nature

https://doi.org/10.1038/s41586-021-03767-x

Accelerated Article Preview

Mapping the human genetic architecture of COVID-19

Received: 2 March 2021

Accepted: 23 June 2021

Accelerated Article Preview Published online 8 July 2021

Cite this article as: COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* https://doi.org/10.1038/s41586-021-03767-x (2021).

COVID-19 Host Genetics Initiative

Mapping the human genetic architecture of COVID-19

https://doi.org/10.1038/s41586-021-03767-x

Received: 2 March 2021

Accepted: 23 June 2021

Published online: 8 July 2021

COVID-19 Host Genetics Initiative*

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[1,2], host genetics may also be important. Identifying host-specific genetic factors may reveal 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 comprised of up to 49,562 COVID-19 patients from 46 studies across 19 countries. We reported 13 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[3–7]. They also represent potentially actionable mechanisms in response to infection. Mendelian Randomization analyses support a causal role for smoking and body mass index for severe COVID-19 although not for type II diabetes. The identification of novel host genetic factors associated with COVID-19, with unprecedented speed, was made possible by the community of human genetic researchers coming together to prioritize sharing of data, results, resources and analytical frameworks. This working model of international collaboration underscores what is possible for future genetic discoveries in emerging pandemics, or indeed for any complex human disease.

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 the variation in consequence ranging from asymptomatic to life-threatening, viral pneumonia and acute respiratory distress syndrome[8]. While established host factors correlate with disease severity (e.g., increasing age, being a man, and higher body mass index[1]), these risk factors alone do not explain all variability in disease severity observed across individuals.

Genetic factors contributing to COVID-19 susceptibility and severity may provide 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. Further suggesting this line of inquiry, rare loss-of-function variants in genes involved in type I interferon (*IFN*) response may be involved in severe forms of COVID-19[9–12]. At the same time, several genome-wide association studies (GWAS) that investigate the contribution of common genetic variation[13–16] to COVID-19 have provided robust support for the involvement of several genomic loci associated with COVID-19 severity and susceptibility, with the strongest and most robust finding for severity being at locus 3p21.31[13–17]. 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/)[18] 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. Here, we report the latest results of meta-analyses of 46 studies from 19 countries (Fig. 1) for COVID-19 host genetic effects.

Meta-analyses of COVID-19

Overall, the COVID-19 Host Genetics Initiative combined genetic data from 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). An overview of the study design is provided in Extended Data Figure 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

*A list of authors and their affiliations appears in the Supplementary Information. ✉email: bneale@broadinstitute.org; mark.daly@helsinki.fi; andrea.ganna@helsinki.fi

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 regardless of symptoms (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). The average age of COVID-19 cases across studies was 55 years (Supplementary Table 1). We report quantile-quantile plots as Supplementary Figure 1 and ancestry principal component plots for contributing studies in Extended Data Figure 2.

Across our three analyses, we reported a total of 13 independent genome-wide significant loci associated with COVID-19 ($P < 1.67 \times 10^{-8}$ threshold adjusted for multiple trait testing) (Supplementary Table 2), most of which were shared between two or more COVID-19 phenotypes. Two of these loci are in very close proximity within the 3p21.31 region, which was previously reported as one single locus associated with COVID-19 severity[13–17] (Extended Data Figure 3). Overall, we find six genome-wide significant associations for critical illness due to COVID-19, using data for 6,179 cases and 1,483,780 controls from 16 studies (Extended Data Figure 4). Nine genome-wide significant loci were detected for moderate to severe hospitalized COVID-19 (including five of the six critical illness loci), from an analysis of 13,641 COVID-19 cases and 2,070,709 controls, across 29 studies (Fig. 2a top panel). Finally, seven 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. 2a bottom panel). The proportion of cases with non-European genetic ancestry for each of the three analyses was 23%, 29% and 22%, respectively. We report the results for the lead variants at the 13 loci in different ancestry-group meta-analyses in Supplementary Table 3. We note that two loci, tagged by lead variants rs1886814 and rs72711165, had higher allele frequencies in South East Asian (rs1886814, 15%) and East Asian genetic ancestry (rs72711165, 8%) whilst the minor allele frequencies in European populations were < 3%. This highlights the value of including data from diverse populations for genetic discovery. We discuss replication of previous findings and the new discoveries from these three analyses in our Supplementary Note.

Variant effects on severity vs. susceptibility

We found no genome-wide significant sex-specific effects at the 13 loci. However, we did identify significant heterogeneous effects ($P<0.004$) across studies for 3 out of the 13 loci (Methods), likely reflecting differential ascertainment of cases (Supplementary Table 2). 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 (Supplementary Information, Extended Data Figure 5).

We next wanted to better understand whether the 13 significant loci were acting through mechanisms increasing susceptibility to infection or by affecting the progression of symptoms towards more severe disease. For all 13 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.

Focusing on the two better powered analyses: all cases with reported infection and all cases hospitalized due to COVID-19, we find four of the loci have similar odds ratios between these two analyses (Methods) (Supplementary Table 2). Such consistency suggests a stronger link to susceptibility to SARS-CoV-2 infection rather than to the development of severe COVID-19. The strongest susceptibility signal was the previously reported *ABO* locus (rs912805253)[13,14,16,17]. Interestingly, and in agreement with the report by Robert and colleagues[16], we also report a locus within the 3p21.31 region that was more strongly associated with susceptibility to SARS-CoV-2 than progression to more severe COVID-19 phenotypes. Rs2271616 showed a stronger association with reported infection ($P=1.79\times10^{-34}$; OR[95%CI]= 1.15 [1.13-1.18]) than hospitalization ($P=1.05\times10^{-5}$; OR[95%CI]=1.12[1.06-1.19]). For this locus, which contains additional independent signals, the linkage-disequilibrium pattern is discordant with the *P*-value expectation (Supplementary Note; Extended Data Figure 6), pointing to a key missing causal variant or to a potentially undiscovered multi-allelic or structural variant in this locus.

In contrast, nine out of the 13 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 (eight loci below threshold $P<0.004$ test for effect size difference, and additionally lead variant rs10774671 had a clear increase in ORs despite not passing this threshold) (Supplementary Table 2). We further compared the ORs for these nine 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 Figure 7a, Supplementary Table 4), but the lower power for association analysis of critically ill COVID-19 means that these results should be considered as suggestive. Overall, these results indicated that these nine 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.

For some of these analyses, the controls were simply existing population controls without knowledge of SARS-CoV-2 infection or COVID-19 status, which may bias effect size estimates as some of these individuals may have either become infected with SARS-CoV-2 or developed COVID-19. We perform several sensitivity analyses (Supplementary Note; Extended Data Figure 7b; Supplementary Table 4) showing that using population controls can be a valid and powerful strategy for host genetic discovery of infectious disease, and particularly those that are widespread and with rare severe outcomes.

Gene prioritization and PheWas

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. Of the 13 genome-wide significant loci, we found nine loci to implicate biologically plausible genes (Supplementary Table 2, Supplementary Table 5). 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 (OR[95%] = 1.43 [1.29, 1.59]; $P = 9.71\times10^{-12}$) and hospitalization due to COVID-19 (OR [95%CI] = 1.27 [1.18, 1.36]; $P = 5.05\times10^{-10}$) 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[3] as this variant is known to reduce function[19,20]. In contrast, this missense variant was previously reported to be protective against autoimmune diseases (Extended Data Figure 8; Supplementary Table 6), including rheumatoid arthritis (OR = 0.74; $P = 3.0\times10^{-8}$; UKB SAIGE), and hypothyroidism (OR = 0.84; $P = 1.8\times10^{-10}$; UK Biobank). 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]; $P = 2.1\times10^{-8}$), is in LD ($r^2 = 0.93$) with a missense variant rs11541192:G>A (p.Gly312Ser) in *PPP1R15A*.

Lung-specific *cis*-eQTL from GTEx v8[21] ($n = 515$) and the Lung eQTL Consortium[22] ($n = 1{,}103$) provided further support for a subset of loci (Supplementary Table 7), including *FOXP4* (6p21.1) and *ABO* (9q34.2), *OAS1/OAS3/OAS2* (12q24.13), and *IFNAR2/IL10RB* (21q22.11), where the COVID-19 associated variants modify gene expression in lung. Furthermore, our PheWAS analysis (Supplementary Table 6) 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, $P = 2.0 \times 10^{-12}$)[5]. The COVID-19 lead variant rs1886814:A>C in *FOXP4* locus is correlated ($r^2 = 0.64$) with a lead variant of lung adenocarcinoma (tag variant=rs7741164; OR=1.2, $P=6.0 \times 10^{-13}$)[6,23] and similarly with a lead variant reporting in subclinical interstitial lung disease[24]. In severe COVID, lung cancer and ILD, the minor, expression increasing allele is associated with increased risk. We also found that intronic variants (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, OR [95%CI] = 0.96 [0.95, 0.97], $P = 1.00 \times 10^{-20}$)[7]. 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[25,26].

Lastly, there are two loci in the 3p21.31 region with varying genes prioritized by different methods for different independent signals. For the severity lead variant rs10490770:T>C, we prioritized *CXCR6* with the Variant2Gene (V2G) algorithm[27], although *LZTFL1* is the closest gene. The *CXCR6* plays a role in chemokine signaling[28], and *LZTFL1* has been implicated in lung cancer[29]. Rs2271616:G>T, associated with susceptibility, tags a complex region including several independent signals (Supplementary Note) all located within a gene body of *SLC6A20* which is known to functionally interact with the SARS-CoV-2 receptor ACE2[30]. However, none of the lead variants in the 3p21.31 region has been previously associated with other traits or diseases in our PheWAS analysis. While these results provide supporting *in-silico* evidence for candidate causal gene prioritization, further functional characterization is strongly needed. Detailed locus descriptions and LocusZoom plots are provided in Supplementary Figure 2.

Polygenic architecture of COVID-19

To further investigate the genetic architecture of COVID-19, we used results from meta-analyses including samples from European ancestries (sample sizes described in Methods and Supplementary Table 1) 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[31] from GTEx dataset[32]. We detected a low, but significant heritability across all three analyses (<1% on observed scale, all *P*-values < 0.0001, LDSC intercept range 1.0024-1.0137; Supplementary Table 8). The values are low compared to previously published studies[15] but may be explained by differences in reported estimate scale (observed *vs*. liability), the specific method used, disease prevalence estimates, phenotypic differences between patient cohorts or ascertainment of controls. Despite the low reported values, 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 9). These findings, together with genome-wide significant loci identified in the meta-analyses, suggest that there is a significant polygenic architecture that can be better leveraged with future, larger, sample sizes.

Genetic correlation Mendelian Randomization

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 hospitalization, critical illness 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 clinical correlation with disease susceptibility, severity, or mortality (Supplementary Table 10).

We found evidence (FDR<0.05) of significant genetic correlations between 9 traits and hospitalized COVID-19 and SARS-CoV-2 reported infection (Fig. 3; Extended Data Figure 9**;** Supplementary Table 11). Interesting findings include that genetic liability to ischemic stroke was only significantly positively correlated with critical illness or hospitalization due to COVID-19, but not with a higher likelihood of reported SARS-CoV-2 infection (infection *rg*= 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).

We next used two-sample Mendelian randomization (MR) to infer potentially causal relationships between these traits. After correcting for multiple testing (FDR < 0.05), 8 exposure – COVID-19 trait-pairs showed suggestive evidence of a causal association (Fig. 3; Supplementary Table 12; Extended Data Figure 10; Supplementary Figure 3). Five of these associations were robust to potential violations of the underlying assumptions of MR. Corroborating our genetic correlation results and evidence from 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], $P = 5.7 \times 10^{-5}$) with a reduced risk of reported infection. Despite the evidence of genetic correlation between type II diabetes and COVID-19 outcomes, there was no evidence of a causal association in the MR analyses, suggesting that the observed genetic correlations are due to pleiotropic effects between BMI and type II diabetes. Further sensitivity analyses relating to sample overlap are discussed in Supplementary Information.

Discussion

The COVID-19 Host Genetics Initiative has brought together investigators from across the world to advance genetic discovery for SARS-CoV-2 infection and severe COVID-19 disease. We report 13 genome-wide significant loci associated with some aspect of SARS-CoV-2 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.

Four out of the 13 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). Of these, one locus was in close proximity, but yet independent, to the major genetic signal for COVID-19 severity at 3p21.31. Surprisingly, this locus was associated with COVID-19 susceptibility rather than severity. The locus overlaps SLC6A20, which encodes an amino acid transporter that interacts with ACE2. Nonetheless, we caution that more data is needed to resolve the nature of the relationship between genetic variation and COVID-19 at this locus, particularly as the physical proximity, linkage disequilibrium structure and patterns of association suggest that untagged genetic variation might be drive the association signal in the region. Our findings support the notion that some genetic variants, most notably at *ABO* and *PPP1R15A* loci, in addition to the aforementioned SLC6A20, might indeed impact susceptibility to infection rather than progression to severe COVID-19 once infected.

Several of the loci reported here, as noted in previous publications[13,15], intersect with well-known genetic variants that have established genetic associations. Examples of these include variants at *DPP9* and FOXP4 which show prior evidence of increasing risk for interstitial lung

disease[5], and missense variants within *TYK2* that show a protective effect on several autoimmune-related diseases[33–36]. 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 association 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 (32%) as well as Admixed American in the Americas (20%) and Middle Eastern samples (7%), but has a low frequency in most European ancestries (2-3%) in our data. Although we cannot be certain of the mechanism of action of *FOXP4* association is an attractive biological target, as it is expressed in the proximal and distal airway epithelium[37], and has been shown to play a role in controlling epithelial cell fate during lung development[38]. The COVID-19 Host genetics Initiative continues to pursue expansion of the datasets included in the consortium's analyses to populations from underrepresented populations in upcoming data releases. We plan to release ancestry-specific results in full once the sample sizes allow for a well-powered meta-analysis.

Care should be taken when interpreting the results from a meta-analysis because of challenges with cases and controls ascertainment and collider bias (see Supplementary Note for a more detailed discussion on study limitations). 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. To accelerate downstream 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 and provides 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[39,40], particularly in the light of newly emerging strains of SARS-CoV-2 virus, which may provoke different host responses leading to disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03767-x.

1. 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).
2. 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* **395**, 1054–1062 (2020).
3. Dendrou, C. A. *et al*. Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* **8**, 363ra149 (2016).
4. 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).
5. Fingerlin, T. E. *et al*. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat. Genet.* **45**, 613–620 (2013).
6. 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).
7. 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).
8. 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).
9. van der Made, C. I. *et al*. Presence of Genetic Variants Among Young Men With Severe COVID-19. *JAMA* (2020) https://doi.org/10.1001/jama.2020.13719.
10. Zhang, Q. *et al*. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, (2020).
11. Bastard, P. *et al*. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370**, (2020).
12. Povysil, G. *et al*. Rare loss-of-function variants in type I IFN immunity genes are not associated with severe COVID-19. *J. Clin. Invest.* (2021) https://doi.org/10.1172/JCI147834.
13. 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).
14. Shelton, J. F. *et al*. Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity. *bioRxiv* (2020) https://doi.org/10.1101/2020.09.04.20188318.
15. Pairo-Castineira, E. *et al*. Genetic mechanisms of critical illness in Covid-19. *Nature* (2020) https://doi.org/10.1038/s41586-020-03065-y.
16. Roberts, G. H. L. *et al*. AncestryDNA COVID-19 host genetic study identifies three novel loci. *bioRxiv* (2020) https://doi.org/10.1101/2020.10.06.20205864.
17. Kosmicki, J. A. *et al*. Genetic association analysis of SARS-CoV-2 infection in 455,838 UK Biobank participants. *bioRxiv* (2020) https://doi.org/10.1101/2020.10.28.20221804.
18. 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).
19. Couturier, N. *et al*. Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility. *Brain* **134**, 693–703 (2011).
20. Li, Z. *et al*. Two rare disease-associated Tyk2 variants are catalytically impaired but signaling competent. *J. Immunol.* **190**, 2335–2344 (2013).
21. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
22. Hao, K. *et al*. Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8**, e1003029 (2012).
23. 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).
24. Manichaikul, A. *et al*. Genome-wide association study of subclinical interstitial lung disease in MESA. *Respir. Res.* **18**, 97 (2017).
25. Stefansson, H. *et al*. A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
26. 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).
27. Ghoussaini, M. *et al*. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* (2020) https://doi.org/10.1093/nar/gkaa840.
28. Xiao, G. *et al*. CXCL16/CXCR6 chemokine signaling mediates breast cancer progression by pERK1/2-dependent mechanisms. *Oncotarget* **6**, 14165–14178 (2015).
29. Wei, Q. *et al*. LZTFL1 suppresses lung tumorigenesis by maintaining differentiation of lung epithelial cells. *Oncogene* **35**, 2655–2663 (2016).
30. Vuille-dit-Bille, R. N. *et al*. Human intestine luminal ACE2 and amino acid transporter expression increased by ACE-inhibitors. *Amino Acids* **47**, 693–705 (2015).
31. Finucane, H. K. *et al*. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
32. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
33. Eyre, S. *et al*. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–1340 (2012).
34. Tsoi, L. C. *et al*. Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. *Nat. Commun.* **8**, 15382 (2017).
35. Langefeld, C. D. *et al*. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).
36. Kichaev, G. *et al*. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
37. 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).
38. 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).
39. 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).
40. 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) https://doi.org/10.1101/2020.04.22.20072124.

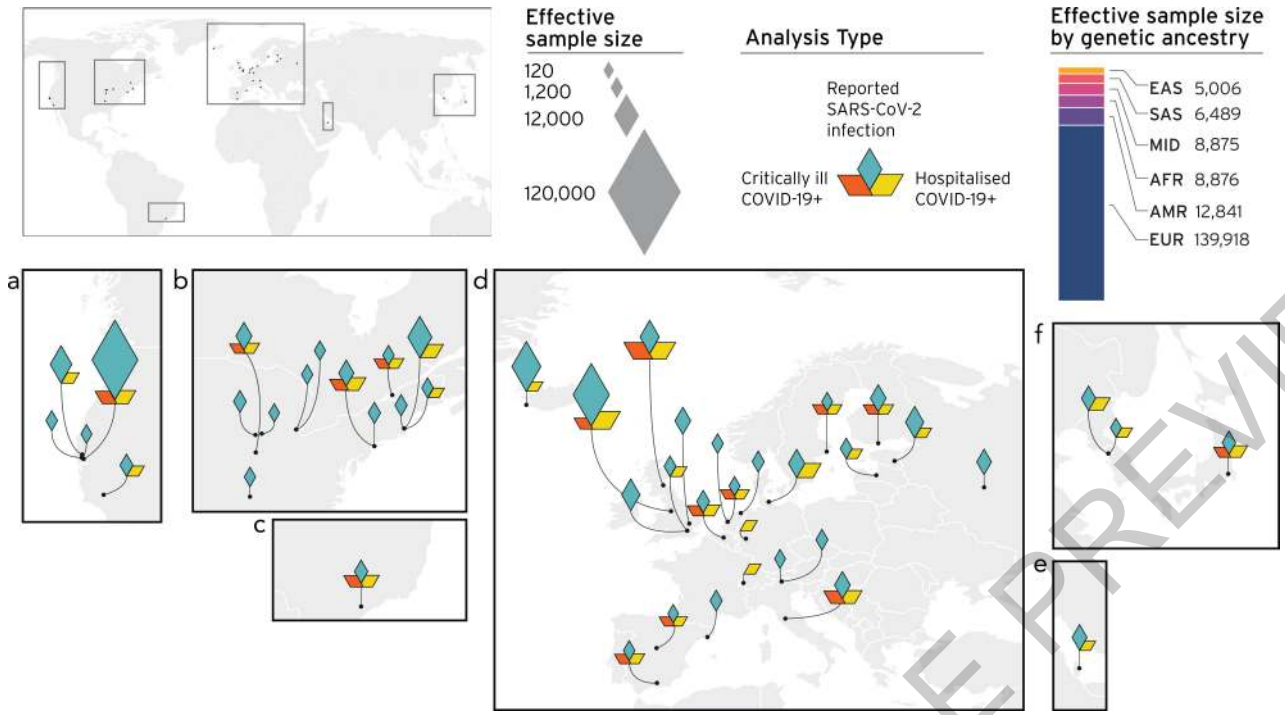


Fig. 1 | Geographical overview of the contributing studies to the COVID-19 HGI and composition by major ancestry groups. Populations are defined as Middle Eastern (MID), South Asian (SAS), East Asian (EAS), African (AFR), Admixed American (AMR), European (EUR).

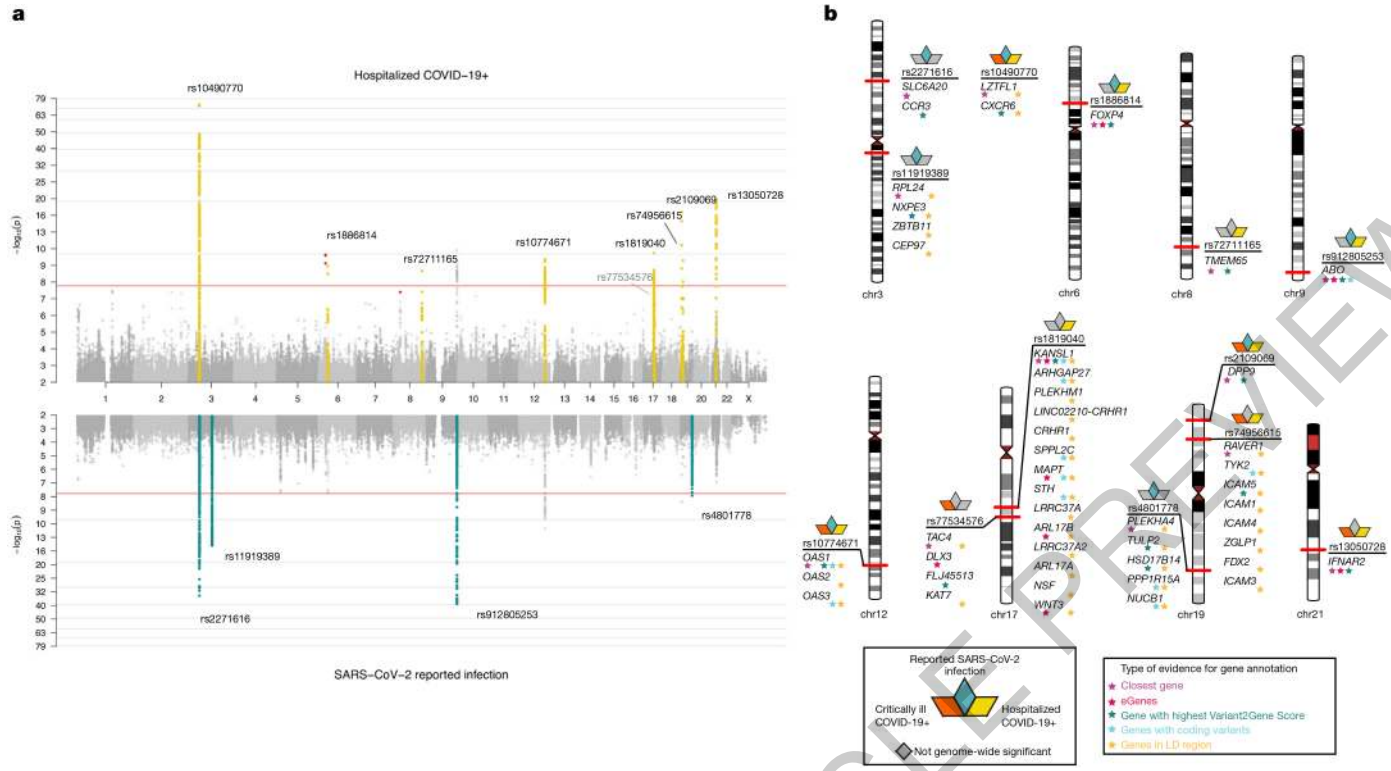


Fig. 2 | Genome-wide association results for COVID-19. a. Top panel shows results of genome-wide association study of hospitalized COVID-19 (n=13,641 cases and =2,070,709 controls), and bottom panel the results of reported SARS-CoV-2 infection (n=49,562 cases and n=1,770,206 controls). Loci highlighted in yellow (top panel) represent regions associated with severity of COVID-19 manifestation i.e. increasing odds for more severe COVID-19 phenotypes. Loci highlighted in green (bottom panel) are regions associated with susceptibility to SARS-CoV-2 infection, i.e. the effect is the same across mild and severe COVID-19 phenotypes. We highlight in red genome-wide significant variants that had high heterogeneity across contributing studies, and were therefore excluded from the list of loci found. **b**. Results of gene prioritization using different evidence measures of gene annotation. Genes in linkage disequilibrium (LD) region, genes with coding variants and eGenes (fine-mapped cis-eQTL variant PIP > 0.1 in GTEx Lung) are annotated if in LD with a COVID-19 lead variant ($r^2 > 0.6$). V2G: Highest gene prioritized by OpenTargetGenetics' V2G score.

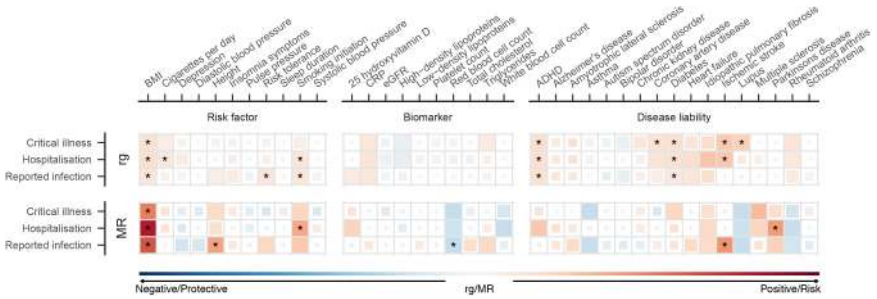


Fig. 3 | Genetic correlations and Mendelian randomization causal estimates between 38 traits and COVID-19 critical illness, hospitalization and SARS-CoV-2 reported infection. Larger squares correspond to more significant P-values, with genetic correlations or MR causal estimates significantly different 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. Two-sided P-values were calculated using LDSC for genetic correlations and Inverse variance weighted analysis for MR.

Methods

Contributing studies

All subjects were recruited following protocols approved by local Institutional Review Boards (IRBs); this information is collected in Supplementary Table 1 for all 46 studies. All protocols followed local ethics recommendations and informed consent was obtained when required. Information about sample numbers, sex and age from for each contributing study is given in Supplementary Table 1. 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. 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. The effective sample sizes for each ancestry group shown in Figure 1 were calculated for display using the formula: ((4 × N_cases × N_controls) / (N_cases + N_controls)). Details of contributing research groups are described in Supplementary Table 1.

Phenotype Definitions

COVID-19 disease status (critical illness, hospitalization status) was assessed following the Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia[41]. 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)[42] on chromosomes 1-22 and X. The recommended analysis tool was SAIGE, but studies also used other software such as PLINK[43]. 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[44] 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 on variants that were present in at least 2/3 of studies contributing to the phenotype analysis. 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.

We report 13 meta-analysis variants that pass genome-wide significance threshold after adjusting the threshold for multiple traits tested ($P < 5 \times 10^{-8}/3$). We report the unadjusted *P*-values for each variant. We tested for heterogeneity between estimates from contributing studies using Cochran's Q test[45,46]. This is calculated for each variant as the weighted sum of squared differences between the 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. Two loci reached genome-wide significance but were excluded from Supplementary Table 2 significant results due to heterogeneity between estimates from contributing studies and missingness between studies at chr6:31057940-31380334 and chr7:54671568-54759789; however these regions are not excluded from the corresponding summary statistics in data release 5. For each of the lead variants reported in Supplementary Table 2, we aimed to find loci specific to susceptibility or severity by testing 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[45,46], 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. A significant *P*-value <0.004 (0.05/13 loci) for multiple tests) indicates that the effect sizes for a particular variant are significantly different in the two analyses (Supplementary Table 2). For the 9 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 4). 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 4). 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.

PC projection

To project every GWAS participant into the same PC space, we used pre-computed PC loadings and reference allele frequencies. For reference, we used unrelated samples from the 1000 Genomes Project and the Human Genome Diversity Project (HGDP) and computed PC loadings and allele frequencies for the 117,221 SNPs that are i) available in every cohort, ii) MAF > 0.1% in the reference, and iii) LD pruned (r2 < 0.8; 500kb window). We then asked each cohort to project their samples using our automated script provided at https://github.com/covid19-hg/. It internally uses PLINK2[47] --score function with variance-standardize option and reference allele frequencies (--read-freq); so that each cohort-specific genotype/dosage matrix is mean-centered and variance-standardized with regards to reference allele frequencies, not cohort-specific allele frequencies. We further normalized the projected PC scores by dividing by a square root of the number of variants used for projection to account for a subtle difference due to missing variants.

Gene prioritization

To prioritize candidate causal genes reported in full in Supplementary Table 2, 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 systematic approach to nominate 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[44] 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[48] 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 5). We retrieved fine-mapped variants from the GTEx v8[21] (https://www.finucanelab.org/) and eQTL catalogue[49]. In addition, we looked up significant associations in the Lung eQTL Consortium[22] ($n = 1{,}103$) to further support findings in lung with a larger sample size (Supplementary Table 7). 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)[27]. 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[50]., 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)[27,51].

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)[52], and from UK Biobank. Summary statistics for UK Biobank traits were extracted from SAIGE[42] 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). We report PheWas results for phenotypes for which the lead variants were in high LD ($r^2 > 0.8$) with the 13 genome-wide significant lead variants from our main COVID-19 meta-analysis (Supplementary Table 6). This conservative approach allowed spurious signals primarily driven by proximity rather than actual colocalization to be removed (see Methods).

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[53] 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, the HLA region on chromosome 6, 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 8.

Partitioned heritability

We used partitioned LD score regression[54] to partition COVID-19 SNP heritability in cell types in our European ancestries 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 10). 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[55] or traits reported to be associated with increased risk of COVID-19 mortality by OpenSafely[56]. 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[54] was also used to estimate genetic correlations between our COVID-19 meta-analysis phenotypes reported using European ancestries 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-CoV-2 infection and COVID-19 severity were compared using a z score method[57].

Mendelian Randomization

Two-sample Mendelian randomization was employed to evaluate the potential for 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[43]. We used a strict r^2 threshold of 0.001, a 10MB clumping window, and the European reference panel from the 1000 Genomes project[58] 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[59]. 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 10 includes the harmonized datasets used in the analyses.

Mendelian Randomization Pleiotropy residual sum and outlier (MR-PRESSO) Global test[60] 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[61] 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 ($I^2_{GX} > 90\%$;[62]). 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[63], 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[59].

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).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Summary statistics generated by COVID-19 HGI are available at https://www.covid19hg.org/results/r5/ and are available on GWAS Catalog (study code GCST011074). 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, but results for 13 loci lead variants are reported in Supplementary Table 3. Individual level data can be requested directly from contributing studies, listed in Supplementary Table 1. We used publicly available data from GTEx (https://gtexportal.org/home/), the Neale lab (http://www.nealelab.is/uk-biobank/), Finucane lab (https://www.finucanelab.org), FinnGen Freeze 4 cohort (https://www.finngen.fi/en/access_results), and eQTL catalogue release 3 (http://www.ebi.ac.uk/eqtl/).

Code availability

The code for summary statistics liftover, projection PCA pipeline including precomputed loadings and meta-analysis are available at https://github.com/covid19-hg/ and the code for Mendelian randomization and genetic correlation pipeline at https://github.com/marcoralab/MRcovid.

41. Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (Trial Version 7). *Chin. Med. J.* **133**, 1087–1095 (2020).
42. Zhou, W. *et al*. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* (2018) https://doi.org/10.1038/s41588-018-0184-y.
43. 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).
44. Karczewski, K. J. *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
45. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet*. **14**, 379–389 (2013).
46. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* **10**, 101–129 (1954).
47. Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
48. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol*. **17**, 122 (2016).
49. 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) https://doi.org/10.1101/2020.01.29.924266.
50. Sun, B. B. *et al*. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
51. 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) https://doi.org/10.1101/2020.09.16.299271.
52. 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).
53. Bulik-Sullivan, B. K. *et al*. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet*. **47**, 291–295 (2015).
54. Finucane, H. K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet*. **47**, 1228–1235 (2015).
55. CDC. COVID-19 and Your Health. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html (2021).
56. Williamson, E. J. *et al*. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020).

57. Zhou, T. *et al*. Educational attainment and drinking behaviors: Mendelian randomization study in UK Biobank. *Mol. Psychiatry* (2019) https://doi.org/10.1038/s41380-019-0596-9.
58. 1000 Genomes Project Consortium *et al*. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
59. Hemani, G. *et al*. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).
60. 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).
61. 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).
62. 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).
63. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol*. **44**, 313–329 (2020).

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. Acknowledgements specific to contributing studies are provided in Supplementary Table 13. We thank G. Butler-Laporte, G. Wojcik, M.-G. Hollm-Delgado, C. Willer and G. Davey Smith for their extensive feedback and discussion.

Author contributions Author contributions are provided within the Authorship list.

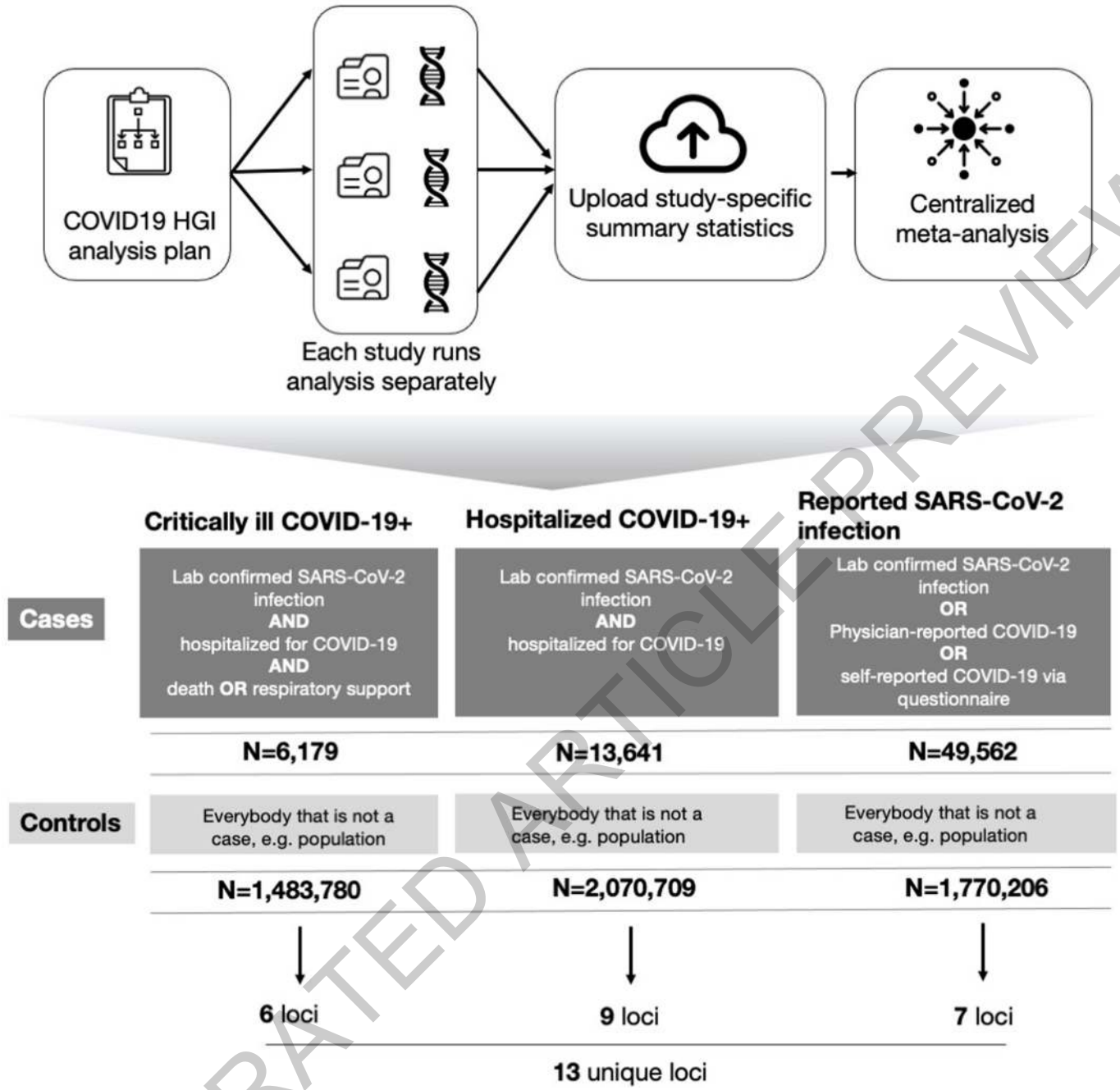
Competing interests A full list of competing interests is supplied as Supplementary Table 13.

Additional information

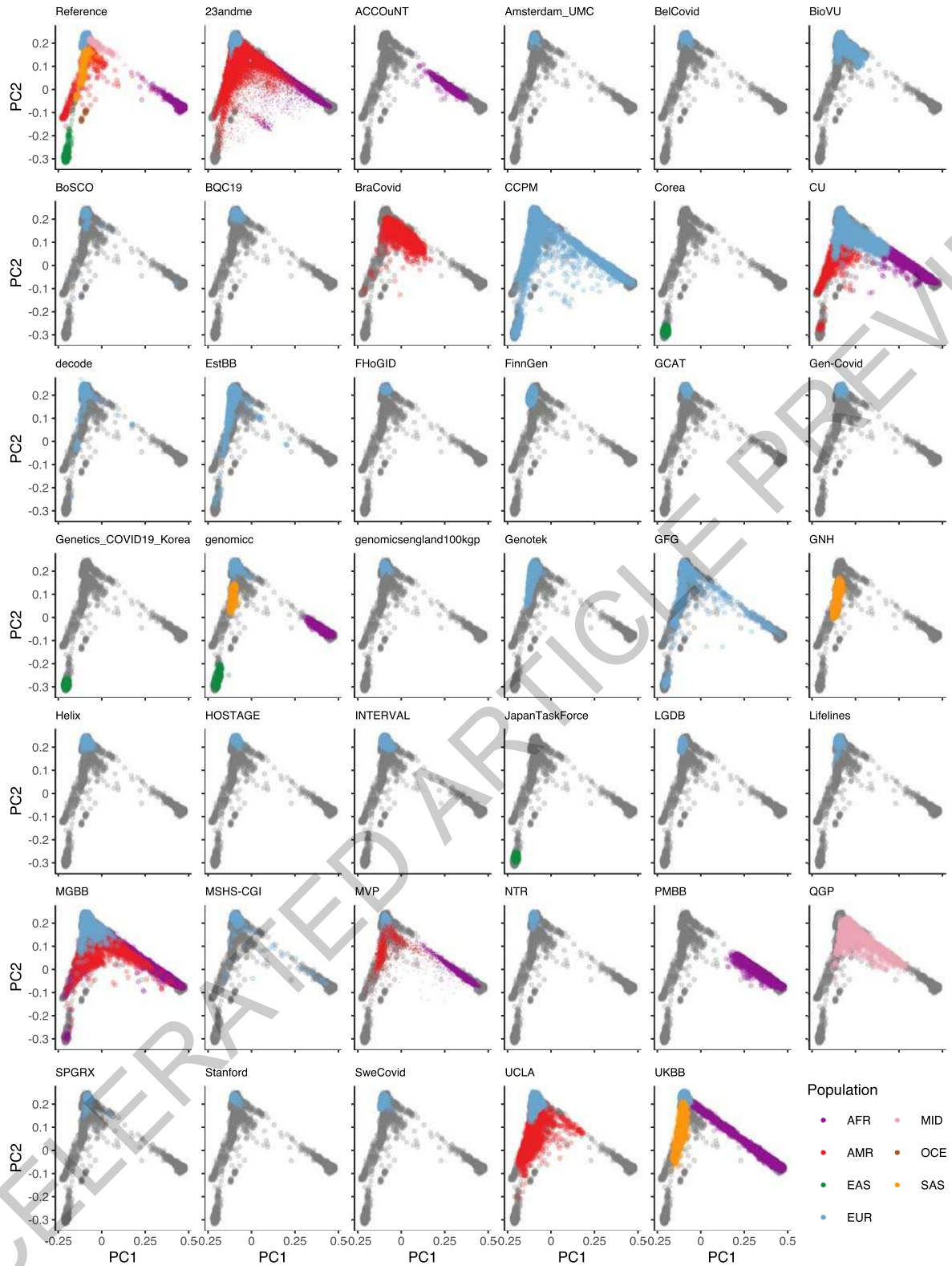
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03767-x.

Correspondence and requests for materials should be addressed to B.M.N, M.D. or A.G.

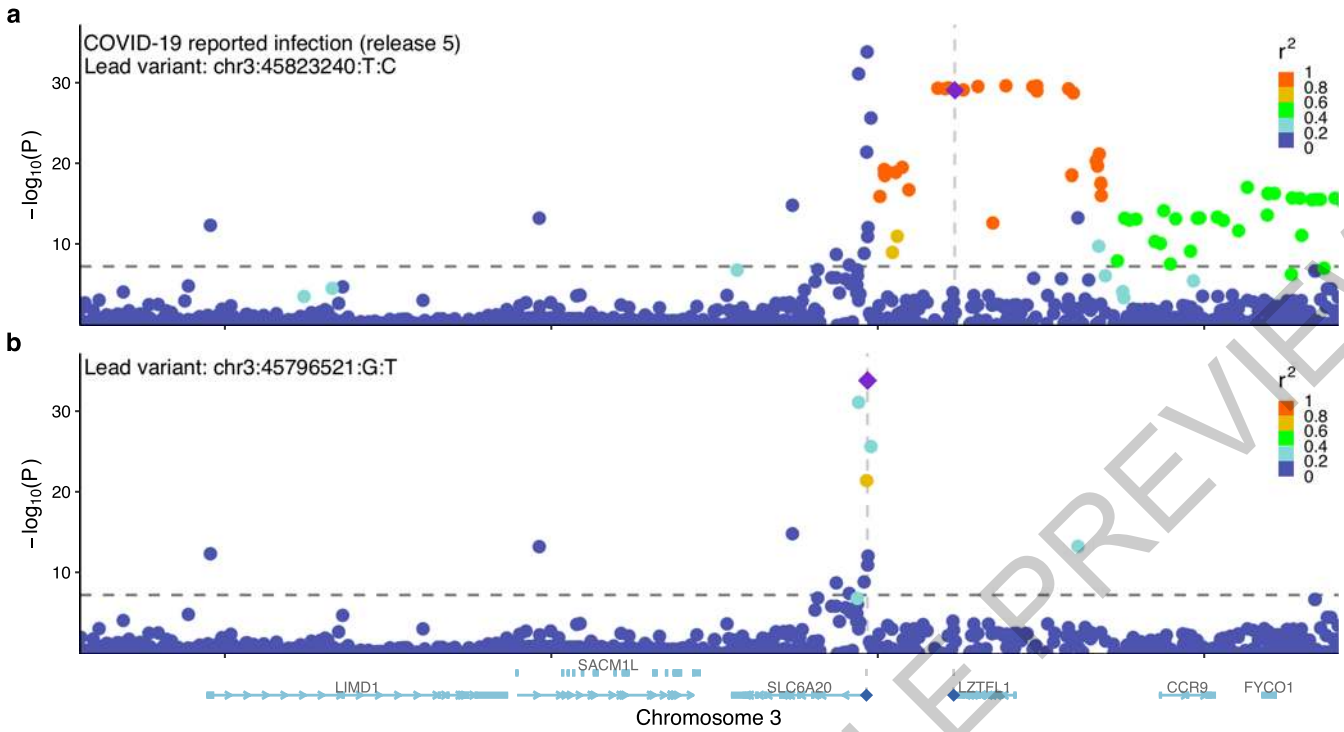
Peer review information *Nature* thanks Samira Asgari, Paul McLaren and Neneh Sallah for their contribution to the peer review of this work. Peer reviewer reports are available.



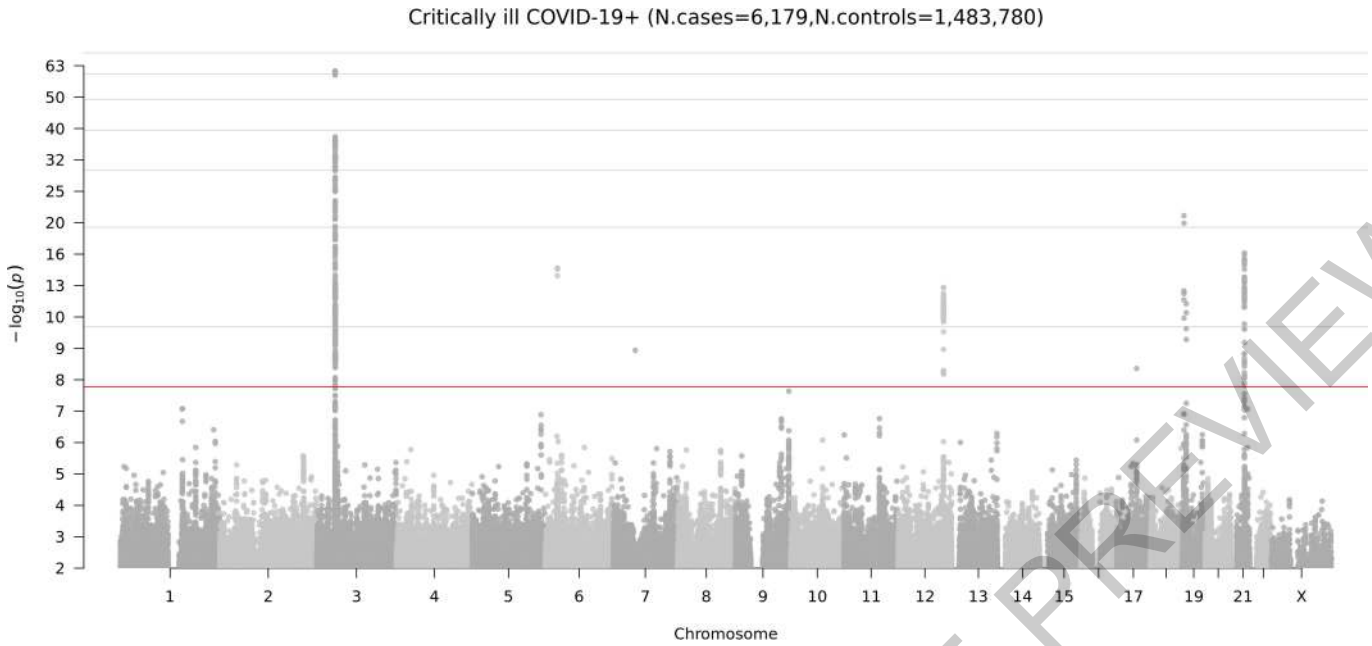
Extended Data Fig. 1 | Analytical summary of the COVID-19 HGI meta-analysis. Using the analytical plan set by the COVID-19 HGI, each individual study runs their analyses and uploads the results to the Initiative, who then runs the meta-analysis. There are three main analyses that each study can contribute summary statistics to; critically ill COVID-19, hospitalized COVID-19 and reported SARS-CoV-2 infection. The phenotypic criteria used to define cases are listed in the dark grey boxes, along with the numbers of cases (N) included in the final all ancestries meta-analysis. Controls were defined in the same way across all three analyses; as everybody that is not a case e.g. population controls (light grey box). Sensitivity analyses, not reported in this Figure, also used mild/asymptomatic COVID-19 cases as controls. Sample number (N) of controls differed between the analyses due to the difference in number of studies contributing data to these.



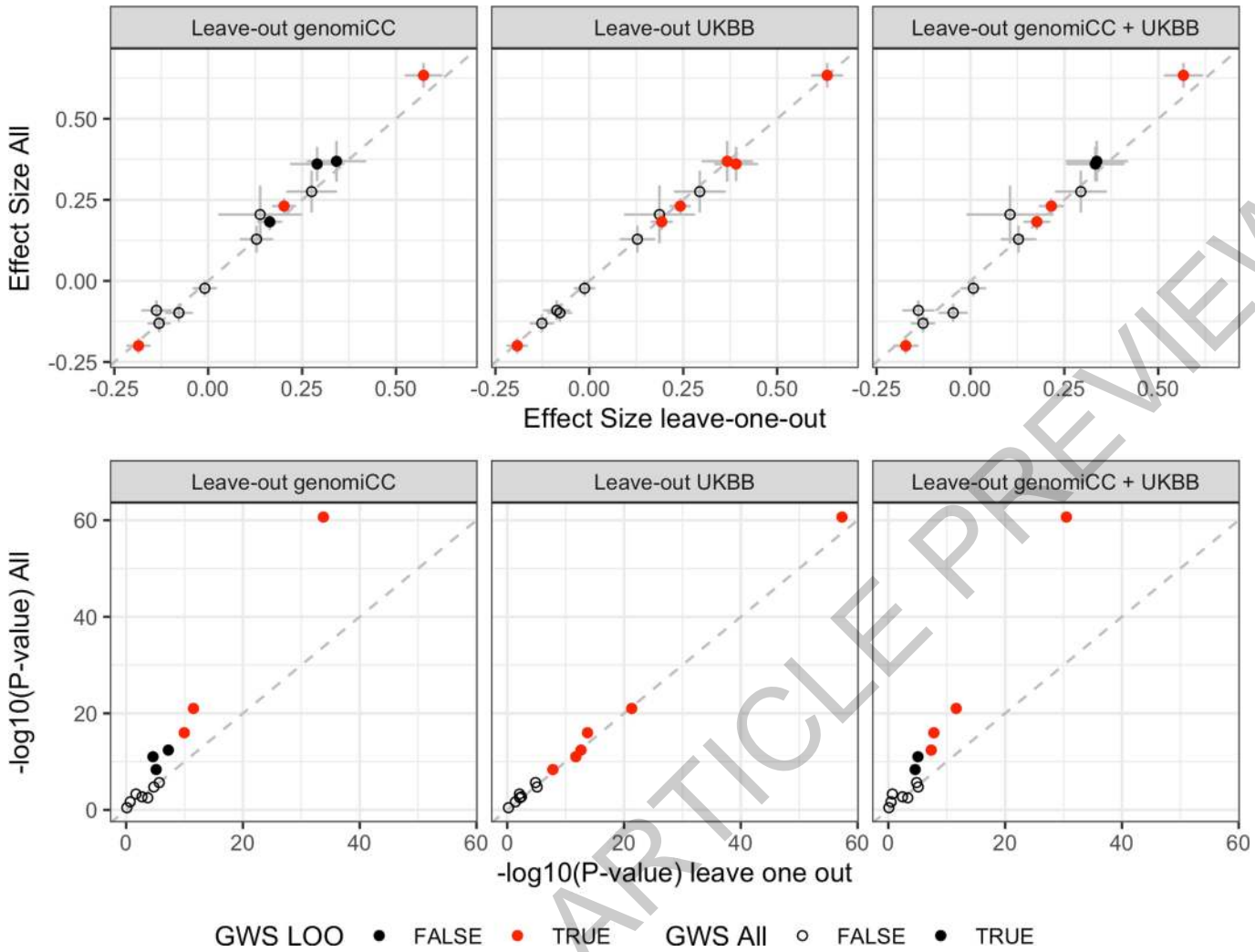
Extended Data Fig. 2 | Projection of contributing studies samples into the same PC space. We asked participating studies to perform PC projection using the 1000 Genomes Project and Human Genome Diversity Project as a reference, with a common set of variants. For each panel (except for the reference), colored points correspond to contributed samples from each cohort, whereas gray points correspond to 1000 Genomes reference samples. Color represents a genetic population that each cohort specified. Since 23andme, genomicsengland100kgp, and MVP only submitted PCA images, we overlaid their submitted transparent images using the same coordinates, instead of directly plotting them.



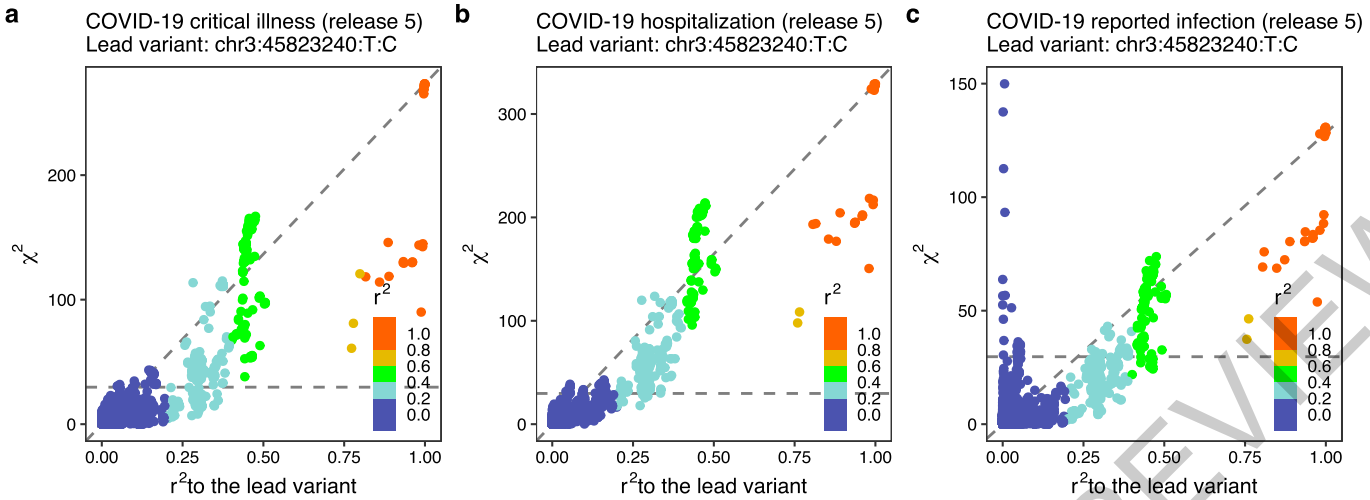
Extended Data Fig. 3 | Locuszoom plots of the 3p21.31 region for reported infection. a. A standard plot without exclusion. Here, the severity lead variant rs10490770 (chr3:45823240:T:C) is shown as a lead variant. b. Additional independent susceptibility signal(s) after excluding variants with $r^2 > 0.05$ with rs10490770. The susceptibility lead variant rs2271616 (chr3:45796521:G:T) is highlighted.



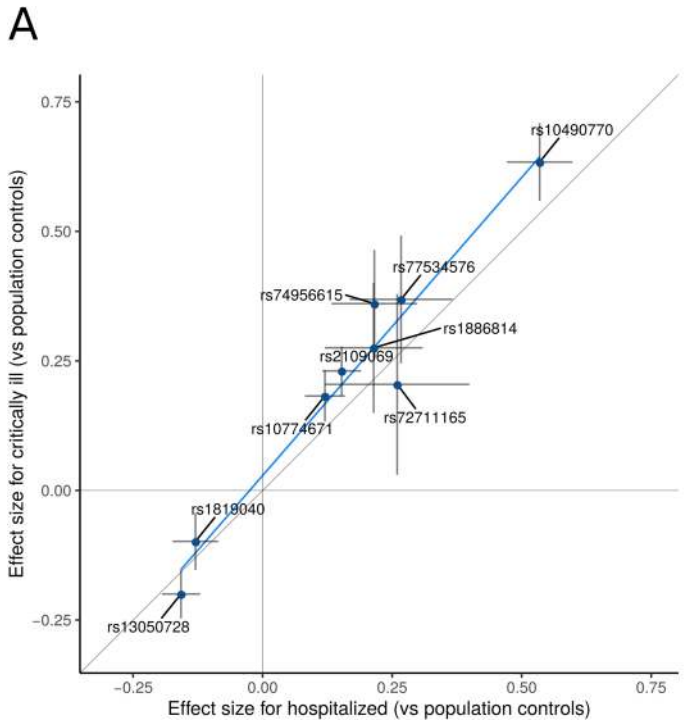
Extended Data Fig. 4 | Genome-wide meta-analysis association results for critical illness due to COVID-19. The locus on chromosome 6 is the HLA locus, which was removed from the list of reported loci in Table 1 due to the high heterogeneity in effect size estimated between studies included in the analysis. The locus on chromosome 7 was also not reported in Table 1 due to missingness across studies, i.e. the high number of studies in the meta-analysis that did not report summary statistics for this region. There are two association peaks on chromosome 19.

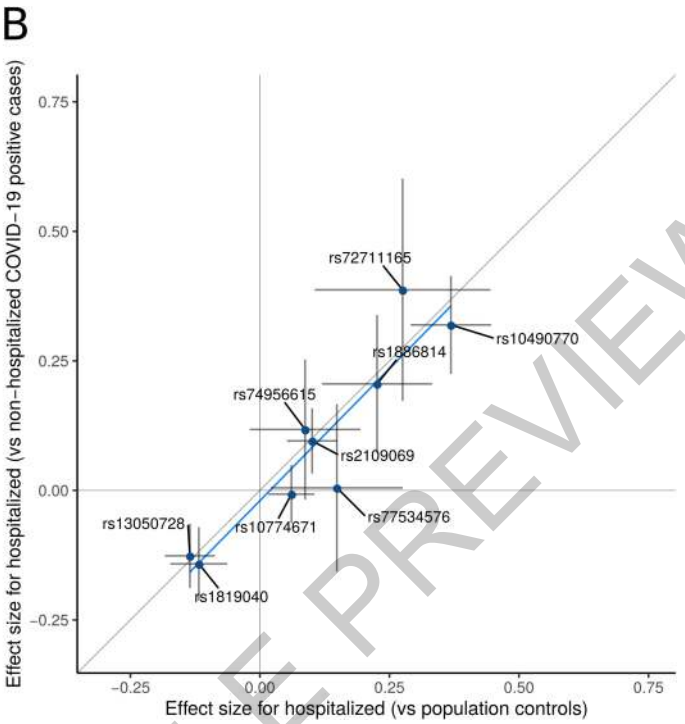


Extended Data Fig. 5 | Sensitivity analyses for overlapping controls in genomiCC and UK Biobank. Comparison of the beta effect sizes (top panel) and unadjusted *P*-values (bottom panel) of the 13 lead variants, using data from the COVID-19 critical illness meta-analysis in all the cohorts (y-axis) to leaving out genomiCC (case = 4,354; control = 1,474,655; total n = 1,479,009), leaving out UK Biobank (UKBB, case = 5,870; control = 1,155,203; total n = 1,161,073) and leaving out genomiCC + UKBB (case = 4,045; control = 1,146,078; total n = 1,150,123), respectively (x-axis). Top panel dots and grey bars represent the beta effect size estimates +/- standard error from the corresponding GWAS meta-analysis, bottom panel dots represent two-sided *P*-values from the corresponding GWAS meta-analysis. Filled dots indicate variants that were genome-wide 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 genome-wide significant in leave-one-out analysis for genomiCC, UKBB or genomiCC + UKBB.

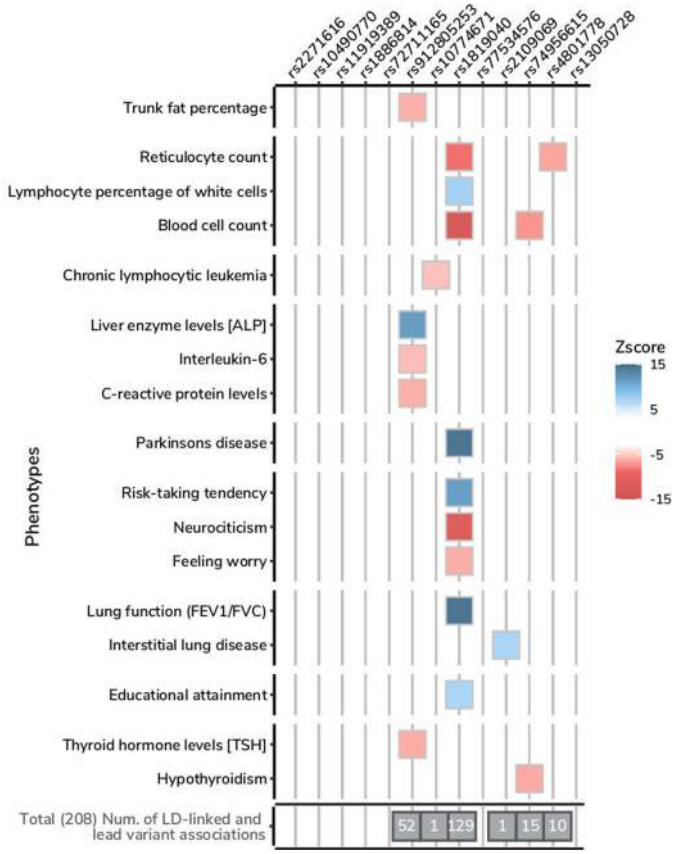


Extended Data Fig. 6 | Comparison of chi-squared statistics vs r^2 values to the lead variant in the 3p21.31 region. For **a**. critical illness **b**. hospitalization, and **c**. reported infection. The left blue peak in panel **c**, which is uncorrelated with the lead variants in the region, indicates that there are independent signals.

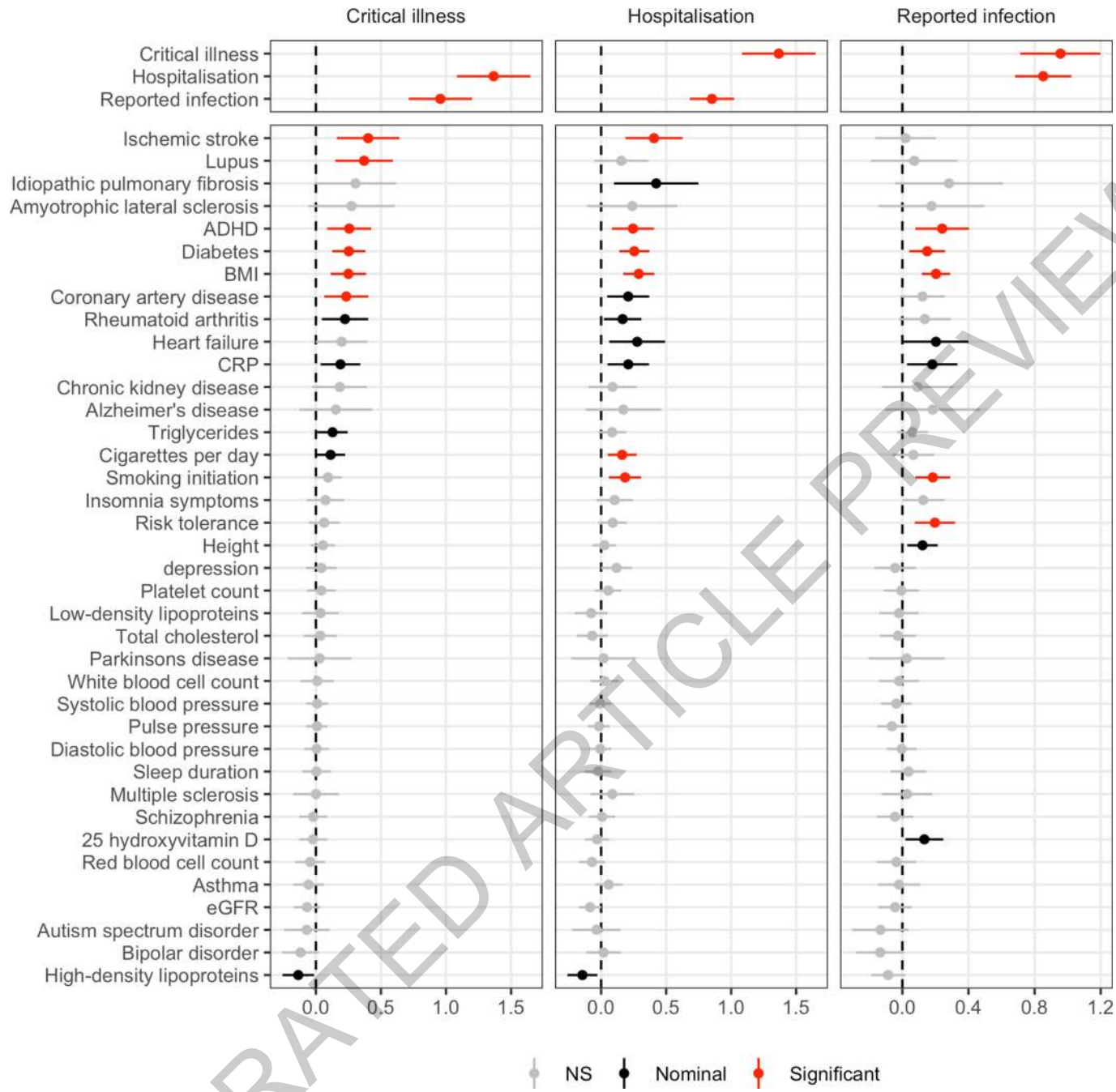




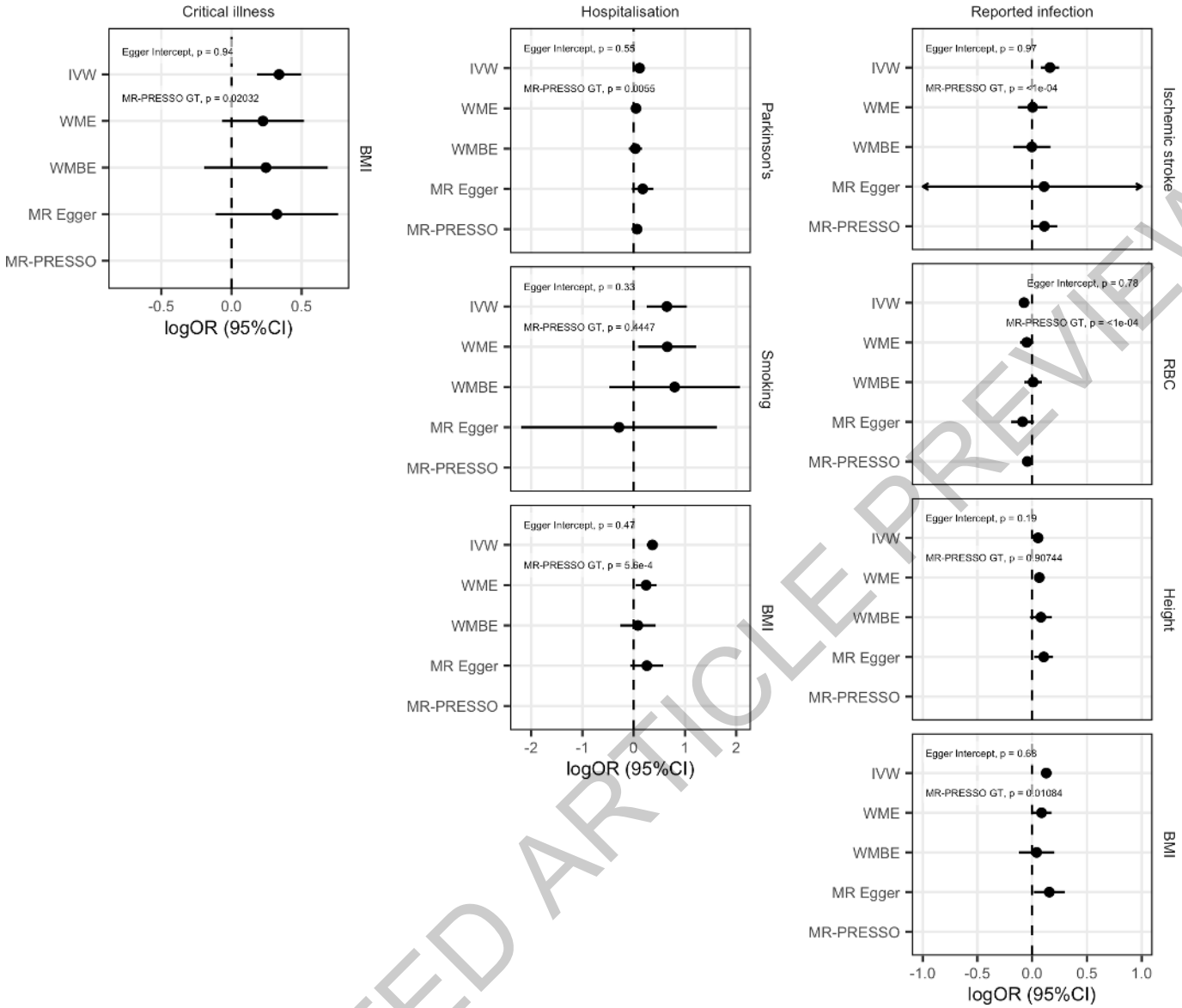
Extended Data Fig. 7 | Comparison of lead variant effect sizes between pairs of COVID-19 meta-analyses. Comparison of effect sizes for the nine variants associated with severity of COVID-19 disease. A. Comparing hospitalized COVID-19 cases vs population controls (x-axis, n=10,428 cases and n=1,483,270 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). Sample sizes for hospitalized COVID-19 cases vs population controls differ between panels A and B due to differences in the sampling of studies selected for the analysis. This selection included all studies that were able to contribute data to the respective analysis that the data were compared to (on the y-axis) in each panel. Dots represent the effect size beta estimates, bars represent the 95% confidence interval of the estimates. Effect size estimates and *P*-values for heterogeneity test (Cochran's Q, two-tailed test) are reported in Supplementary Table 3.



Extended Data Fig. 8 | PheWas for genome-wide significant lead variants. Selected phenotypes associated with genome-wide significant COVID-19 variants (see Supplementary Table 6 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.



Extended Data Fig. 9 | Genetic correlation with COVID-19 phenotypes. Each column shows genetic correlation results for the three COVID-19 phenotypes (European ancestry analyses only): critical illness, hospitalization and reported infection. The traits the genetic correlation is run against are listed on the left. Significant correlations (FDR<0.05) are shown with their 95% confidence intervals in red, nominally significant (P<0.05) in black and non-significant in grey. Two-sided P-values were calculated using LDSC for genetic correlations and exact estimates, unadjusted standard errors and two-sided P-values are available in Supplementary Table 11.



Extended Data Fig. 10 | Mendelian Randomization sensitivity analyses. Genetic correlations and Forest plots displaying 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% . Two-sided *P*-values were estimated using Inverse variance weighted analysis (IVW), Weighted median estimator (WME), weighted mode based estimator (WMBE), and Mendelian Randomization Pleiotropy RESidual Sum and Outlier (MR-PRESSO). RBC: Red blood cell count.

nature research

Corresponding author(s): Andrea Ganna

Last updated by author(s): Jun 1, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No code was used to collect data in the study.
Data analysis	Each individual study that contributed genetic-phenotype association summary statistics to the consortium carried out their association analyses independently of the consortium (study-specific information outlined in Supplementary Table 1). However, the consortium did release phenotyping and analysis guidelines as a recommendation (https://www.covid19hg.org/). For quality control of genotype data we recommended using the Ricopili pipeline (PMID: 31393554). For genotype phasing and imputation we recommended the TopMed Imputation Server (PMID: 27571263) or Michigan Imputation Server (PMID: 27571263). For genome-wide association study (GWAS), we recommended SAIGE (PMID: 30104761), but some studies used PLINK (PMID: 17701901). Each study then submitted their GWAS summary statistics to the consortium for meta-analysis. LD score regression v 1.0.1 [PMID: 25642630] was used for heritability and partitioned heritability analyses. Variants for Mendelian randomization instruments were selected using PLINK version 1.90b6.18 (PMID: 17701901). Exposure and outcome datasets were harmonized, and MR statistical analysis conducted using R version 4.0.3. with the R-package TwoSampleMR version 0.5.5 (PMID: 29846171) (which included Fixed-effects IVW analysis (PMID: 24114802), weighted median estimator (WME) (PMID: 27061298), weighted mode based estimator (WMBE) and MR Egger regression (PMID: 26050253)) and additionally MR-PRESSO version 1.0 (PMID: 29686387). Code availability statement: The code for summary statistics liftover, projection PCA pipeline including precomputed loadings and meta-analysis are available at https://github.com/covid19-hg/ and the code for Mendelian randomization and genetic correlation pipeline at https://github.com/marcoralab/MRcovid.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability statement:
Summary statistics generated by COVID-19 HGI are available at https://www.covid19hg.org/results/r5/ and are available on GWAS Catalog (study code GCST011074). 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, but results for 13 loci lead variants are reported in Supplementary Table 3. Individual level data can be requested directly from contributing studies, listed in Supplementary Table 1. We used publicly available data from GTEx (https://gtexportal.org/home/), the Neale lab (http://www.nealelab.is/uk-biobank/), Finucane lab (https://www.finucanelab.org), FinnGen Freeze 4 cohort (https://www.finngen.fi/en/access_results), and eQTL catalogue release 3 (http://www.ebi.ac.uk/eqtl/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The consortium meta-analysed genome-wide association study (GWAS) summary statistics from any individual study that had included a minimum of n=50 cases and n=50 controls in their analysis. The cutoff at n=50 cases and n=50 controls was aimed at reducing noise to the meta-analysis, but also to be inclusive of studies that had not yet accumulated large numbers of COVID-19 patient data. No statistical calculation for adequate sample size was performed, but the results identifying multiple genomic regions at genome-wide significance threshold indicates adequate power for genetic discovery.
Data exclusions	Individual level phenotype and genotype data exclusions were performed by each individual study, following the consortium analysis plan recommendations (www.covid19hg.org). Possible reasons for sample exclusion included removing genetic ancestry outliers within a study (using principal components analysis), poor quality of genetic data or lack of phenotypic data for a sample. The consortium manually examined GWAS summary statistics data submitted by each study (for each submitted analysis separately), including sample size used for analysis, allele frequency check against gnomad reference panel, and distribution of test statistics. After meta-analysis, the results were checked for heterogeneity variant effects between contributing studies, and Table 1 excludes two genome-wide significant loci that were deemed to have extremely heterogeneous effects, but these variants are reported in the released consortium summary statistics (with heterogeneity test values).
Replication	No replication was performed. The consortium meta-analysed GWAS summary statistics, bringing together as many studies as possible to achieve the largest possible sample size and statistical power for association. this meant that the consortium included most large studies of COVID-19 host genetics that have been performed to date, so it was not possible to perform replication analyses in external cohorts. Therefore we performed manual checks on each study contributing summary statistics before entering them into the meta-analysis. In addition, after meta-analysis, we performed a check for heterogeneity between variant association estimates across studies contributing data. This allowed us to better understand whether the variant effects differed much between individual studies.
Randomization	No randomization was performed because there was no allocation of samples to experimental groups.
Blinding	Blinding was not relevant to the study. The case status and severity of symptoms was evaluated for each sample by investigators from each study respectively. The consortium recommended using covariates to control for confounding: age + age2 + sex + age*sex + 20 principal components (obtained using genetic data) + study specific covariates (if any). The consortium meta-analysed summary statistics from these case/control studies, not individual level data. Details of which variables each study used and how the calculated PCs for their analysis are available in Supplementary Table 1.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Summary statistics from 46 independent studies were included in consortium meta-analyses. Mean age of cases across studies was 55.3 years. The effective sample size for genetic ancestry populations was: n=11,598 Middle Eastern; n=28,918 South Asian; 43,332 East Asian; 48,714 African; 70,902 Ad-mixed American; 738,538 European. Population characteristics regarding age, sex and exact case and control sample numbers for each contributing study are given in Supplementary Table 1.
Recruitment	The consortium pre-defined phenotype criteria for cases and controls, but the specific recruitment was carried out independently by each contributing study. COVID-19 disease status (critical illness, hospitalization status) was assessed following the Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (PMID: 32358325). 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.
Ethics oversight	Ethical statements for each contributing study are given in Supplementary Table 1.

Note that full information on the approval of the study protocol must also be provided in the manuscript.