from a prior GWAS results was in high LD ($r^2 > 0.8$) with the index COVID-19 variants. The colour represents the Z-scores of correlated risk increasing alleles for the trait. The total number of associations for each COVID-19 variant is highlighted in the grey box.*

Polygenic architecture of COVID-19

To further investigate the genetic architecture of COVID-19, we used results from meta-analyses including only European ancestry samples (sample sizes described in **Methods and Supplementary Table 1**). We applied linkage disequilibrium (LD) score regression [35] to the summary statistics to estimate SNP heritability, i.e. proportion of variation in the two phenotypes that was attributable to common genetic variants, and to determine whether heritability for COVID-19 phenotypes was enriched in genes specifically expressed in certain tissues [36] from GTEx dataset [37]. We detected a low, but significant heritability across all three analyses (<1% on observed scale, all *P*-values < 0.0001; **Supplementary Table 6**). Despite these low values, which interpretation is complicated by the use of population controls and variation in the disease prevalence estimates, we found that heritability for reported infection was significantly enriched in genes specifically expressed in the lung ($P = 5.0 \times 10^{-4}$) (**Supplementary Table 7**). These findings, together with genome-wide significant loci identified in the meta-analyses, illustrate that there is a significant polygenic or oligogenic architecture that can be better leveraged with future, larger, sample sizes.

Genetic correlation and causal relationship between COVID-19 and other traits

Genetic correlations (*rg*) between the three COVID-19 phenotypes was high, though lower correlations were observed between hospitalized COVID-19 and reported infection (critical illness *vs.* hospitalized: *rg* [95%CI] = 1.37 [1.08, 1.65], $P = 2.9 \times 10^{-21}$; critical illness *vs.* reported infection, *rg* [95%CI] = 0.96 [0.71, 1.20], $P = 1.1 \times 10^{-14}$; hospitalized *vs.* reported infection: *rg* [95%CI] = 0.85 [0.68, 1.02], $P = 1.1 \times 10^{-22}$). To better understand which traits are genetically correlated and/or potentially causally associated with COVID-19 severity and SARS-CoV-2 reported infection, we chose a set of 38 disease, health and neuropsychiatric phenotypes as potential COVID-19 risk factors based on their putative relevance to the disease susceptibility, severity, or mortality (**Supplementary Table 8**).

We found evidence (FDR<0.05) of significant genetic correlations between 9 traits and hospitalized COVID-19 and SARS-CoV-2 reported infection (**Fig. 4; Supplementary Table 9**). Genetic correlation results for COVID-19 severity partially overlap with reported SARS-CoV-2 infection, with genetic liability to BMI, type 2 diabetes, smoking, and attention deficit hyperactivity disorder showing significant positive correlations (*rg* range between 0.16 - 0.26). However some results were significantly different between COVID-19 severity and reported infection. For example, genetic liability to ischemic stroke, was only significantly positively correlated with critical illness or hospitalization due to to COVID-19, but not with a higher likelihood of reported SARS-CoV-2 infection (infection *r g*= 0.019 *vs.* hospitalization *rg* = 0.41, z = 2.7, *P* = 0.006; infection *rg* = 0.019 *vs.* critical illness *rg* = 0.40, z = 2.49, *P* = 0.013). In addition, coronary artery disease, and systemic lupus erythematosus showed positive genetic correlations with critical illness or hospitalization due to COVID-19. Genetic liability to risk tolerance, on the hand, was the only trait specifically associated with SARS-CoV-2 infection. This potentially reflects that risk taking behavior could be associated with a higher chance of infection, but is not, *per-se*, impacting the chances to develop a severe form of COVID-19. With improved phenotyping of cases and controls, methods to deconvolute the effects specific to SARS-CoV-2 infection - a proxy for disease susceptibility - and those specific for progression to severe disease can be applied to better interpret these results.

We next used two-sample Mendelian randomization (MR) to infer potentially causal relationships between these traits. Fixed-effects IVW analysis was used as the primary analysis [38], with weighted median estimator (WME) [39], weighted mode based estimator (WMBE) [40], MR Egger regression [41] and MR-PRESSO [42] outlier corrected estimates used as additional sensitivity analyses.

After correcting for multiple testing (FDR < 0.05), 8 exposure — COVID-19 trait-pairs showed suggestive evidence of a causal association (**Fig. 4; Supplementary Table 10a**). Five of these associations were robust to potential violations of the underlying assumptions of MR. Corroborating our genetic correlation results and evidence from traditional epidemiological studies, genetically predicted higher BMI (OR [95%CI] 1.4 [1.3, 1.6], $P = 8.5 \times 10^{-11}$) and smoking (OR [95%CI] = 1.9 [1.3, 2.8], $P = 0.0012$) were associated with increased risk of COVID-19 hospitalization, with BMI also being associated with increased risk of SARS-CoV-2 infection (OR [95%CI] = 1.1 [1.1, 1.2], $P = 4.8 \times 10^{-7}$). Genetically predicted increased height (OR [95%CI] = 1.1 [1, 1.1]), $P = 8.9 \times 10^{-4}$) was associated with an increased risk of reported infection, and genetically predicted higher red blood cell count (OR [95%CI] = 0.93 [0.89, 0.96], $P = 5.7 \times 10^{-5}$) was associated with a reduced risk of reported infection.

We noted that there was sample overlap between some datasets used to generate exposures used in the previous analysis, and the samples contributing to our meta-analysis of hospitalized COVID-19, as a result of inclusion of samples from the UK Biobank. We therefore conducted an additional sensitivity analysis, using new hospitalized COVID-19 summary statistics in which the UK Biobank study had been removed (**Supplementary Table 10b**). In this analysis, genetically predicted BMI, height, and red blood cell counts remained significantly associated with COVID-19 outcomes ($p < 0.05$).

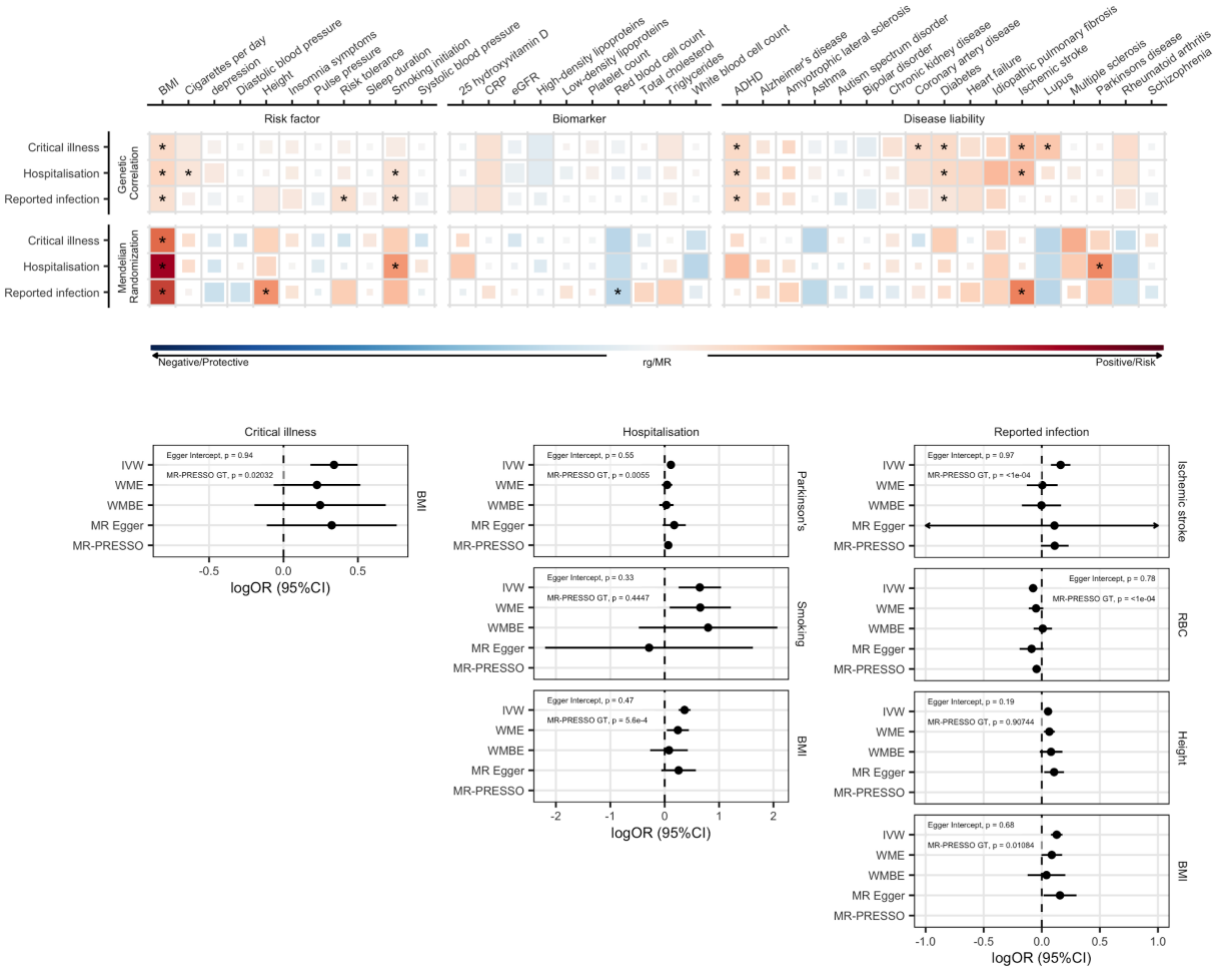


Figure 4. *Genetic correlations and Mendelian randomization causal estimates between 43 traits and COVID-19 severity and SARS-CoV-2 reported infection. Blue, negative genetic correlation and protective Mendelian randomization (MR) causal estimates; red, positive genetic correlation and risk MR causal estimates. Larger squares correspond to more significant P values, with genetic correlations or MR causal estimates significantly different from zero at a P < 0.05 shown as a full-sized square. Genetic correlations or causal estimates that are significantly different from zero at a false discovery rate (FDR) of 5% are marked with an asterisk. Forest plots display the causal estimates for each of the sensitivity analyses used in the MR analysis for trait pairs that were significant at an FDR of 5%. Individual scatter and funnel plots for each pair of traits are available in Extended Data Fig. 4.*

IVW: Inverse variance weighted analysis; WME: Weighted median estimator; WMBE: weighted mode based estimator; MR-PRESSO: Mendelian Randomization Pleiotropy RESidual Sum and Outlier. RBC: Red blood cell count

DISCUSSION

The COVID-19 Host Genetics Initiative has brought together investigators from across the world to to advance genetic discovery for SARS Cov2 infection and severe COVID-19 disease.We report 15 genome-wide significant loci associated with some aspect of SARS Cov2 infection or COVID-19. Many of these loci overlap with previously reported associations with lung-related phenotypes or autoimmune/inflammatory diseases, but some loci have no obvious candidate gene as yet.

Four out of the 15 genome-wide significant loci showed similar effects in the reported infection analysis (a proxy for disease susceptibility) and all-hospitalized COVID-19 (a proxy for disease severity). This supports the notion that some genetic variants, most notably at *ABO* and *PPP1R15A* loci, might indeed impact susceptibility to infection rather than progression to a severe form of the disease once infected. Whilst our ability to draw definitive conclusions is impaired by incomplete capture of who has been infected with SARS-CoV-2, a recent study based on self-reported exposure to COVID-19 positive housemate and consequent development of the disease, support our findings [43].

Several of the loci reported here, as noted in previous publications, intersect with well-known genetic variants that have established genetic associations. Variants at *DPP9* show prior evidence of increasing risk for interstitial lung disease. Missense variants within *TYK2* show a protective effect on several autoimmune-related diseases. Variants overlapping the well-known structural variants-rich 17q21.31 locus have been previously associated with pulmonary function. Together with the heritability enrichment observed in genes expressed in lung tissues, these results highlight the involvement of lung-related biological pathways in developing severe COVID-19. Several other loci show no prior documented genome-wide significant associations, even despite the high significance and attractive candidate genes for COVID-19 (e.g., *CXCR6*, *LZTFL1*, *IFNAR2* and *OAS1/2/3* loci). The previously reported associations for the strongest signal for COVID-19 severity at 3p21.31 and monocytes count are likely to be due to proximity and not a true co-localization.

Increasing the global representation in genetic studies enhances the ability to detect novel associations. Two of the loci affecting disease severity were only discovered by including the four studies of individuals with East Asian ancestry. One of these loci, close to *FOXP4*, is common particularly in East Asian (40%) as well as Middle Eastern and Admixed American samples in the Americas but has a low frequency in most European populations (2-3%). Previous studies have reported association between this locus and lung cancer [29,44] and interstitial lung disease [45]. Although we cannot be certain of the mechanism of action of this association *FOXP4* is an attractive biological target, as it is expressed in the proximal and distal airway epithelium [46], and has been shown to play a role in controlling epithelial cell fate during lung development [47].

A central challenge for the COVID-19 HGI was the harmonization of phenotype definitions, analytic pipelines and cohorts with extremely heterogeneous designs, sample ascertainment and control populations. Large-scale biobanks with existing genotype resources and connections to medical systems, newly enrolled hospital-based studies (particularly well-powered to study the extremes of severity by through the recruitment of individuals from intensive care units), and direct-to-consumer genetics studies with customer surveys each contributed different aspects to understanding the genetic basis of susceptibility and severity

traits. Indeed, working together through aligning phenotype definitions and sharing results accelerated progress and has enhanced the robustness of the reported findings.

Nevertheless, the differences in study sample size, ascertainment and phenotyping of COVID-19 cases are unavoidable and care should be taken when interpreting the results from a meta-analysis. First, studies enriched with severe cases or studies with antibody-tested controls may disproportionately contribute to genetic discovery despite potentially smaller sample sizes. Second, differences in genomic profiling technology, imputation, and sample size across the constituent studies can have dramatic impacts on replication and downstream analyses (particularly fine-mapping where differential missing patterns in the reported results can muddy the signal). Third, the use of population controls with no complete information about SARS-CoV-2 exposure might result in cases of misclassification or reflect ascertainment biases in testing and reporting rather than true susceptibility to infection. Genotyping large numbers of control samples who have been exposed to the virus but remained asymptomatic or experienced only mild symptoms is challenging. Therefore, many studies prefer to use pre-existing datasets of genetically ancestry-matched samples as their controls, protecting against population stratification, but potentially introducing some of these biases. Our analysis comparing the discovery meta-analysis effects to one where controls were phenotypically refined, indicated that, for genome-wide significant variants, such bias was limited.

Drawing a comprehensive and reproducible map of the host genetics factors associated with COVID-19 severity and SARS-CoV-2 requires a sustained international effort to include diverse ancestries and study designs. The number of COVID-19 study participants and studies contributing data to this study illustrate the benefits of worldwide international collaboration, open governance and planning, and sharing of technological and analytical resources. To expedite downstream scientific research and therapeutic discovery, the COVID-19 Host Genetic Initiative regularly publishes meta-analysis results from periodic data freezes on the website www.covid19hg.org as new data are included in the study. We also provide an interactive explorer where researchers can browse the results and the genomic loci in more detail. Future work will be required to better understand the biological and clinical value of these findings. Continued efforts to collect more samples and detailed phenotypic data should be endorsed globally; allowing for more thorough investigation of variable, heritable symptoms [48,49], particularly in the light of newly emerging strains of SARS-CoV-2 virus, which may provoke different host responses leading to disease, and with the enrollment of vaccines.

METHODS

Contributing studies

In total 16 studies contributed data to analysis of critical illness due to COVID-19, 29 studies contributed data to hospitalized COVID-19 analysis, and 44 studies contributed to the analysis of all COVID-19 cases. Details of contributing research groups are described in **Supplementary Table 1**. All subjects were recruited following protocols approved by local Institutional Review Boards (IRBs). All protocols followed local ethics recommendations and informed consent was obtained when required.

Phenotype Definitions

COVID-19 disease status (critical illness, hospitalization status) was assessed following the Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia [50]. The critically ill COVID-19 group included patients who were hospitalized due to symptoms associated with laboratory-confirmed SARS-CoV-2 infection and who required respiratory support or whose cause of death was associated with COVID-19. The hospitalized COVID-19 group included patients who were hospitalized due to symptoms associated with laboratory-confirmed SARS-CoV-2 infection.
The reported infection cases group included individuals with laboratory-confirmed SARS-CoV-2 infection or electronic health record, ICD coding or clinically confirmed COVID-19, or self-reported COVID-19 (e.g. by questionnaire), with or without symptoms of any severity. Genetic ancestry-matched controls for the three case definitions were sourced from population-based cohorts, including individuals whose exposure status to SARS-CoV-2 was either unknown or infection- negative for questionnaire/electronic health record-based cohorts. Additional information regarding individual studies contributing to the consortium are described in **Supplementary Table 1**.

GWAS and meta-analysis

Each contributing study genotyped the samples and performed quality controls, data imputation and analysis independently, but following consortium recommendations (information available at www.covid19hg.org). We recommended to run GWAS analysis using Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) [51] on chromosomes 1-22 and X. The recommended analysis tool was SAIGE, but studies also used other software such as PLINK [52]. The suggested covariates were age, age2, sex, age*sex, and 20 first principal components. Any other study-specific covariates to account for known technical artefacts could be added. SAIGE automatically accounts for sample relatedness and case-control imbalances. Individual study quality control and analysis approaches are reported in **Supplementary Table 1**.

Study-specific summary statistics were then processed for meta-analysis. Potential false positives, inflation, and deflation were examined for each submitted GWAS. Standard error values as a function of effective sample size was used to find studies which deviated from the expected trend. Summary statistics passing this manual quality control were included in the meta-analysis. Variants with allele frequency of >0.1% and

imputation INFO>0.6 were carried forward from each study. Variants and alleles were lifted over to genome build GRCh38, if needed, and harmonized to gnomAD 3.0 genomes [53] by finding matching variants by strand flipping or switching ordering of alleles. If multiple matching variants, the best match was chosen by minimum absolute allele frequency fold change. Meta-analysis was performed using the inverse-variance weighted method. The method summarizes effect sizes across the multiple studies by computing the mean of the effect sizes weighted by the inverse variance in each individual study.

For each of the 15 independent lead variants reported in **Table 1**, we tested whether there was heterogeneity between the effect sizes associated with hospitalized COVID-19 (progression to severe disease) and reported SARS-CoV-2 infection. We used Cochran's Q measure [54,55], which is calculated for each variant as the weighted sum of squared differences between the two analysis effects sizes and their meta-analysis effect, the weights being the inverse variance of the effect size. Q is distributed as a chi-square statistic with k (number of studies) minus 1 degrees of freedom. A significant *P*-value <0.003 (0.05/15 for multiple tests) indicates that the effect sizes for a particular variant are significantly different in the two analyses. For the 11 loci, where the lead variant effect size was significantly higher for hospitalized COVID-19, we carried out the same test again but comparing effect sizes from hospitalized COVID-19 with critically ill COVID-19 (**Supplementary Table 3**). Further, we carried out the same test comparing meta-analyzed hospitalized COVID-19 (population as controls) and hospitalized COVID-19 (SARS-Cov-2 positive but non-hospitalized as controls) (**Supplementary Table 3)**. For these pairs of phenotype comparisons, we generated new meta-analysis summary statistics to use; including only those studies that could contribute data to both phenotypes that were under comparison.

Gene prioritization

To prioritize candidate causal genes, we employed various gene prioritization approaches using both locus-based and similarity-based methods. Because we only referred *in-silico* gene prioritization results without characterizing actual functional activity *in-vitro/vivo*, we aimed to provide a conservative list of any potential causal genes in a locus using the following criteria:

1. Closest gene: a gene that is closest to a lead variant by distance to the gene body
2. Genes in LD region: genes that overlap with a genomic range containing any variants in LD (r^2 > 0.6) with a lead variant. For LD computation, we retrieved LD matrices provided by the gnomAD v2.1.1 [53] for each population analyzed in this study (except for Admixed American, Middle Eastern, and South Asian that are not available). We then constructed a weighted-average LD matrix by per-population sample sizes in each meta-analysis, which we used as a LD reference.
3. Genes with coding variants: genes with at least one loss of function or missense variant (annotated by VEP [56] v95 with GENCODE v29) that is in LD with a lead variant (r^2 > 0.6).
4. eGenes: genes with at least one fine-mapped *cis*-eQTL variant (PIP > 0.1) that is in LD with a lead variant (r^2 > 0.6) (**Supplementary Table 4**). We retrieved fine-mapped variants from the GTEx v8[26] (https://www.finucanelab.org/) and eQTL catalogue[57]. In addition, we looked up significant associations in the Lung eQTL Consortium [27] (n = 1,103) to further support findings in lung with a larger sample size (**Supplementary Table 11**). We note that, unlike the GTEx or eQTL catalogue, we only looked at associations and didn't finemap in the Lung eQTL Consortium data.

5. V2G: a gene with the highest overall Variant-to-Gene (V2G) score based on the Open Targets Genetics (OTG) [23]. For each variant, the overall V2G score aggregates differentially weighted evidence of variant-gene association from several data sources, including molecular cis-QTL data (e.g., cis-pQTLs from [58]., cis-eQTLs from GTEx v7 etc.), interaction-based datasets (e.g., Promoter Capture Hi-C), genomic distance, and variant effect predictions (VEP) from Ensembl. A detailed description of the evidence sources and weights used is provided in the OTG documentation (https://genetics-docs.opentargets.org/our-approach/data-pipeline) [23,59].

Phenome-wide association study

To investigate the evidence of shared effects of 15 index variants for COVID-19 and previously reported phenotypes, we performed a phenome-wide association study. We considered phenotypes in (Open Target) OTG obtained from the GWAS catalog (this included studies with and without full summary statistics, n = 300 and 14,013, respectively) [60], and from UK Biobank. Summary statistics for UK Biobank traits were extracted from SAIGE [51] for binary outcomes (n = 1,283 traits), and Neale v2 (n = 2,139 traits) for both binary and quantitative traits (http://www.nealelab.is/uk-biobank/)and FinnGen Freeze 4 cohort (https://www.finngen.fi/en/access_results). To remove plausible spurious associations, we retrieved phenotypes for GWAS lead variants that were in LD (r^2>0.8) with COVID-19 index variants.

Heritability

LD score regression v 1.0.1 [35] was used to estimate SNP heritability of the phenotypes from the meta-analysis summary statistic files. As this method depends on matching the linkage disequilibrium (LD) structure of the analysis sample to a reference panel, the European-only summary statistics were used. Sample sizes were n = 5,101 critically ill COVID-19 cases and n = 1,383,241 controls, n = 9,986 hospitalized COVID-19 cases and n = 1,877,672 controls, and n = 38,984 cases and n = 1,644,784 controls for all cases analysis, all including the 23andMe cohort. Pre-calculated LD scores from the 1000 Genomes European reference population were obtained online (https://data.broadinstitute.org/alkesgroup/LDSCORE/). Analyses were conducted using the standard program settings for variant filtering (removal of non-HapMap3 SNPs, non-autosomal, chi-square > 30, MAF < 1%, or allele mismatch with reference). We additionally report SNP heritability estimates for the all-ancestries meta-analyses, calculated using European panel LD scores, in **Supplementary Table 6**.

Partitioned heritability

We used partitioned LD score regression [61] to partition COVID-19 SNP heritability in cell types in our European-only summary statistics. We ran the analysis using the baseline model LD scores calculated for European populations and regression weights that are available online. We used the COVID-19 European only summary statistics for the analysis.

Genome-wide association summary statistics

We obtained genome-wide association summary statistics for 43 complex disease, neuropsychiatric, behavioural, or biomarker phenotypes (**Supplementary Table 8**). These phenotypes were selected based on their putative relevance to COVID-19 susceptibility, severity, or mortality, with 19 selected based on the Centers for Disease Control list of underlying medical conditions associated with COVID-19 severity [62] or traits reported to be associated with increased risk of COVID-19 mortality by OpenSafely [63]. Summary statistics generated from GWAS using individuals of European ancestry were preferentially selected if available. These summary statistics were used in subsequent genetic correlation and Mendelian randomization analyses.

Genetic Correlation

LD score regression [61] was also used to estimate genetic correlations between our COVID-19 meta-analysis phenotypes reported using European-only samples, and between these and the curated set of 38 summary statistics. Genetic correlations were estimated using the same LD score regression settings as for heritability calculations. Differences between the observed genetic correlations of SARS-COV2 infection and COVID-19 severity were compared using a z score method [64].

Mendelian Randomization

Two-sample Mendelian randomization was employed to evaluate the causal association of the 38 traits on COVID-19 hospitalization, on COVID-19 severity and SARS-CoV-2 reported infection using European-only samples. Independent genome-wide significant SNPs robustly associated with the exposures of interest ($P < 5 \times 10^{-8}$) were selected as genetic instruments by performing LD clumping using PLINK [52]. We used a strict r^2 threshold of 0.001, a 10MB clumping window, and the European reference panel from the 1000 Genomes project [65] to discard SNPs in linkage disequilibrium with another variant with smaller p-value association. For genetic variants that were not present in the hospitalized COVID analysis, PLINK was used to identify proxy variants that were in LD ($r^2 > 0.8$). Next, the exposure and outcome datasets were harmonized using the R-package TwoSampleMR [66]. Namely, we ensured that the effect of a variant on the exposure and outcome corresponded to the same allele, we inferred positive strand alleles and dropped palindromes with ambiguous allele frequencies, as well as incompatible alleles. **Supplementary Table 8** includes the harmonized datasets used in the analyses.

Mendelian Randomization Pleiotropy residual sum and outlier (MR-PRESSO) Global test [42] was used to investigate overall horizontal pleiotropy. In short, the standard IVW meta-analytic framework was employed to calculate the average causal effect by excluding each genetic variant used to instrument the analysis. A global statistic was calculated by summing the observed residual sum of squares, i.e., the difference between the effect predicted by the IVW slope excluding the SNP, and the observed SNP-effect on the outcome. Overall horizontally pleiotropy was subsequently probed by comparing the observed residual sum of squares, with the residual sum of squares expected under the null hypothesis of no pleiotropy. The MR-PRESSO Global test was shown to perform well when the outcome and exposure

GWASs are not disjoint (although the power to detect horizontal pleiotropy is slightly reduced by complete sample overlap). We also used the MR-Egger regression intercept [41] to evaluate potential bias due to directional pleiotropic effects. This additional check was employed in MR analyses with an I^2_{GX} index surpassing the recommended threshold (I^2_{GX} > 90%; [67]). Contingent on the MR-PRESSO Global test results we probed the causal effect of each exposure on COVID-19 hospitalization by using a fixed effect inverse-weighted (IVW) meta-analysis as the primary analysis, or, if pleiotropy was present, the MR-PRESSO outlier corrected test. The IVW approach estimates the causal effect by aggregating the single-SNP causal effects (obtained using the ratio of coefficients method, i.e., the ratio of the effect of the SNP on the outcome on the effect of the SNP on the exposure) in a fixed effects meta-analysis. The SNPs were assigned weights based on their inverse variance. The IVW method confers the greatest statistical power for estimating causal associations [68], but assumes that all variants are valid instruments and can produce biased estimates if the average pleiotropic effect differs from zero. Alternatively, when horizontal pleiotropy was present, we used MR-PRESSO Outlier corrected method to correct the IVW test by removing outlier SNPs. We conducted further sensitivity analyses using alternative MR methods that provide consistent estimates of the causal effect even when some instrumental variables are invalid, at the cost of reduced statistical power including: 1) Weighted Median Estimator (WME); 2) Weighted Mode Based Estimator (WMBE); 3) MR-Egger regression. Robust causal estimates were defined as those that were significant at an FDR of 5% and either 1) showed no evidence of heterogeneity (MR-PRESSO Global test $P > 0.05$) or horizontal pleiotropy (Egger Intercept $P > 0.05$), or 2) in the presence of heterogeneity or horizontal pleiotropy, either the WME, WMBE, MR-Egger or MR-PRESSO corrected estimates were significant ($P < 0.05$). All statistical analyses were conducted using R version 4.0.3. MR analysis was performed using the "TwoSampleMR" version 0.5.5 package [66].

Website and data distribution

In anticipation of the need to coordinate many international partners around a single meta-analysis effort, we created the COVID-19 HGI website (https://covid19hg.org). We were able to centralize information, recruit partner studies, rapidly distribute summary statistics, and present preliminary interpretations of the results to the public. Open meetings are held on a monthly basis to discuss future plans and new results; video recordings and supporting documents are shared (https://covid19hg.org/meeting-archive). This centralized resource provides a conceptual and technological framework for organizing global academic and industry groups around a shared goal. The website source code and additional technical details are available at https://github.com/covid19-hg/covid19hg.

To recruit new international partner studies, we developed a workflow whereby new studies are registered and verified by a curation team (https://covid19hg.org/register). Users can explore the registered studies using a customized interface to find and contact studies with similar goals or approaches (https://covid19hg.org/partners). This helps to promote organic assembly around focused projects that are adjacent to the centralized effort (https://covid19hg.org/projects). Visitors can query study information, including study design and research questions. Registered studies are visualized on a world map and are searchable by institutional affiliation, city, and country.

To encourage data sharing and other forms of participation, we created a rolling acknowledgements page (https://covid19hg.org/acknowledgements) and directions on how to contribute data to the central meta-analysis effort (https://covid19hg.org/data-sharing). Upon the completion of each data freeze, we post summary statistics, plots, and sample size breakdowns for each phenotype and contributing cohort (https://covid19hg.org/results). The results can be explored using an interactive web browser (https://app.covid19hg.org). Several computational research groups carry out follow-up analyses, which are made available for download (https://covid19hg.org/in-silico). To enhance scientific communication to the public, preliminary results are described in blog posts by the scientific communications team and shared on Twitter. The first post was translated to 30 languages with the help of 85 volunteering translators. We compile publications and preprints submitted by participating groups and summarize genome-wide significant findings from these publications (https://covid19hg.org/publications).

ACKNOWLEDGEMENTS

We thank the entire COVID-19 Host Genetics Initiative community for their contributions and continued collaboration. The work of the contributing studies was supported by numerous grants from governmental and charitable bodies, and study specific acknowledgements will be released with the publication. We thank G. Butler-Laporte, G. Wojcik, M.-G. Hollm-Delgado, C. Willer and G. Davey Smith for their extensive feedback and discussion.

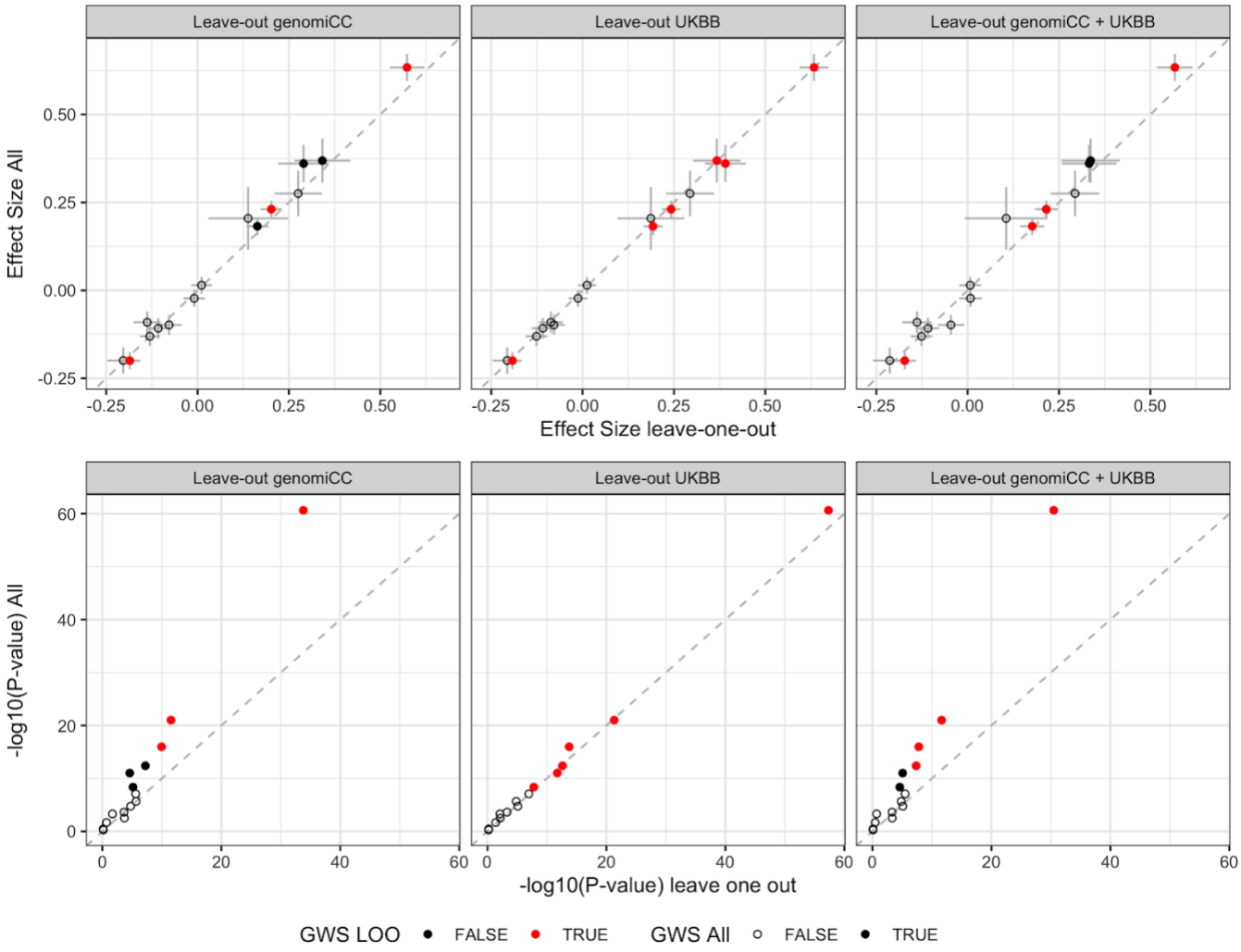
DATA AVAILABILITY

Summary statistics generated by COVID-19 HGI are available at https://www.covid19hg.org/results/r5/ and will be made available on GWAS Catalog. The analyses described here utilize the freeze 5 data. COVID-19 HGI continues to regularly release new data freezes. Summary statistics for non-European ancestry samples are not currently available due to the small individual sample sizes of these groups.

CODE AVAILABILITY

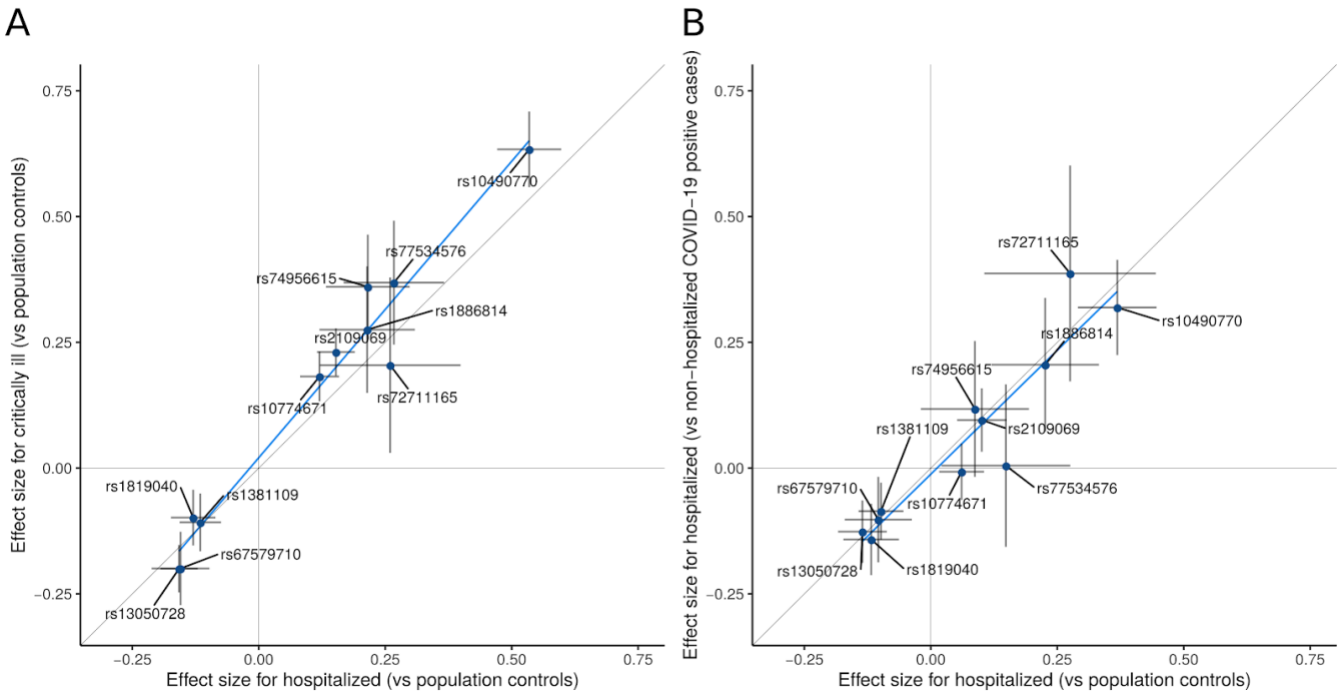
Meta-analysis code is available at https://github.com/covid19-hg/META_ANALYSIS/.

EXTENDED DATA FIGURES



Extended Data Figure 1. Sensitivity analyses for overlapping controls in genomiCC and UK Biobank.

Comparison of the effect sizes and *P*-values of the 15 lead variant, using data from the COVID-19 critical illness meta-analysis in all the cohorts (y-axis) to leaving out genomiCC, leaving out UK Biobank (UKBB) and leaving out genomiCC + UKBB, respectively (x-axis). Dots represent the effect size estimates (top panels) and *P*-values (bottom panels), and bars represent the standard error. Filled dots indicate variants that were significant in the full meta-analysis of critical illness due to COVID-19, and empty dots represent variants that were not significant for critical illness but were significant for either hospitalization due to COVID-19 or SARS-CoV-2 reported infection. Red dots represent variants that were significant in leave-one-out analysis for genomiCC, UKBB or genomiCC + UKBB.



Extended Data Figure 2. Comparison of lead variant effect sizes between pairs of COVID-19 meta-analyses.

Comparison of effect sizes for the 11 variants associated with severity of COVID-19 disease. A. Comparing hospitalized COVID-19 cases vs population controls (x-axis, n=10,428 cases and n=1,483,270 controls) and critically ill COVID-19 cases vs population controls (y-axis, n=6,179 cases and n=1,483,780 controls). B. hospitalized COVID-19 cases vs population controls (x-axis,n=5,806 cases and n=1,144,263 controls)and hospitalized COVID-19 cases vs non-hospitalized COVID-19 cases (y-axis, n=5,773 and n=15,497 controls). Dots represent the effect size estimates, bars represent the confidence interval of the estimates. Effect size estimates and *P*-values for heterogeneity test are reported in Supplementary Table 3.

Please find the corresponding figure in the supplementary PDF "Extended Data Figure 3"

Extended Data Figure 3. LocusZoom plots for each COVID-19 locus in three meta-analyses.

For each genome-wide significant locus in three meta-analyses: critical illness (labelled as Analysis A2), hospitalization (labelled as Analysis B2), and reported infection (labelled as Analysis C2), we showed 1) a manhattan plot of each locus where a color represents a weighted-average r^2 value (see **Methods**) to a lead variant; 2) r2 values to a lead variant across gnomAD v2 populations, i.e., African/African-American (AFR), Latino/Admixed American (AMR), Ashkenazi Jewish (ASJ), East Asian (EAS), Estonian (EST), Finnish (FIN), Non-Finish Europeans (NFE), North-Western Europeans (NWE), and Southern Europeans (SEU); 3) genes at a locus; and 4) genes prioritized by each gene prioritization metric where a size of circles represents a rank in each metric. Note that the COVID-19 lead variants were chosen across all the

meta-analyses (**Table 1**; see **Methods**) and were not necessarily a variant with the most significant *P*-value in each meta-analysis.

Please find the corresponding figure in the supplementary PDF "Extended Data Figure 4"

Extended Data Figure 4. Scatter and funnel plots for each for exposure - COVID-19 outcome pair. Scatter plots show the exposure variant effect size against the COVID-19 outcome variant effect size and corresponding standard errors. Funnel plots show the Mendelian randomization (MR) causal estimates for each variant against their precision, with asymmetry in the plot indicating potential violations of the assumptions of MR. Regression lines show the corresponding causal estimates fixed effect inverse-weighted (IVW, red-solid line) meta-analysis; MR-Egger regression (blue-dashed); Weighted median estimator (WME, green-dashed); weighted mode based estimator (WMBE, purple dashed); and Mendelian Randomization Pleiotropy RESidual Sum and Outlier corrected (MR-PRESSO, orange dashed). Variants highlighted in red were flagged as outliers by MR-PRESSO.

REFERENCES

1. Buitrago-Garcia, D. *et al.* Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS Med.* **17**, e1003346 (2020).
2. Docherty, A. B. *et al.* Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* **369**, m1985 (2020).
3. Casanova, J.-L. & Abel, L. Lethal Infectious Diseases as Inborn Errors of Immunity: Toward a Synthesis of the Germ and Genetic Theories. *Annu. Rev. Pathol.* **16**, 23–50 (2021).
4. Ovsyannikova, I. G., Haralambieva, I. H., Crooke, S. N., Poland, G. A. & Kennedy, R. B. The role of host genetics in the immune response to SARS-CoV-2 and COVID-19 susceptibility and severity. *Immunol. Rev.* **296**, 205–219 (2020).
5. Everitt, A. R. *et al.* IFITM3 restricts the morbidity and mortality associated with influenza. *Nature* **484**, 519–523 (2012).
6. Samson, M. *et al.* Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **382**, 722–725 (1996).
7. 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.* **79**, 15351–15355 (2005).
8. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).
9. Thomas, D. L. *et al.* Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature* **461**, 798–801 (2009).
10. Kenney, A. D. *et al.* Human Genetic Determinants of Viral Diseases. *Annu. Rev. Genet.* **51**, 241–263 (2017).

11. van der Made, C. I. *et al.* Presence of Genetic Variants Among Young Men With Severe COVID-19. *JAMA* (2020) doi:10.1001/jama.2020.13719.
12. Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, (2020).
13. Bastard, P. *et al.* Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370**, (2020).
14. Povysil, G. *et al.* Failure to replicate the association of rare loss-of-function variants in type I IFN immunity genes with severe COVID-19. *medRxiv* (2020) doi:10.1101/2020.12.18.20248226.
15. Severe Covid-19 GWAS Group *et al.* Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N. Engl. J. Med.* **383**, 1522–1534 (2020).
16. Shelton, J. F. *et al.* Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity. *bioRxiv* (2020) doi:10.1101/2020.09.04.20188318.
17. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in Covid-19. *Nature* (2020) doi:10.1038/s41586-020-03065-y.
18. Roberts, G. H. L. *et al.* AncestryDNA COVID-19 host genetic study identifies three novel loci. *bioRxiv* (2020) doi:10.1101/2020.10.06.20205864.
19. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet.* **28**, 715–718 (2020).
20. Kosmicki, J. A. *et al.* Genetic association analysis of SARS-CoV-2 infection in 455,838 UK Biobank participants. *bioRxiv* (2020) doi:10.1101/2020.10.28.20221804.
21. Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).
22. Finer, S. *et al.* Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *Int. J. Epidemiol.* **49**, 20–21i (2020).
23. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkaa840.
24. Dendrou, C. A. *et al.* Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* **8**, 363ra149 (2016).
25. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415–1429.e19 (2016).
26. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
27. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8**, e1003029 (2012).
28. Fingerlin, T. E. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat. Genet.* **45**, 613–620 (2013).
29. Wang, Z. *et al.* Meta-analysis of genome-wide association studies identifies multiple lung cancer susceptibility loci in never-smoking Asian women. *Hum. Mol. Genet.* **25**, 620–629 (2016).
30. Shrine, N. *et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493 (2019).
31. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).

32. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44**, 881–885 (2012).
33. Xiao, G. *et al.* CXCL16/CXCR6 chemokine signaling mediates breast cancer progression by pERK1/2-dependent mechanisms. *Oncotarget* **6**, 14165–14178 (2015).
34. Wei, Q. *et al.* LZTFL1 suppresses lung tumorigenesis by maintaining differentiation of lung epithelial cells. *Oncogene* **35**, 2655–2663 (2016).
35. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
36. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
37. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
38. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
39. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
40. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985–1998 (2017).
41. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
42. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
43. Roberts, G. H. L. *et al.* Novel COVID-19 phenotype definitions reveal phenotypically distinct patterns of genetic association and protective effects. *bioRxiv* (2021) doi:10.1101/2021.01.24.21250324.
44. Dai, J. *et al.* Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *Lancet Respir Med* **7**, 881–891 (2019).
45. Manichaikul, A. *et al.* Genome-wide association study of subclinical interstitial lung disease in MESA. *Respir. Res.* **18**, 97 (2017).
46. Lu, M. M., Li, S., Yang, H. & Morrisey, E. E. Foxp4: a novel member of the Foxp subfamily of winged-helix genes co-expressed with Foxp1 and Foxp2 in pulmonary and gut tissues. *Gene Expr. Patterns* **2**, 223–228 (2002).
47. Li, S. *et al.* Foxp1/4 control epithelial cell fate during lung development and regeneration through regulation of anterior gradient 2. *Development* **139**, 2500–2509 (2012).
48. Meng, X., Deng, Y., Dai, Z. & Meng, Z. COVID-19 and anosmia: A review based on up-to-date knowledge. *Am. J. Otolaryngol.* **41**, 102581 (2020).
49. Williams, F. M. K. *et al.* Self-reported symptoms of covid-19 including symptoms most predictive of SARS-CoV-2 infection, are heritable. *bioRxiv* (2020) doi:10.1101/2020.04.22.20072124.
50. Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (Trial Version 7). *Chin. Med. J.* **133**, 1087–1095 (2020).
51. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0184-y.

52. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
53. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
54. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
55. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* **10**, 101–129 (1954).
56. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
57. Kerimov, N. *et al.* eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. *Cold Spring Harbor Laboratory* 2020.01.29.924266 (2021) doi:10.1101/2020.01.29.924266.
58. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
59. Mountjoy, E. *et al.* Open Targets Genetics: An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Cold Spring Harbor Laboratory* 2020.09.16.299271 (2020) doi:10.1101/2020.09.16.299271.
60. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
61. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
62. CDC. COVID-19 and Your Health. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html (2021).
63. Williamson, E. J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020).
64. Zhou, T. *et al.* Educational attainment and drinking behaviors: Mendelian randomization study in UK Biobank. *Mol. Psychiatry* (2019) doi:10.1038/s41380-019-0596-9.
65. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
66. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).
67. Bowden, J. *et al.* Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int. J. Epidemiol.* **45**, 1961–1974 (2016).
68. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* **44**, 313–329 (2020).