Accelerated Article Preview

Whole genome sequencing reveals host factors underlying critical Covid-19

Received: 2 September 2021

Accepted: 23 February 2022

Accelerated Article Preview

Published online: 07 March 2022

Cite this article as: Kousathanas, A. et al. Whole genome sequencing reveals host factors underlying critical Covid-19. *Nature* <https://doi.org/10.1038/s41586-022-04576-6> (2022).

Athanasios Kousathanas, Erola Pairo-Castineira, Konrad Rawlik, Alex Stuckey, Christopher A. Odhams, Susan Walker, Clark D. Russell, Tomas Malinauskas, Yang Wu, Jonathan Millar, Xia Shen, Katherine S. Elliott, Fiona Griffiths, Wilna Oosthuyzen, Kirstie Morrice, Sean Keating, Bo Wang, Daniel Rhodes, Lucija Klaric, Marie Zechner, Nick Parkinson, Afshan Siddiq, Peter Goddard, Sally Donovan, David Maslove, Alistair Nichol, Malcolm G. Semple, Tala Zainy, Fiona Maleady-Crowe, Linda Todd, Shahla Salehi, Julian Knight, Greg Elgar, Georgia Chan, Prabhu Arumugam, Christine Patch, Augusto Rendon, David Bentley, Clare Kingsley, Jack A. Kosmicki, Julie E. Horowitz, Aris Baras, Goncalo R. Abecasis, Manuel A. R. Ferreira, Anne Justice, Tooraj Mirshahi, Matthew Oetjens, Daniel J. Rader, Marylyn D. Ritchie, Anurag Verma, Tom A. Fowler, Manu Shankar-Hari, Charlotte Summers, Charles Hinds, Peter Horby, Lowell Ling, Danny McAuley, Hugh Montgomery, Peter J. M. Openshaw, Paul Elliott, Timothy Walsh, Albert Tenesa, GenOMICC Investigators, 23andMe, Covid-19 Human Genetics Initiative, Angie Fawkes, Lee Murphy, Kathy Rowan, Chris P. Ponting, Veronique Vitart, James F. Wilson, Jian Yang, Andrew D. Bretherick, Richard H. Scott, Sara Clohisey Hendry, Loukas Moutsianas, Andy Law, Mark J. Caulfield & J. Kenneth Baillie

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

Whole genome sequencing reveals host factors underlying critical Covid-19

<https://doi.org/10.1038/s41586-022-04576-6>

Received: 2 September 2021

Accepted: 23 February 2022

Published online: 07 March 2022

Athanasios Kousathanas^{1,36}, Erola Pairo-Castineira^{2,3,36}, Konrad Rawlik², Alex Stuckey¹, Christopher A. Odhams¹, Susan Walker¹, Clark D. Russell^{2,4}, Tomas Malinauskas⁵, Yang Wu⁶, Jonathan Millar², Xia Shen^{7,8}, Katherine S. Elliott⁵, Fiona Griffiths², Wilna Oosthuyzen², Kirstie Morrice⁹, Sean Keating¹⁰, Bo Wang², Daniel Rhodes¹, Lucija Klaric³, Marie Zechner², Nick Parkinson², Afshan Siddiq¹, Peter Goddard¹, Sally Donovan¹, David Maslove¹¹, Alistair Nichol¹², Malcolm G. Semple^{13,14}, Tala Zainy¹, Fiona Maleady-Crowe¹, Linda Todd¹, Shahla Salehi¹, Julian Knight⁵, Greg Elgar¹, Georgia Chan¹, Prabhu Arumugam¹, Christine Patch¹, Augusto Rendon¹, David Bentley¹⁵, Clare Kingsley¹⁵, Jack A. Kosmicki¹⁶, Julie E. Horowitz¹⁶, Aris Baras¹⁶, Goncalo R. Abecasis¹⁶, Manuel A. R. Ferreira¹⁶, Anne Justice¹⁷, Tooraj Mirshahi¹⁷, Matthew Oetjens¹⁷, Daniel J. Rader¹⁸, Marylyn D. Ritchie¹⁸, Anurag Verma¹⁸, Tom A. Fowler^{1,19}, Manu Shankar-Hari²⁰, Charlotte Summers²¹, Charles Hinds²², Peter Horby²³, Lowell Ling²⁴, Danny McAuley^{25,26}, Hugh Montgomery²⁷, Peter J. M. Openshaw^{28,29}, Paul Elliott³⁰, Timothy Walsh¹⁰, Albert Tenesa^{2,3,8}, GenOMICC Investigators*, 23andMe*, Covid-19 Human Genetics Initiative*, Angie Fawkes⁹, Lee Murphy⁹, Kathy Rowan³¹, Chris P. Ponting³, Veronique Vitart³, James F. Wilson^{3,8}, Jian Yang^{32,33}, Andrew D. Bretherick³, Richard H. Scott^{1,34}, Sara Clohisey Hendry^{2,37}, Loukas Moutsianas^{1,37}, Andy Law^{2,37}, Mark J. Caulfield^{1,35,37}✉ & J. Kenneth Baillie^{2,3,4,10,37}✉

Critical Covid-19 is caused by immune-mediated inflammatory lung injury. Host genetic variation influences the development of illness requiring critical care¹ or hospitalisation^{2–4} following SARS-CoV-2 infection. The GenOMICC (Genetics of Mortality in Critical Care) study enables the comparison of genomes from critically-ill cases with population controls in order to find underlying disease mechanisms. Here, we use whole genome sequencing in 7,491 critically-ill cases compared with 48,400 controls to discover and replicate 23 independent variants that significantly predispose to critical Covid-19. We identify 16 new independent associations, including variants within genes involved in interferon signalling (*IL10RB*, *PLSCR1*), leucocyte differentiation (*BCL11A*), and blood type antigen secretor status (*FUT2*). Using transcriptome-wide association and colocalisation to infer the effect of gene expression on disease severity, we find evidence implicating multiple genes, including reduced expression of a membrane flippase (*ATP11A*), and increased mucin expression (*MUC1*), in critical disease. Mendelian randomisation provides evidence in support of causal roles for myeloid cell adhesion molecules (*SELE*, *ICAM5*, *CD209*) and coagulation factor *F8*, all of which are potentially druggable targets. Our results are broadly consistent with a multi-component model of Covid-19 pathophysiology, in which at least two distinct mechanisms can predispose to life-threatening disease: failure to control viral replication, or an enhanced tendency towards pulmonary inflammation and intravascular coagulation. We show that comparison between critically-ill cases and population controls is highly efficient for detection of therapeutically-relevant mechanisms of disease.

a relatively homogeneous clinical definition including patients with hypoxaemic respiratory failure⁵ with acute lung injury⁶, and excluding many patients with non-pulmonary clinical presentations⁷ who are known to have divergent responses to therapy⁸. In the UK, the critically-ill patient group is younger, less likely to have significant comorbidity, and more severely affected than a general hospitalised cohort⁵, characteristics which may amplify observed genetic effects.

In addition, since development of critical illness is in itself a key clinical endpoint for therapeutic trials⁸, using critical illness as a phenotype in genetic studies enables detection of directly therapeutically-relevant genetic effects¹.

Using microarray genotyping in 2,244 cases, we previously discovered that critical Covid-19 is associated with genetic variation in the host immune response to viral infection (*OAS1*, *IFNAR2*, *TYK2*) and the

inflammasome regulator *DPP9*¹. In collaboration with international groups, we extended these findings to include a variant near *TAC4* (rs77534576)³. Several variants have been associated with milder phenotypes, including the ABO blood type locus², a pleiotropic inversion in chr17q21.31⁹, and associations in 5 additional loci including the T lymphocyte-associated transcription factor, *FOXP4*³. An enrichment of rare loss-of-function variants in candidate interferon signalling genes has been reported⁴, but this has yet to be replicated at genome-wide significance thresholds^{10,11}.

We performed whole genome sequencing (WGS) in partnership with Genomics England, to improve resolution and deepen fine-mapping of significant signals to enhance the biological insights into critical Covid-19. Here, we present results from a cohort of 7,491 critically-ill patients from 224 intensive care units, compared with 48,400 controls, describing discovery and validation of 23 gene loci for susceptibility to critical Covid-19 (Extended Data Fig. 1).

GWAS analysis

Following quality control procedures, we used a logistic mixed model regression, implemented in SAIGE¹², to perform association analyses with unrelated individuals (critically-ill cases: $n = 7,491$, controls: $n = 48,400$ (100k cohort: $n = 46,770$ and mild Covid-19: $n = 1,630$) (Methods, Supplementary Table 2). 1,339 of these cases were included in the primary analysis for our previous report¹. Genome wide association studies (GWAS) were performed separately for genetic ancestry groups ($n_{\text{cases}}/n_{\text{controls}}$): European(EUR) 5,989/42,891; South Asian(SAS) 788/3,793; African(AFR) 440/1,350; East Asian(EAS) 274/366), and combined by inverse-variance weighted fixed effects meta-analysis using METAL (Methods). We established independence of signals using GCTA-cojo which we validated with conditional analysis using individual-level data with SAIGE (Methods, Supplementary Table 6). In order to reduce the risk of spurious associations arising from genotyping or pipeline errors, we required supporting evidence from variants in linkage disequilibrium for all genome-wide significant variants: observed z-scores for each variant were compared with imputed z-scores for the same variant, with discrepant values being excluded (see Methods, Supplementary Figure 2).

In population-specific analyses, we discovered 22 independent genome-wide significant associations in the EUR ancestry group (Fig. 1, Supplementary Figure 11 and Table 1) at a P -value threshold adjusted for multiple testing (2.2×10^{-8} ; Supplementary Table 5). In multi-ancestry meta-analysis, we identified an additional three independent genome-wide significant association signals (Fig. 1, Table 1).

To assess the sensitivity of our results to mismatches of demographic characteristics between cases and controls (Supplementary Figures 9,10), we performed an age-, sex- and body mass index (BMI)-matched case-control analysis (Supplementary Figures 18–21). Since there is a theoretical risk of mismatch between cases and 100k participants in risk factors for exposure (e.g. shielding behavior) or susceptibility to critical Covid-19 (e.g. immunosuppression), we performed a sensitivity analysis using only the mild cohort (see above; Supplementary Table 10). In both of these analyses, allele frequencies and directions of effect were concordant for all lead signals.

We inferred credible sets of variants using Bayesian fine-mapping with susieR¹³, by analysing the GWAS summaries of 17 regions of genomic length 3Mb that were flanking groups of lead signals. We obtained 22 independent credible sets of variants for EUR and an additional two from the transancestry meta-analysis with posterior inclusion probability > 0.95 (Extended Data Table 1 and Supplementary File: GWAS.xlsx). Fine mapping of the association signals revealed putative causal variants for both previously reported and novel association signals (See Supplementary Information, Extended Data Table 1). In 12 out of the 24 fine-mapped signals, the credible sets included five or fewer variants, and for 8 signals we detected variants with predicted

missense or worse consequence across each credible set (Extended Data Table 1). We were able to fine-map multiple independent signals at previously identified loci (Fig. 3, Extended Data Figures 2,29,4). For example, the signal in the 3p21.31 region, first reported by Ellinghaus *et al*², was fine-mapped into two independent associations, with the credible set for the first refined to a single variant in the 5' UTR region of *SLC6A20* (chr3:45796521:G:T, rs2271616, OR:1.29, 95%CI:1.21,1.37) and the second credible set including multiple variants in downstream and intronic regions of *LZTFL1* (Fig. 3). Among the novel signals, at 3q24 and 9p21.3 we detected missense variants that impact *PLSCR1* and *IFNA10* respectively (chr3:146517122:G:A, rs343320, p.His262Tyr, OR:1.24, 95%CI:1.15-1.33, CADD:22.6; chr9:21206606:C:G, rs28368148, p.Trp164Cys, OR:1.74, 95%CI:1.45-2.09, CADD:23.9). Both are predicted to be deleterious by the Combined Annotation Dependent Depletion (CADD) tool¹⁴. Structural predictions for these variants suggest functional effects (Extended Data Fig. 5). We assessed whether the main signals of this study were underlain by rarer variants with lower minor allele frequency (MAF) ($> 0.02\%$) than our GWAS default threshold ($> 0.5\%$), by including rarer variant summaries when fine-mapping, but no additional variants were added to the main credible sets (Supplementary Table 9).

Consistent with our expectation that genetic susceptibility plays a stronger role in younger patients, age-stratified analysis (< 60 vs. ≥ 60) in EUR revealed a signal in the 3p21.31 region with a significantly stronger effect in the younger age group (chr3:45801750:G:A, rs13071258, OR=3.34, 95%CI:2.98-3.75 vs. OR=2.1, 95%CI:1.88-2.34) that is in strong LD ($r^2=0.947$) with main GWAS signal indexed by rs73064425. Sex-specific analysis did not reveal significant effects (Supplementary Figure 17).

Replication

For replication, we performed a meta-analysis of summary statistics generously shared by 23andMe, Inc. and the Covid-19 Host Genetics Initiative (HGI) data freeze 6(B2). Since a previous analysis of GenOMICC¹ contributes a substantial part of the signal at each locus in HGIv6, and leave-one-out analyses were not available, we removed the signal from GenOMICC cases in HGIv6 using mathematical subtraction to ensure independence (Methods). Using LD clumping to find variants genotyped in both the discovery and replication studies, we required $P < 0.002$ (0.05/25) and concordant direction of effect (Table 1 and Supplementary Table 8) for replication. We interrogated two variants which failed replication in this set in a second GWAS meta-analysis of hospitalised Covid-19 patients from UKB, AncestryDNA, Penn Medicine Biobank, and Geisinger Health Systems, totaling 9937 hospitalized Covid-19 cases and 1,059,390 controls. This led to a further successful replicated finding, in *IFNA10* (Table 1).

We replicated 23 of the 25 significant associations identified in the population specific and/or multi-ancestry GWAS. One of the non-replicated signals (rs4424872) corresponds to a rare variant that may not be well represented in the replication datasets which are dominated by SNP genotyping data, but also had significant heterogeneity among ancestries. The second non-replicated signal is within the human leukocyte antigen (HLA) locus which has complex LD (see below).

HLA region

The lead variant in the HLA region, rs9271609, lies upstream of *HLA-DQA1* and *HLA-DRB1* genes. To investigate the contribution of specific HLA alleles to the observed association in the HLA region, we imputed HLA alleles at a four digit (two-field) level using HIBAG¹⁵. The only allele that reached genome-wide significance was *HLA-DRB1*04:01* (OR = 0.80, 95%CI:0.75 – 0.86, $P = 1.6 \times 10^{-10}$ in EUR), which has a stronger P -value than the lead SNP in the region (OR: 0.88, 95%CI:0.84 – 0.92, $P = 3.3 \times 10^{-9}$ in EUR) and is a better fit to

the data ($AIC_{DRB1*04:01} = 30241.34$, $AIC_{leadSNP} = 30252.93$) (Extended Data Fig. 6). *HLA-DRB1*04:01* has been previously reported to confer protection against severe disease in a small cohort of European ancestry¹⁶.

Gene burden testing

To assess the contribution of rare variants to critical illness, we performed gene-based analysis using SKAT-O as implemented in SAIGE-GENE¹⁷, using a subset of 12,982 individuals from our cohort (7,491 individuals with critical Covid-19 and 5,391 controls) for which the genome sequencing data were processed with the same alignment and variant calling pipeline. We tested the burden of rare ($MAF < 0.5\%$) variants considering the predicted variant consequence type (tested variant counts provided in sheet E of Supplementary File AVTsuppinfo.xlsx). We assessed burden using a strict definition for damaging variants (high-confidence loss-of-function (pLoF) variants as identified by LOFTEE¹⁸) and a lenient definition (pLoF plus missense variants with $CADD \geq 10$)¹⁴, but found no significant associations at a gene-wide significance level. Moreover, all individual rare variants included in the tests had P -values $> 10^{-5}$.

Consistent with other recent work¹¹, we did not find any significant gene burden test associations among the 13 genes previously reported from an interferon pathway-focused study⁴ (tests for all genes had P -value > 0.05 , Supplementary File AVTsuppinfo.xlsx), and we did not replicate the reported association^{19–21} in *TLR7* (EUR $P = 0.30$ for pLoF and $P = 0.075$ for missense variants).

Transcriptome-wide association study (TWAS)

In order to infer the effect of genetically-determined variation in gene expression on disease susceptibility, we performed a TWAS using gene expression data (GTExv8²²) for two disease-relevant tissues, lung and whole blood. We found significant associations between critical Covid-19 and predicted expression in lung (14) and blood (6; Supplementary Figure 23) and all-tissue meta-analysis (GTExv8, 51; Supplementary Figure 24). Expression signals for 16 genes significantly colocalised with susceptibility (Fig. 2). As the LD structure of the HLA is complex, we only assessed colocalisation for the significant association, *HLA-DRB1*. Although it was not significant in our TWAS analysis, eQTLs in the proximity of the association significantly colocalise with the GWAS signal for both blood and lung (both $PP_{H4} > 0.8$, Supplementary File: TWAS.xlsx).

We repeated the TWAS analysis using models of intron excision rate from GTExv8 to obtain splicing TWAS, revealing significant signals in lung (16 genes) and whole blood (9 genes) and all-tissue meta-analysis (33 genes); 11 of these had strongly colocalising splicing signals (Supplementary File: TWAS.xlsx).

Mendelian randomisation

We performed generalised summary-data-based Mendelian randomisation (GSMR)²³ in a replicated outcome study design using the pQTLs from the INTERVAL study²⁴. GSMR incorporates information from multiple independent SNPs and provides stronger evidence of a causal relationship than single SNP based approaches. Of 16 proteome-wide significant associations in this study, 8 were replicated in an external dataset at a Bonferroni-corrected p -value threshold of $P < 0.0031$ ($P < 0.05/16$; Extended Data Table 2, Extended Data Fig. 7).

Discussion

We report 23 replicated genetic associations with critical Covid-19, discovered in only 7,491 cases. This demonstrates the efficiency of the design of the GenOMICC study, an open-source²⁵ international research programme (<https://genomicc.org>) focusing on extreme

phenotypes: patients with life-threatening infectious disease, sepsis, pancreatitis and other critical illness phenotypes. GenOMICC detects greater heritability and stronger effect sizes than other study designs across all variants (Supplementary Figures 22,14). In Covid-19, critical illness is not only an extreme susceptibility phenotype, but also a more homogeneous one: we have shown previously that critically-ill Covid-19 patients are more likely to have the primary disease process - hypoxaemic respiratory failure²⁶ - and that this group have a divergent response to immunosuppressive therapy from other hospitalised patients²⁷. We detect distinct signals at several of the associated loci, in some cases implicating different biological mechanisms.

Five critical Covid-19-associated variants have direct roles in interferon signalling and broadly concordant predicted biological effects. These include a probable destabilising amino acid substitution in a ligand, *IFNA10* (Trp164Cys, Extended Data Fig. 5), and (as we reported previously¹) reduced expression of a subunit of its receptor *IFNAR2* (Fig. 2). *IFNAR2* signals through a kinase encoded by *TYK2*¹. Although the lead variant in *TYK2* in whole genome sequencing is a protein-coding variant with reduced STAT1 phosphorylation activity²⁸, it is also associated with significantly increased *TYK2* expression (Fig. 2, Methods). Fine-mapping reveals a significant association with an independent missense variant in *IL1ORB*, a receptor for Type III (lambda) interferons (rs8178521, Table 1). Finally, we detected a lead risk variant in phospholipid scramblase 1 (chr3:146517122:G:A, rs343320; *PLSCR1*) which disrupts a nuclear localisation signal important for the antiviral effect of interferon (Extended Data Fig. 5)²⁹. *PLSCR1* controls replication of other RNA viruses including vesicular stomatitis virus, encephalomyocarditis virus and Influenza A virus^{29,30}.

Although our genome-wide gene-based association tests did not replicate any findings from a previous pathway-specific study of rare deleterious variants⁴, our results provide robust evidence implicating reduced interferon signalling in susceptibility to critical Covid-19. Importantly, systemic administration of interferon in two large clinical trials, albeit late in disease, did not reduce mortality^{31,32}.

We found significant associations in genes implicated in lymphopoiesis, and differentiation of myeloid cells. *BCL11A* is essential in B- and T-lymphopoiesis³³ and promotes plasmacytoid dendritic cell differentiation³⁴. *TAC4*, reported previously³, encodes a regulator of B-cell lymphopoiesis³⁵ and antibody production³⁶, and promotes survival of dendritic cells³⁷. Finally, although the strongest fine mapping signal at 5q31.1 (chr5:131995059:C:T, rs56162149) is in an intron of *ACSL6* with significant effects on expression (Supplementary Material: TWAS.xlsx), the credible set includes a missense variant in *CSF2* (granulocyte-macrophage colony stimulating factor, GM-CSF) of uncertain significance (chr5:132075767:T:C, Extended Data Table 1). We have previously shown that GM-CSF is strongly up-regulated in critical Covid-19³⁸, and it is already under investigation as a target for therapy³⁹. Mendelian randomisation results are consistent with a direct link between plasma levels of a closely-related cytokine receptor subunit, IL3ra, and critical Covid-19 (Extended Data Table 2).

Fine mapping, colocalisation and TWAS provide evidence for increased expression of *MUC1* as the mediator of the association with rs41264915 (Supplementary Table 12). This suggests a potentially therapeutically-important role for mucins in the development of critical illness in Covid-19.

Mendelian randomisation reveals the first genetic evidence in support of a causal role for coagulation factors (*F8*) and platelet activation (*PDGFR1*) in critical Covid-19 (Extended Data Table 2, Extended Data Fig. 7), consistent with autopsy⁴⁰, proteomic⁴¹, and therapeutic⁴² evidence. Perhaps more importantly, we identify specific and closely-related intercellular adhesion molecules with known roles in inflammatory cell recruitment to sites of inflammation, including E-selectin (*SELE*), intercellular adhesion molecule 5 (*ICAM5*), and dendritic cell-specific intercellular adhesion molecule-3-Grabbing non-integrin (DC-SIGN, *CD209*), which may present additional

therapeutic targets. DC-SIGN (*CD209*) mediates pathogen endocytosis and antigen presentation, and has known roles in multiple viral infections, including SARS-CoV and influenza A virus. It has affinity for SARS-CoV-2^{43–45}.

Our previous report, in 2020, of an association between the *OAS* gene cluster and severe disease was robustly replicated in an external cohort¹, but does not meet genome-wide significance in the present analysis (Supplementary Table 7). This may indicate a change in the observed effect size because any effect that is detected in GWAS is more likely to have been sampled from the larger end of the range of possible effect sizes – the “winner’s curse”. Alternatively it may indicate either a change in the patient population (cases or controls) or a change in the pathogen. For example it is possible that, as with the other coronaviruses known to infect humans⁴⁶, more recent variants of SARS-CoV-2 have evolved to overcome this host antiviral defence mechanism.

Limitations

In contrast to microarray genotyping, whole genome sequencing is a rapidly evolving and relatively new technology for genome-wide association studies, with relatively few sources of population controls. We selected a control cohort from the 100,000 Genomes Project sequenced and analysed using a different platform and bioinformatics pipeline compared with the case cohort (Extended Data Fig. 1). However, to minimise the risk of false positive associations due to technical artifacts, extensive quality measures were utilised (See Methods): briefly, we masked low-quality genotypes, filtered for genotype signal using a low threshold for missingness, and performed a control-control relative allele frequency filter using a subset of samples processed with both bioinformatics pipelines. Finally, we required all significant associations to be supported by local variants in linkage disequilibrium, which may be excessively stringent (see Methods). Although this approach may remove some true associations, our priority is to maximise confidence in the reported signals. Of 25 variants meeting this requirement, 23 are externally replicated, and the remaining 2 may be true associations that are yet to be replicated due to a lack of coverage or power in the replication datasets.

The design of our study incorporates genetic signals for every stage in the disease progression into a single phenotype. This includes establishment of infection, viral replication, inflammatory lung injury and hypoxaemic respiratory failure. Although we can have considerable confidence that the replicated associations with critical Covid-19 we report are robust, we cannot determine at which stage in the disease process, or in which tissue, the relevant biological mechanisms are active.

Conclusions

These genetic associations implicate new biological mechanisms underlying the development of life-threatening Covid-19, several of which may be amenable to therapeutic targeting. Furthermore, we demonstrate the value of whole genome sequencing in to fine map loci in a complex trait. In the context of the ongoing global pandemic, translation to clinical practice is an urgent priority. As with our previous work, biological and molecular studies, and, where appropriate, large-scale randomised trials, will be essential before translating our findings into clinical practice.

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/s41586-022-04576-6>.

- Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
- Ellinghaus, D. et al. Genomewide association study of severe covid-19 with respiratory failure. *The New England Journal of Medicine* **383**, 1522–1534 (2020).
- COVID-19 Host Genetics Initiative Mapping the human genetic architecture of COVID-19. *Nature* (2021). <https://doi.org/10.1038/s41586-021-03767-x>
- Zhang, Q. et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science (New York, N.Y.)* **370**, eabd4570 (2020).
- 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).
- Dorward, D. A. et al. Tissue-Specific Immunopathology in Fatal COVID-19. *American Journal of Respiratory and Critical Care Medicine* **203**, 192–201 (2021).
- Millar, J. E., et al. Robust, reproducible clinical patterns in hospitalised patients with COVID-19. Preprint at *medRxiv* <https://doi.org/10.1101/2020.08.14.20168088> (2020).
- Horby, P. et al. Dexamethasone in Hospitalized Patients with Covid-19 Preliminary Report. *New England Journal of Medicine* (2020). <https://doi.org/10.1056/NEJMoa2021436>
- Degenhardt, F., et al. New susceptibility loci for severe COVID-19 by detailed GWAS analysis in European populations. Preprint at *medRxiv* <https://doi.org/10.1101/2021.07.21.21260624> (2021).
- Povysil, G., et al. Rare loss-of-function variants in type I IFN immunity genes are not associated with severe COVID-19. *The Journal of clinical investigation* **131**, (2021).
- Kosmicki, J. A. et al. Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. *American Journal of Human Genetics* **108**, 1350–1355 (2021).
- Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335–1341 (2018).
- Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273–1300 (2020).
- Rentsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**, D886–D894 (2018).
- Zheng, X. et al. HIBAG - HLA genotype imputation with attribute bagging. *Pharmacogenomics Journal* **14**, 192–200 (2014).
- Langton, D. J. et al. The influence of HLA genotype on the severity of COVID-19 infection. *HLA* **98**, 14–22 (2021).
- Zhou, W. et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature Genetics* **52**, 634–639 (2020).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Asano, T., et al. X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19. *Science immunology* **6**, (2021).
- Fallerini, C., et al. Association of toll-like receptor 7 variants with life-threatening COVID-19 disease in males: Findings from a nested case-control study. *eLife* **10**, (2021).
- Made, C. I. V. D. et al. Presence of genetic variants among young men with severe COVID-19. *JAMA* **324**, 663–673 (2020).
- Consortium, T. G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature communications* **9**, 224 (2018).
- Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
- Dunning, J. W. et al. Open source clinical science for emerging infections. *The Lancet Infectious Diseases* **14**, 8–9 (2014).
- 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**, (2020).
- Horby, P. et al. Dexamethasone in hospitalized patients with covid-19. *The New England Journal of Medicine* **384**, 693–704 (2021).
- Dendrou, C. A. et al. Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Science Translational Medicine* **8**, 363ra149 (2016).
- Dong, B. et al. Phospholipid scramblase 1 potentiates the antiviral activity of interferon. *Journal of virology* **78**, 8983–93 (2004).
- Luo, W. et al. Phospholipid scramblase 1 interacts with influenza A virus NP, impairing its nuclear import and thereby suppressing virus replication. *PLoS pathogens* **14**, e1006851 (2018).
- Repurposed Antiviral Drugs for Covid-19 Interim WHO Solidarity Trial Results. *New England Journal of Medicine* **0**, null (2020).
- Kalil, A. C., et al. Efficacy of interferon beta-1a plus remdesivir compared with remdesivir alone in hospitalised adults with COVID-19: A double-blind, randomised, placebo-controlled, phase 3 trial. *The Lancet Respiratory Medicine* **0**, (2021).
- Yu, Y. et al. Bcl11a is essential for lymphoid development and negatively regulates p53. *The Journal of experimental medicine* **209**, 2467–83 (2012).
- Reizis, B. Plasmacytoid Dendritic Cells: Development, Regulation, and Function. *Immunity* **50**, 37–50 (2019).
- Zhang, Y., Lu, L., Furlonger, C., Wu, G. E. & Paige, C. J. Hemokinin is a hematopoietic-specific tachykinin that regulates b lymphopoiesis. *Nature immunology* **1**, 392–7 (2000).
- Wang, W. et al. Hemokinin-1 activates the MAPK pathway and enhances b cell proliferation and antibody production. *Journal of immunology (Baltimore, Md. : 1950)* **184**, 3590–7 (2010).
- Janelsins, B. M. et al. Proinflammatory tachykinins that signal through the neurokinin 1 receptor promote survival of dendritic cells and potent cellular immunity. *Blood* **113**, 3017–26 (2009).
- Thwaites, R. S., & Investigators*, et al., on behalf of the I. Inflammatory profiles across the spectrum of disease reveal a distinct role for GM-CSF in severe COVID-19. *Science Immunology* **6**, (2021).

39. Lang, F. M., Lee, K. M.-C., Teijaro, J. R., Becher, B. & Hamilton, J. A. GM-CSF-based treatments in COVID-19: Reconciling opposing therapeutic approaches. *Nature reviews. Immunology* **20**, 507–514 (2020).
40. DA, D. et al. Tissue-specific immunopathology in fatal COVID-19. *American journal of respiratory and critical care medicine* **203**, 192–201 (2021).
41. Reyes, L. et al. a type I IFN, prothrombotic hyperinflammatory neutrophil signature is distinct for COVID-19 ARDS-. *Wellcome open research* **6**, 38 (2021).
42. Lawler, P. R. et al. Therapeutic anticoagulation with heparin in noncritically ill patients with covid-19. *The New England journal of medicine* **385**, 790–802 (2021).
43. Amraei, R. et al. CD209L/L-SIGN and CD209/DC-SIGN Act as Receptors for SARS-CoV-2. *ACS Central Science* **7**, 1156–1165 (2021).
44. Amraei, R., et al. CD209L/L-SIGN and CD209/DC-SIGN act as receptors for SARS-CoV-2. Preprint at *bioRxiv : the preprint server for biology* <https://doi.org/10.1101/2020.06.22.165803> (2021).
45. Thépaut, M. et al. DC/L-SIGN recognition of spike glycoprotein promotes SARS-CoV-2 trans-infection and can be inhibited by a glycomimetic antagonist. *PLOS Pathogens* **17**, e1009576 (2021).
46. Silverman, R. H. & Weiss, S. R. Viral Phosphodiesterases That Antagonize Double-Stranded RNA Signaling to RNase L by Degrading 2-5A. *Journal of Interferon & Cytokine Research* **34**, 455–463 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

¹Genomics England, London, UK. ²Roslin Institute, University of Edinburgh, Easter Bush, Edinburgh, UK. ³MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, UK. ⁴Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, 47 Little France Crescent, Edinburgh, UK. ⁵Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, UK. ⁶Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia. ⁷Biostatistics Group, Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University, Guangzhou, China. ⁸Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, Teviot Place,

Edinburgh, UK. ⁹Edinburgh Clinical Research Facility, Western General Hospital, University of Edinburgh, Edinburgh, UK. ¹⁰Intensive Care Unit, Royal Infirmary of Edinburgh, 54 Little France Drive, Edinburgh, UK. ¹¹Department of Critical Care Medicine, Queen's University and Kingston Health Sciences Centre, Kingston, ON, Canada. ¹²Clinical Research Centre at St Vincent's University Hospital, University College Dublin, Dublin, Ireland. ¹³NIHR Health Protection Research Unit for Emerging and Zoonotic Infections, Institute of Infection, Veterinary and Ecological Sciences University of Liverpool, Liverpool, UK. ¹⁴Respiratory Medicine, Alder Hey Children's Hospital, Institute in The Park, University of Liverpool, Alder Hey Children's Hospital, Liverpool, UK. ¹⁵illumina Cambridge, 19 Granta Park, Great Abington, Cambridge, UK. ¹⁶Regeneron Genetics Center, 777 Old Saw Mill River Rd., Tarrytown, USA. ¹⁷Geisinger, Danville, PA, USA. ¹⁸Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ¹⁹Test and Trace, the Health Security Agency, Department of Health and Social Care, Victoria St, London, UK. ²⁰Department of Intensive Care Medicine, Guy's and St. Thomas NHS Foundation Trust, London, UK. ²¹Department of Medicine, University of Cambridge, Cambridge, UK. ²²William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK. ²³Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Roosevelt Drive, Oxford, UK. ²⁴Department of Anaesthesia and Intensive Care, The Chinese University of Hong Kong, Prince of Wales Hospital, Hong Kong, China. ²⁵Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, Northern Ireland, UK. ²⁶Department of Intensive Care Medicine, Royal Victoria Hospital, Belfast, Northern Ireland, UK. ²⁷UCL Centre for Human Health and Performance, London, UK. ²⁸National Heart and Lung Institute, Imperial College London, London, UK. ²⁹Imperial College Healthcare NHS Trust: London, London, UK. ³⁰Imperial College, London, UK. ³¹Intensive Care National Audit & Research Centre, London, UK. ³²School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China. ³³Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China. ³⁴Great Ormond Street Hospital, London, UK. ³⁵William Harvey Research Institute, Queen Mary University of London, Charterhouse Square, London, UK. ³⁶These authors contributed equally: Athanasios Kousathanas, Erola Pairo-Castineira. ³⁷These authors jointly supervised this work: Sara Clohissey Hendry, Loukas Moutsianas, Andy Law, Mark J Caulfield, J. Kenneth Baillie. *A list of authors and their affiliations appears in the Supplementary Information. [✉]e-mail: m.j.caulfield@qmul.ac.uk; j.k.baillie@ed.ac.uk

Fig. 1 | GWAS results for the EUR ancestry group, and multi-ancestry meta-analysis. Manhattan plots are shown on the left and quantile–quantile (QQ) plots of observed versus expected P values are shown on the right, with genomic inflation (λ) displayed for each analysis. Highlighted results in blue in the Manhattan plots indicate variants that are LD-clumped ($r^2=0.1, P_2=0.01$, EUR LD) with the lead variants at each locus. Gene name annotation indicates

genes impacted by the predicted worst consequence type of each lead variant (annotation by Variant Effect Predictor (VEP)). For the HLA locus, the gene that was identified by HLA allele analysis is annotated. GWAS was performed using logistic regression and meta-analysed by the inverse variant method. The red dashed line shows the Bonferroni-corrected P -value $= 2.2 \times 10^{-8}$.

Fig. 2 | Gene-level Manhattan plot showing results from TWAS meta-analysis and highlighting genes that colocalise with GWAS signals or have strong metaTWAS associations. Highlighting color is different for lung and blood tissue data that were used for colocalisation and we also distinguish loci that were significant in both. Results are grouped according to two classes for the posterior probability of colocalisation (PP_{H4}): $P > 0.5$ and $P > 0.8$. If a

variant is placed in both classes then the color corresponding to the higher probability class is displayed. Arrows show direction of change in gene expression associated with an increased disease risk. Red dashed line shows bonferroni-corrected significance threshold for the metaTWAS analysis at $P < 2.3 \times 10^{-6}$.

Fig. 3 | Regional detail showing fine-mapping to identify two adjacent independent signals on Chromosome 3. Top two panels: variants in linkage disequilibrium with the lead variants shown. The variants that are included in two independent credible sets are displayed with black outline circles. r^2 values in the legend denote upper limits, *i.e.* $0.2=[0,0.2]$, $0.4=[0.2,0.4]$, $0.6=[0.4,0.6]$, $0.8=[0.6,0.8]$, $1=[0.8,1]$. Bottom panel: locations of protein-coding genes, coloured by TWAS P -value. The red dashed line shows the Bonferroni-corrected P -value= 2.2×10^{-8} for Europeans.

Table 1 | Lead variants from independent association signals in the per-population GWAS and multi-ancestry meta-analysis

chr:pos (hg38)	rsid	REF	ALT	RAF	OR	OR _{CI}	P	P _{hgib2.23m}	P _{reg}	Consequence	Gene	Cit
1:155066988	rs114301457	C	T*	0.0058	2.4	1.82-3.16	6.8×10 ⁻¹⁰	0.00011*	-	synonymous	EFNA4	-
1:155175305 [†]	rs7528026	G	A*	0.032	1.4	1.24-1.55	7.16×10 ⁻⁹	0.00012*	-	intron	TRIM46	-
1:155197995	rs41264915	A*	G	0.89	1.3	1.19-1.37	1.02×10 ⁻¹²	1.51×10 ^{-9*}	-	intron	THBS3	(³)
2:60480453 [†]	rs1123573	A*	G	0.61	1.1	1.09-1.18	9.85×10 ⁻¹⁰	0.000018*	-	intron	BCL11A	-
3:45796521	rs2271616	G	T*	0.14	1.3	1.21-1.37	9.9×10 ⁻¹⁷	4.95×10 ^{-9*}	-	5' UTR	SLC6A20	(³)
3:45859597	rs73064425	C	T*	0.077	2.7	2.51-2.94	1.97×10 ⁻¹³³	1.02×10 ^{-77*}	-	intron	LZTFL1	²
3:146517122	rs343320	G	A*	0.081	1.2	1.16-1.35	4.94×10 ⁻⁹	0.00028*	-	missense	PLSCR1	-
5:131995059	rs56162149	C	T*	0.17	1.2	1.13-1.26	7.65×10 ⁻¹¹	0.00074*	-	intron	ACSL6	-
6:32623820	rs9271609	T*	C	0.65	1.1	1.09-1.19	3.26×10 ⁻⁹	0.89	-	-	HLA-DRB1	-
6:41515007 [†]	rs2496644	A*	C	0.015	1.4	1.32-1.60	7.59×10 ⁻¹⁵	3.17×10 ^{-7*}	-	intron	LINC01276	³
9:21206606	rs28368148	C	G*	0.013	1.7	1.45-2.09	1.93×10 ⁻⁹	0.0024	0.00089	missense	IFNA10	-
11:34482745	rs61882275	G*	A	0.62	1.1	1.10-1.20	1.61×10 ⁻¹⁰	1.9×10 ^{-10*}	-	intron	ELF5	-
12:132489230	rs56106917	GC	G*	0.49	1.1	1.09-1.18	2.08×10 ⁻⁹	0.00047*	-	upstream	FBRSL1	-
13:112889041	rs9577175	C	T*	0.23	1.2	1.12-1.24	3.71×10 ⁻¹¹	1.29×10 ^{-6*}	-	downstream	ATP11A	-
15:93046840 [†]	rs4424872	T*	A	0.0079	2.4	1.87-3.01	8.61×10 ⁻¹³	-	0.29	intron	RGMA	-
16:89196249	rs117169628	G	A*	0.15	1.2	1.12-1.26	4.4×10 ⁻⁹	6.57×10 ^{-9*}	-	missense	SLC22A31	-
17:46152620	rs2532300	T*	C	0.77	1.2	1.10-1.22	4.19×10 ⁻⁹	2.49×10 ^{-9*}	-	intron	KANSL1	⁹
17:49863260	rs3848456	C	A*	0.029	1.5	1.33-1.70	4.19×10 ⁻¹¹	1.34×10 ^{-7*}	-	regulatory	.	³
19:4717660	rs12610495	A	G*	0.31	1.3	1.27-1.38	3.91×10 ⁻³⁶	5.74×10 ^{-19*}	-	intron	DPP9	¹
19:10305768	rs73510898	G	A*	0.093	1.3	1.19-1.37	1.57×10 ⁻¹¹	0.00016*	-	intron	ZGLP1	-
19:10352442	rs34536443	G	C*	0.05	1.5	1.36-1.65	6.98×10 ⁻¹⁷	4.06×10 ^{-11*}	-	missense	TYK2	¹
19:48697960	rs368565	C	T*	0.44	1.1	1.1-1.2	3.55×10 ⁻¹¹	0.00087*	-	intron	FUT2	-
21:33230000	rs17860115	C	A*	0.32	1.2	1.19-1.3	9.69×10 ⁻²²	1.77×10 ^{-18*}	-	5' UTR	IFNAR2	¹
21:33287378	rs8178521	C	T*	0.27	1.2	1.12-1.23	3.53×10 ⁻¹²	8.02×10 ^{-6*}	-	intron	IL10RB	-
21:33959662	rs35370143	T	TAC*	0.083	1.3	1.17-1.36	1.24×10 ⁻⁹	2.33×10 ^{-7*}	-	intron	LINC00649	-

Variants and the reference and alternate allele are reported according to GRCh38. The three variants discovered in multi-ancestry meta-analysis but not in the European ancestry GWAS are labelled with †, and † indicates genome-wide significant heterogeneity. REF and ALT columns indicate the reference and alternative alleles; an asterisk (*) indicates the risk allele. For each variant, we report the risk allele frequency in Europeans (RAF), the odds ratio and 95% confidence interval (OR and OR_{CI}), and the association P-value. Consequence indicates the predicted worst consequence type across GENCODE basic transcripts predicted by VEP(v104), and Gene indicates the VEP-predicted gene, but not necessarily the causal mediator. For the HLA locus, the gene that was identified by HLA allele analysis is displayed. An asterisk (*) next to replication P-value (P_{hgib2.23m} or P_{reg}) indicates that the lead signal (from multi-ancestry meta-analysis) is replicated with a Bonferroni-corrected P < 0.002 (0.05/25) with a concordant direction of effect. Cit column lists citation numbers for the first publication of confirmed genome-wide associations with critical illness or (in brackets) any Covid-19 phenotype.

Ethics

GenOMICC Study: GenOMICC was approved by the following research ethics committees: Scotland "A" Research Ethics Committee, 15/SS/0110; Coventry and Warwickshire Research Ethics Committee (England, Wales and Northern Ireland), 19/WM/0247). Current and previous versions of the study protocol are available at <https://genomicc.org/protocol/>. **100,000 Genomes project:** the 100,000 Genomes project was approved by East of England – Cambridge Central REC REF 20/EE/0035. Only individuals from the 100,000 Genomes project for whom Whole Genome Sequencing data were available and who consented for their data to be used for research purposes were included in the analyses. **UK Biobank study:** ethical approval for the UK Biobank was previously obtained from the North West Centre for Research Ethics Committee (11/NW/0382). The work described herein was approved by UK Biobank under application number 26041. **GHS study:** approval for DiscovEHR analyses was provided by the Geisinger Health System Institutional Review Board under project number 2006-0258. **AncestryDNA study:** all data for this research project was from subjects who provided prior informed consent to participate in AncestryDNA's Human Diversity Project, as reviewed and approved by our external institutional review board, Advarra (formerly Quorum). All data was de-identified prior to use. **PMBB study:** appropriate consent was obtained from each participant regarding storage of biological specimens, genetic sequencing and genotyping, and access to all available EHR data. This study was approved by the Institutional Review Board of the University of Pennsylvania and complied with the principles set out in the Declaration of Helsinki. Informed consent was obtained for all study participants. **23andMe study:** Participants in this study were recruited from the customer base of 23andMe, Inc., a personal genetics company. All individuals included in the analyses provided informed consent and answered surveys online according to 23andMe human subjects research protocol, which was reviewed and approved by Ethical and Independent Review Services, a private institutional review board (<http://www.eandireview.com>).

Recruitment of cases

Patients were recruited to the GenOMICC (Genetics Of Mortality In Critical Care) study in 224 UK intensive care units (<https://genomicc.org>). All cases had confirmed Covid-19 according to local clinical testing and were deemed, in the view of the treating clinician, to require continuous cardiorespiratory monitoring. In UK practice this kind of monitoring is undertaken in high-dependency or intensive care units.

Recruitment of controls

Mild/asymptomatic controls. Participants were recruited to the mild Covid-19 cohort on the basis of having experienced mild (non-hospitalised) or asymptomatic Covid-19. Participants volunteered to take part in the study via a microsite and were required to self-report the details of a positive Covid-19 test. Volunteers were prioritised for genome sequencing based on demographic matching with the critical Covid-19 cohort considering self-reported ancestry, sex, age and location within the UK. We refer to this cohort as the covid-mild cohort.

100,000 Genomes project controls. Participants were enrolled in the 100,000 Genomes Project from families with a broad range of rare diseases, cancers and infection by 13 regional NHS Genomic Medicine Centres across England and in Northern Ireland, Scotland and Wales. For this analysis, participants for whom a positive SARS-CoV-2 test had been recorded as of March, 2021 were not included due to uncertainty in the severity of Covid-19 symptoms. Only participants for whom genome sequencing was performed from blood derived DNA were included and participants with haematological malignancies were excluded to avoid potential tumour contamination.

DNA extraction

For severe Covid-19 cases and mild cohort controls, DNA was extracted from whole blood either manually using Nucleon Kit (Cytiva) and re-suspended in 1 ml TE buffer pH 7.5 (10mM Tris-Cl pH 7.5, 1mM EDTA pH 8.0), or automated on the Chemagic 360 platform using Chemagic DNA blood kit (Perkin Elmer) and re-suspended in 400µL Elution Buffer. The yield of the DNA was measured using Qubit and normalised to 50ng/µl before sequencing. For the 100,000 Genomes Project samples, DNA was extracted from whole blood at designated extraction centres following sample handling guidance provided by Genomics England and NHS England.

Whole Genome Sequencing

Sequencing libraries were generated using the Illumina TruSeq DNA PCR-Free High Throughput Sample Preparation kit and sequenced with 150bp paired-end reads in a single lane of an Illumina HiSeq X instrument (for 100,000 Genomes Project samples) or NovaSeq instrument (for the Covid-19 critical and mild cohorts).

Sequencing data QC. All genome sequencing data were required to meet minimum quality metrics and quality control measures were applied for all genomes as part of the bioinformatics pipeline. The minimum data requirements for all genomes were $>85 \times 10^9$ bases with $Q \geq 30$ and $\geq 95\%$ of the autosomal genome covered at $\geq 15\times$ calculated from reads with mapping quality >10 after removing duplicate reads and overlapping bases, after adaptor and quality trimming. Assessment of germline cross-sample contamination was performed using Verify-BamID and samples with $>3\%$ contamination were excluded. Sex checks were performed to confirm that the sex reported for a participant was concordant with the sex inferred from the genomic data.

WGS Alignment and variant calling

Covid-19 cohorts. For the critical and mild Covid-19 cohorts, sequencing data alignment and variant calling was performed with Genomics England pipeline 2.0 which uses the DRAGEN software (v3.2.22). Alignment was performed to genome reference GRCh38 including decoy contigs and alternate haplotypes (ALT contigs), with ALT-aware mapping and variant calling to improve specificity.

100,000 Genome Project cohort (100K-genomes). All genomes from the 100,000 Genomes Project cohort were analysed with the Illumina North Star Version 4 Whole Genome Sequencing Workflow (NSV4, version 2.6.53.23); which is comprised of the ISAAC Aligner (version 03.16.02.19) and Starling Small Variant Caller (version 2.4.7). Samples were aligned to the Homo Sapiens NCBI GRCh38 assembly with decoys.

A subset of the genomes from the Cancer program of the 100,000 Genomes Project were reprocessed (alignment and variant calling) using the same pipeline used for the Covid-19 cohorts (DRAGEN v3.2.22) for equity of alignment and variant calling.

Aggregation

Aggregation was conducted separately for the samples analysed with Genomics England pipeline 2.0 (severe-cohort, mild-cohort, cancer-realigned-100K), and those analysed with the Illumina North Star Version 4 pipeline (100K-Genomes).

For the first three, the WGS data were aggregated from single sample gVCF files to multi-sample VCF files using GVCfGenotyper (GG) v3.8.1, which accepts gVCF files generated via the DRAGEN pipeline as input. GG outputs multi-allelic variants (several ALT variants per position on the same row), and for downstream analyses the output was decomposed to bi-allelic variants per row using software vt v0.57721. We refer to the aggregate as aggCOVID_vX, where X is the specific freeze. The analysis in this manuscript uses data from freeze v4.2 and the respective aggregate is referred to as aggCOVID_v4.2.

Aggregation for the 100K-Genomes cohort was performed using Illumina's gvcfgenotyper v2019.02.26, merged with bcftools v1.10.2 and normalised with vt v0.57721.

Sample Quality Control (QC)

Samples that failed any of the following four BAM-level QC filters: free-mix contamination (>3%), mean autosomal coverage (<25X), percent mapped reads (<90%), and percent chimeric reads (>5%) were excluded from the analysis.

Additionally, a set of VCF-level QC filters were applied post-aggregation on all autosomal bi-allelic SNVs (akin to gnomAD v3.1)¹⁸. Samples were filtered out based on the residuals of eleven QC metrics (calculated using bcftools) after regressing out the effects of sequencing platform and the first three ancestry assignment principal components (including all linear, quadratic, and interaction terms) taken from the sample projections onto the SNP loadings from the individuals of 1000 Genomes Project phase 3 (1KGP3). Samples were removed that were four median absolute deviations (MADs) above or below the median for the following metrics: ratio heterozygous-homozygous, ratio insertions-deletions, ratio transitions-transversions, total deletions, total insertions, total heterozygous snps, total homozygous snps, total transitions, total transversions. For the number of total singletons (snps), samples were removed that were more than 8 MADs above the median. For the ratio of heterozygous to homozygous alternate snps, samples were removed that were more than 4 MADs above the median.

After quality control, 79,803 individuals were included in the analysis with the breakdown according to cohort shown in Supplementary Table 2.

Selection of high-quality (HQ) independent SNPs

We selected high-quality independent variants for inferring kinship coefficients, performing PCA, assigning ancestry and for the conditioning on the Genetic Relatedness matrix by the logistic mixed model of SAIGE and SAIGE-GENE. To avoid capturing platform and/or analysis pipeline effects for these analyses, we performed very stringent variant QC as described below.

HQ common SNPs. We started with autosomal, bi-allelic SNPs which had frequency > 5% in aggV2 (100K participant aggregate) and in the 1KGP3. We then restricted to variants that had missingness <1%, median genotype quality QC>30, median depth (DP) >=30 and >= 90% of heterozygote genotypes passing an ABRatio binomial test with P -value > 10^{-2} for aggV2 participants. We also excluded variants in complex regions from the list available in [https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_\(LD\)_lifted_over_for_GRCh38](https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)_lifted_over_for_GRCh38), and variants where the ref/alt combination was CG or AT (C/G, G/C, A/T, T/A). We also removed all SNPs which were out of Hardy Weinberg Equilibrium (HWE) in any of the AFR, EAS, EUR or SAS super-populations of aggV2, with a P -value cutoff of $pHWE < 10^{-5}$. We then LD-pruned using plink v1.9 with an $r^2 = 0.1$ and in 500kb windows. This resulted in a total of 63,523 high-quality sites from aggV2.

We then extracted these high-quality sites from the aggCOVID_v4.2 aggregate and further applied variant quality filters (missingness <1%, median QC>30, median depth >=30 and >= 90% of heterozygote genotypes passing an ABRatio binomial test with P -value > 10^{-2}), per batch of sequencing platform (i.e. HiSeqX, NovaSeq6000).

After applying variant filters in aggV2 and aggCOVID_v4.2, we merged the genomic data from the two aggregates for the intersection of the variants which resulted in a final total of 58,925 sites.

HQ rare SNPs. We selected high-quality rare ($MAF < 0.005$) bi-allelic SNPs to be used with SAIGE for aggregate variant testing analysis. To create this set, we applied the same variant QC procedure as with the common variants: We selected variants that had missingness <1%, median QC>30, median depth >=30 and >= 90% of heterozygote genotypes passing an ABRatio binomial test with P -value > 10^{-2} per batch of se-

quencing and genotyping platform (i.e. HiSeq+NSV4, HiSeq+Pipeline 2.0, NovaSeq+Pipeline 2.0). We then subsetted those to the following groups of MAC/MAF categories: MAC 1, 2, 3, 4, 5, 6-10, 11-20, MAC 20- MAF 0.001, MAF 0.001- 0.005.

Relatedness, ancestry and principal components

Kinship. We calculated kinship coefficients among all pairs of samples using software plink2 and its implementation of the KING robust algorithm. We used a kinship cutoff < 0.0442 to select unrelated individuals with argument "-king-cutoff".

Genetic Ancestry Prediction. To infer the ancestry of each individual we performed principal components analysis (PCA) on unrelated 1KGP3 individuals with GCTA v1.93.1_beta software using HQ common SNPs⁴⁷ and inferred the first 20 PCs. We calculated loadings for each SNP which we used to project aggV2 and aggCOVID_v4.2 individuals onto the 1KGP3 PCs. We then trained a random forest algorithm from R-package random-Forest with the first 10 1KGP3 PCs as features and the super-population ancestry of each individual as labels. These were 'AFR' for individuals of African ancestry, 'AMR' for individuals of American ancestry, 'EAS' for individuals of East Asian ancestry, 'EUR' for individuals of European ancestry, and 'SAS' for individuals of South Asian ancestry. We used 500 trees for the training. We then used the trained model to assign probability of belonging to a certain super-population class for each individual in our cohorts. We assigned individuals to a super-population when class probability >= 0.8. Individuals for which no class had probability >= 0.8 were labelled as "unassigned" and were not included in the analyses.

Principal component analysis. After labelling each individual with predicted genetic ancestry, we calculated ancestry-specific PCs using GCTA v1.93.1_beta⁴⁷. We computed 20 PCs for each of the ancestries that were used in the association analyses (AFR, EAS, EUR, and SAS).

Variant Quality Control

Variant QC was performed to ensure high quality of variants and to minimise batch effects due to using samples from different sequencing platforms (NovaSeq6000 and HiSeqX) and different variant callers (Strelka2 and DRAGEN). We first masked low-quality genotypes setting them to missing, merged aggregate files and then performed additional variant quality control separately for the two major types of association analyses, GWAS and AVT, which concerned common and rare variants, respectively.

Masking. Prior to any analysis we masked low quality genotypes using bcftools setGT module. Genotypes with DP<10, GQ<20, and heterozygote genotypes failing an AB-ratio binomial test with P -value < 10^{-3} were set to missing.

We then converted the masked VCF files to plink and bgen format using plink v2.0.

Merging of aggregate samples. Merging of aggV2 and aggCOVID_v4.2 samples was done using plink files with masked genotypes and the merge function of plink v1.9⁴⁸. for variants that were found in both aggregates.

GWAS analyses

Variant QC. We restricted all GWAS analyses to common variants applying the following filters using plink v1.9: $MAF > 0$ in both cases and controls, $MAF > 0.5\%$ and $MAC > 20$, missingness < 2%, Differential missingness between cases and controls, mid- P -value < 10^{-5} , HWE deviations on unrelated controls, mid- P -value < 10^{-6} , Multi-allelic variants were additionally required to have $MAF > 0.1\%$ in both aggV2 and aggCOVID_v4.2.

Control-control QC filter. 100K aggV2 samples that were aligned and genotype called with the Illumina North Star Version 4 pipeline

Article

represented the majority of control samples in our GWAS analyses, whereas all of the cases were aligned and called with Genomics England pipeline 2.0 (Supplementary Table 1). Therefore, the alignment and genotyping pipelines partially match the case/control status which necessitates additional filtering for adjusting for between-pipeline differences in alignment and variant calling. To control for potential batch effects, we used the overlap of 3,954 samples from the Genomics England 100K participants that were aligned and called with both pipelines. For each variant, we computed and compared between platforms the inferred allele frequency for the population samples. We then filtered out all variants that had > 1% relative difference in allele frequency between platforms. The relative difference was computed on a per-population basis for EUR (n=3,157), SAS (n=373), AFR (n=354) and EAS (n=81).

Model. We used a 2-step logistic mixed model regression approach as implemented in SAIGE v0.44.5 for single variant association analyses. In step 1, SAIGE fits the null mixed model and covariates. In step 2, single variant association tests are performed with the saddlepoint approximation (SPA) correction to calibrate unbalanced case-control ratios. We used the HQ common variant sites for fitting the null model and sex, age, age², age*sex and 20 principal components as covariates in step 1. The principal components were computed separately by predicted genetic ancestry (i.e., EUR-specific, AFR-specific, etc.), to capture subtle structure effects.

Analyses. All analyses were done on unrelated individuals with pairwise kinship coefficient < 0.0442. We conducted GWAS analyses per predicted genetic ancestry, for all populations for which we had >100 cases and >100 controls (AFR, EAS, EUR, and SAS).

Multiple testing correction. As our study is testing variants that were directly sequenced by WGS and not imputed, we calculated the *P*-value significance threshold by estimating the effective number of tests. After selecting the final filtered set of tested variants for each population, we LD-pruned in a window of 250Kb and $r^2 = 0.8$ with plink 1.9. We then computed the Bonferroni-corrected *P*-value threshold as 0.05 divided by the number of LD-pruned variants tested in the GWAS. The *P*-value thresholds that were used for declaring statistical significance are given in Supplementary Table 5.

LD-clumping. We used plink 1.9 to do clumping of variants that were genome-wide significant for each analysis with *P*1 set to per-population *P*-value from Supplementary Table 5, *P*2 = 0.01, clump distance 1500Kb and $r^2 = 0.1$.

Conditional analysis and signal independence. To find the set of independent variants in the per-population analyses, we performed a step-wise conditional analysis with the GWAS summary statistics for each population using GCTA 1.9.3 -cojo-slc function⁴⁷. The parameters for the function were *pval* = 2.2×10^{-8} , a distance of 10,000 kb and a colinear threshold of 0.9⁴⁹. For establishing independence of multi-ancestry meta-analysis signals from per-population discovered signals, we performed LD-clumping using the meta-analysis summaries and identified signals with no overlap with the LD-clumped results from the per-population analyses. In addition to the GCTA-cojo analysis, we also performed confirmatory individual-level conditional analysis as implemented in SAIGE. For every lead variant signal (including the multi-ancestry meta-analysis signals), we conditioned on the lead variants of all other signals identified as independent by GCTA-cojo and located on the same chromosome with option -condition of SAIGE (Supplementary Table 6).

Fine-mapping. We performed fine-mapping for genome-wide significant signals using Rpackage SusieR v0.11.42⁵⁰. For each genome-wide

significant variant locus, we selected the variants 1.5 Mbp on each side and computed the correlation matrix among them with plink v1.9. We then ran the susieR summary-statistics based function susie_rss and provided the summary z-scores from SAIGE (i.e., effect size divided by its standard error) and the correlation matrix computed with the same samples that were used for the corresponding GWAS. We required coverage ≥ 0.95 for each identified credible set and minimum and median absolute correlation coefficients (purity) of $r=0.1$ and 0.5, respectively.

Functional annotation of credible sets. We annotated all variants included in each credible set identified by SusieR using the online VEP v104 and selected the worst consequence across GENCODE basic transcripts (Supplementary File: GWAS.xlsx). We also ranked each variant within each credible set according to the predicted consequence and the ranking was based on the table provided by Ensembl: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html.

Multi-ancestry meta-analysis. We performed a meta-analysis across all ancestries using an inverse-variance weighting method and control for population stratification for each separate analysis in the METAL software⁵¹. The meta-analysed variants were filtered for variants with heterogeneity *P*-value $p < 2.22 \times 10^{-8}$ and variants that are not present in at least half of the individuals. We used the meta R package to plot forest plots of the clumped multi-ancestry meta-analysis variants⁵².

LD-based validation of lead GWAS signals. In order to quantify the support for genome-wide significant signals from nearby variants in LD, we assessed the internal consistency of GWAS results of the lead variants and their surroundings. To this end, we compared observed z-scores at lead variants with the expected z-scores based on those observed at neighbouring variants. Specifically, we computed the observed z-score for a variant *i* as $\hat{s}_i = \hat{\beta} / \hat{\sigma}_{\hat{\beta}}$ and, following the approach of⁵³, the imputed z-score at a target variant *t* as

$$\hat{s}_t = \Sigma_{t,P} (\Sigma_{P,P} + \lambda \mathbf{I})^{-1} \mathbf{s}_P$$

where $\mathbf{s}_P$ are the observed z-scores at a set *P* of predictor variants, $\Sigma_{x,y}$ is the empirical correlation matrix of dosage coded genotypes computed on the GWAS sample between the variants in *x* and *y*, and λ is a regularization parameter set to 10^{-5} . The set *P* of predictor variants consisted of all variants within 100 kb of the target variant with a genotype correlation with the target variant greater than 0.25. This approach is similar to one proposed recently by Chen et al.⁵⁴.

Stratified analysis. We performed sex-specific analysis (male and females separately) as well as analysis stratified by age (i.e., participants <60 and ≥60 years old) for the EUR ancestry group. To compare effect of variants within groups for the age and sex stratified analysis we first adjusted the effect and error of each variant for the standard deviation of the trait in each stratified group and then used the following t-statistic, as in previous studies^{55,56}

$$t = \frac{b_1 - b_2}{\sqrt{se_1^2 + se_2^2 - 2 \cdot r \cdot se_1 \cdot se_2}}$$

where b_1 is the adjusted effect for group 1, b_2 is the adjusted effect for group 2, se_1 and se_2 are the adjusted standard errors for group 1 and 2 respectively and *r* is the Spearman rank correlation between groups across all genetic variants.

Replication. In order to generate a replication set we conducted a meta-analysis of data from 23andMe, together with Host Genetic Initiative (HGI) GWAS meta-analysis round 6 hospitalised COVID vs population (B2 analysis), including all genetic ancestries. Although the HGI

programme included an analysis designed to mirror the GenOMICC study (analysis "A2"), most of these cases come from GenOMICC are already included in the discovery cohort. We therefore used the broader hospitalised phenotype ("B2") for replication.

In order to account for signal due to sample overlap we performed a mathematical subtraction form HGIv6B2, of the GenOMICC GWAS of European genetic ancestry. Publicly-available HGI data was downloaded from <https://www.covid19hg.org/results/r6/>. The subtraction was performed using MetaSubtract package (version 1.60) for R (version 4.0.2) after removing variants with the same genomic position and using the lambda.cohorts with genomic inflation calculated on the GenOMICC summary statistics.

We calculated a multi-ancestry meta-analysis for the three ancestries with summary statistics in 23andMe: African, Latino and European using variants that passed the 23andMe ancestry QC, with imputation score >0.6 and with $\text{maf} > 0.005$, before performing a final meta-analysis of 23andMe and HGI B2 without GenOMICC to create the final replication set. Meta-analysis were performed using METAL⁵¹, with the inverse-variance weighting method (STDERR mode) and genomic control ON. We considered that a hit was replicated if the direction of effect in the GenOMICC-subtracted HGI summary statistics was the same as in our GWAS, and the P -value was significant after Bonferroni correction for the number of attempted replications ($p_{\text{val}} < 0.05/25$). If the main hit was not present in the HGI-23andMe meta-analysis or if the hit was not replicating we looked for replication in variants in high LD with the top variant ($r^2 > 0.9$), which helped replicate two regions.

In order to attempt additional replication of two associations, we performed a multi-ancestry meta-analysis across 5 continental ancestry groups in UKB, AncestryDNA, Penn Medicine Biobank (PMBB), and Geisinger Health Systems (GHS) totaling 9937 hospitalized Covid-19 cases and 1,059,390 controls (Covid-19 negative or unknown). Hospitalization status (positive, negative or unknown) was determined based on Covid-19-related ICD10 codes U071, U072, U073 in variable 'diag_icd10' (table 'hesin_diag') in the UKB study; self-reported hospitalization due to Covid-19 in the AncestryDNA study; medical records in the GHS and PMBB studies. Association analyses in each study were performed using the genome-wide Firth logistic regression test implemented in REGENIE. In this implementation, Firth's approach is applied when the P -value from standard logistic regression score test is below 0.05. We included in step 1 of REGENIE (i.e., prediction of individual trait values based on the genetic data) directly genotyped variants with a minor allele frequency (MAF) $>1\%$, $<10\%$ missingness, Hardy-Weinberg equilibrium test $P > 1 \times 10^{-15}$ and linkage-disequilibrium (LD) pruning (1000 variant windows, 100 variant sliding windows and $r^2 < 0.9$). The association model used in step 2 of REGENIE included as covariates age, age^2 , sex, age-by-sex, and the first 10 ancestry-informative principal components (PCs) derived from the analysis of a stricter set of LD-pruned (50 variant windows, 5 variant sliding windows and $r^2 < 0.5$) common variants from the array (imputed for the GHS study) data. Within each study, association analyses were performed separately for five different continental ancestries defined based on the array data: African (AFR), Hispanic or Latin American (HLA), East Asian (EAS), European (EUR) and South Asian (SAS). Results were subsequently meta-analyzed across studies and ancestries using an inverse variance-weighted fixed-effects meta-analysis.

HLA Imputation and Association Analysis

HLA types were imputed at two field (4-digit) resolution for all samples within aggV2 and aggCOVID_v4.2 for the following seven loci: HLA-A, HLA-C, HLA-B, HLA-DRB1, HLA-DQA1, HLA-DQB1, and HLA-DPB1 using the HIBAG package in R¹⁵. At time of writing, HLA types were also imputed for 8% of samples using HLA*LA⁵⁷. Inferred HLA alleles between HIBAG and HLA*LA were $>96\%$ identical at 4-digit resolution. HLA association analysis was run under an additive model using SAIGE,

in an identical fashion to the SNV GWAS. The multi-sample VCF of aggregated HLA type calls from HIBAG were used as input where any allele call with posterior probability (T) < 0.5 were set to missing.

Aggregate variant testing (AVT)

Aggregate variant testing on aggCOVID_v4.2 was performed using SKAT-O as implemented in SAIGE-GENE v0.44.5¹⁷ on all protein-coding genes. Variant and sample QC for the preparation and masking of the aggregate files has been described elsewhere. We further excluded SNPs with differential missingness between cases and controls (mid- P value $< 10^{-5}$) or a site-wide missingness above 5%. Only bi-allelic SNPs with a MAF $< 0.5\%$ were included.

We filtered the variants to include in the aggregate variant testing by applying two functional annotation filters: A putative loss of function ($pLoF$) filter, where only variants that are annotated by LOFTEE¹⁸ as high confidence loss of function were included, and a more lenient (*missense*) filter where variants that have a consequence of missense or worse as annotated by VEP, with a CADD PHRED score of ≥ 10 , were also included. All variants were annotated using VEP v99. SAIGE-GENE was run with the same covariates used in the single variant analysis: sex, age, age^2 , $\text{age} \times \text{sex}$ and 20 (population-specific) principal components generated from common variants (MAF $\geq 5\%$).

We ran the tests separately by genetically predicted ancestry, as well as across all four ancestries as a mega-analysis. We considered a gene-wide significant threshold on the basis of the genes tested per ancestry, correcting for the two masks ($pLoF$ and *missense*, Supplementary Table 14).

Post-GWAS analysis

Transcriptome-wide Association Studies (TWAS). We performed TWAS in the MetaXcan framework and the GTExv8 eQTL and sQTL MASHR-M models available for download in <http://predictdb.org/>. We first calculated, using the European summary statistics, individual TWAS for whole blood and lung with the S-PrediXcan function^{58,59}. Then we performed a meta-TWAS including data from all tissues to increase statistical power using s-MultiXcan⁶⁰. We applied Bonferroni correction to the results in order to choose significant genes and introns for each analysis.

Colocalisation analysis. Significant genes from TWAS, splicing TWAS, metaTWAS and splicing metaTWAS, as well as genes where one of the top variants was a significant eQTL or sQTL were selected for a colocalisation analysis using the coloc R package⁶¹. We chose the lead SNPs from the European ancestry GWAS summary statistics and a region of ± 200 kb around each SNP to do the colocalisation with the identified genes in the region. GTExv8 whole blood and lung tissue summary statistics and eqtlGen (which has blood eQTL summary statistics for $>30,000$ individuals) were used for the analysis^{22,62}. We first performed a sensitivity analysis of the posterior probability of colocalisation (PP_{H4}) on the prior probability of colocalisation (P_{12}), going from $P_{12} = 10^{-8}$ to $P_{12} = 10^{-4}$ with the default threshold being $P_{12} = 10^{-5}$. eQTL signal and GWAS signals were deemed to colocalise if these two criteria were met: (1) At $P_{12} = 5 \times 10^{-5}$ the probability of colocalisation $PP_{H4} > 0.5$ and (2) At $P_{12} = 10^{-5}$ the probability of independent signal (PP_{H3}) was not the main hypothesis ($PP_{H3} < 0.5$). These criteria were chosen to allow eQTLs with weaker P -values due to lack of power in GTExv8, to be colocalised with the signal when the main hypothesis using small priors was that there was not any signal in the eQTL data.

As the chromosome 3 associated interval is larger than 200kb, we performed additional colocalisation including a region up to 500 kb, but no further colocalisations were found.

Mendelian Randomisation. We performed generalised summary-data-based Mendelian randomisation (GSMR)²³ in a replicated outcome study design. As exposures, we used the pQTLs from the INTERVAL

Article

study²⁴. We used the 1000 Genomes Project imputed data of the Health and Retirement Study (HRS) (n = 8,557) as the LD reference data required for GSMR analysis. The HRS data are available from dbGap (accession number: phs000428).

GSMR was undertaken using all exposures for which we were able to identify two or more independent SNPs associated with the exposure (P -value(exposure) $< 5 \times 10^{-8}$; linkage disequilibrium clumping ± 1 Mb, $r^2 < 0.05$; HEIDI-outlier filtering test, for the removal of SNPs with evidence of horizontal pleiotropy, was performed at the default threshold value of 0.01). Using GSMR, we identified those proteins implicated in determining Covid-19 severity in the new GenOMICC results (following genomic-control correction for inflation) at a false-discovery rate (FDR) < 0.05 , and attempted replication in the GWAS of "Hospitalized covid vs. population" (phenotype B2) of COVID19 HGI⁶³ having excluded the previous GenOMICC results. We achieved this by mathematically removing the contribution of GenOMICC¹ from the meta-analysis. We considered as replicated those results that passed a Bonferroni-corrected P -value threshold, correcting for the total number replication tests attempted (i.e. the number of observations from the discovery set with FDR < 0.05).

Heritability. For the SNP-based narrow-sense heritabilities of severe Covid-19 and HGI COVID phenotypes, both HDL and LD score regression (LDSC)⁶⁴ methods were applied. The HGI summary statistics were based on the GWAS analysis of all available samples, where the majority were European populations (see <https://www.covid19hg.org/results/r6/>). The munge_sumstats.py procedure in the LDSC software was used to harmonize the summary statistics, and in LDSC, the reference panel was built using the 1000 Genome European samples with SNPs that have minor allele frequencies (MAFs) > 0.05 . As both HDL and LDSC are based on GWAS summary Z-score statistics, the estimated heritabilities are thus on the observed scale.

Enrichment analysis. Enrichment analysis was performed to identify ontologies in which discovery genes were overrepresented. Using the XGR algorithm (<http://galahad.well.ox.ac.uk/XGR>)⁶⁵, 19 genes identified through lead variant proximity, credible variant sets, mutation consequence and TWAS analyses were tested for enrichment in disease ontology⁶⁶, gene ontologies (biological process, molecular function and cellular component)⁶⁷ and KEGG⁶⁸ and Reactome⁶⁹ pathways using default settings. This generated a P -value and FDR for overrepresentation of genes within each of the ontologies (Supplementary Table 15).

Reporting summary

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

Data availability

All data is available through <https://genomicc.org/data>. This includes downloadable summary data tables and instructions for applying to access individual-level data. Individual-level genome sequence data for the Covid-19 severe and mild cohorts can be analysed by qualified researchers in the UK Outbreak Data Analysis Platform at the University of Edinburgh by application at <https://genomicc.org/data>. Genomic data for the 100,000 genomes participants and a subset of Covid-19 cases are also available through the Genomics England research environment which can be accessed by application at <https://www.genomicsengland.co.uk/join-a-gecip-domain>. The full GWAS summary statistics for the 23andMe discovery dataset are available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. More information and access to the data are provided at <https://research.23andMe.com/dataset-access/>.

Code availability

Code to calculate the imputation of P -value based on LD SNPs is available at https://github.com/baillielab/GenOMICC_GWAS

47. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76–82 (2011).
48. Purcell, S. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
49. Yang, J. et al. Genetic Investigation of ANthropometric Traits (GIANT) Consortium & DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369–375 (2012).
50. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **82**, 1273–1300 (2020).
51. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* **26**, 2190–2191 (2010).
52. Balduzzi, S., Rücker, G. & Schwarzer, G. How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health* **22**, 153–160 (2019).
53. Pasaniuc, B. et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
54. Chen, W., et al Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. 2020.07.09.196535 (2020). <https://doi.org/10.1101/2020.07.09.196535>
55. Bernabeu, E., et al Sexual differences in genetic architecture in UK biobank. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.07.20.211813> (2020).
56. Winkler, A. E. A. G. & Thomas, W. AND Justice The influence of age and sex on genetic associations with adult body size and shape: A large-scale genome-wide interaction study. *PLOS Genetics* **11**, 1–42 (2015).
57. Dilthey, A. T. et al. HLA*LA—HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–4396 (2019).
58. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**, 1091–1098 (2015).
59. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**, 1825 (2018).
60. Barbeira, A. N. et al. Integrating predicted transcriptome from multiple tissues improves association detection. *PLOS Genetics* **15**, 1–20 (2019).
61. Giambartolomei, C. et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).
62. Vösa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* 2021 53:9 **53**, 1300–1310 (2021).
63. 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. *European journal of human genetics: EJHG* **28**, 715–718 (2020).
64. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47**, 291–5 (2015).
65. Fang, H., Knezevic, B., Burnham, K. L. & Knight, J. C. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Medicine* **8**, 129 (2016).
66. Schriml, L. M. et al. Disease Ontology: A backbone for disease semantic integration. *Nucleic Acids Research* **40**, D940–D946 (2012).
67. Ashburner, M. et al. Gene Ontology: Tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).
68. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
69. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Research* **48**, D498–D503 (2020).
70. Chen, M.-H. et al. Phospholipid scramblase 1 contains a nonclassical nuclear localization signal with unique binding site in importin alpha. *The Journal of Biological Chemistry* **280**, 10599–10606 (2005).
71. Thomas, C. et al. Structural linkage between ligand discrimination and receptor activation by type I interferons. *Cell* **146**, 621–32 (2011).
72. Chen, C.-W., Sowden, M., Zhao, Q., Wiedmer, T. & Sims, P. J. Nuclear phospholipid scramblase 1 prolongs the mitotic expansion of granulocyte precursors during G-CSF-induced granulopoiesis. *Journal of Leukocyte Biology* **90**, 221–233 (2011).

Acknowledgements We thank the patients and their loved ones who volunteered to contribute to this study at one of the most difficult times in their lives, and the research staff in every intensive care unit who recruited patients at personal risk during the most extreme conditions we have ever witnessed in UK hospitals. GenOMICC was funded by the Department of Health and Social Care (DHSC), Illumina, LifeArc, the Medical Research Council, UKRI, Sepsis Research (the Fiona Elizabeth Agnew Trust), the Intensive Care Society, a Wellcome Trust Senior Research Fellowship (J.K.Baillie, 223164/Z/21/Z) a BBSRC Institute Program Support Grant to the Roslin Institute (BBS/E/D/20002172, BBS/E/D/10002070 and BBS/E/D/30002275) and UKRI grants MC_PC_200004, MC_PC_19025, MC_PC_1905, and MRN02995X/1. Whole-genome sequencing was performed by Illumina at Illumina Laboratory Services (ILS) and was overseen by Genomics England. We would like to thank all at Genomics England who have contributed to the sequencing, clinical and genomic data analysis. This research is supported in part by the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK

Research and Innovation (grant ref MC_PC_20029). ADB would like to acknowledge funding from the Wellcome PhD training fellowship for clinicians (204979/Z/16/Z), the Edinburgh Clinical Academic Track (ECAT) programme. We would like to thank the research participants and employees of 23andMe for making this work possible. Genomics England and the 100,000 Genomes Project was funded by the National Institute for Health Research, the Wellcome Trust, the Medical Research Council, Cancer Research UK, the Department of Health and Social Care and NHS England. We are grateful for the support from Professor Dame Sue Hill and the team in NHS England and the 13 Genomic Medicine Centres that delivered the 100,000 Genomes Project which provide the majority of the control genome sequences for this study. We thank the participants of the 100,000 Genomes Project who made this study possible and the Genomics England Participant Panel for their strategic advice, involvement and engagement. We acknowledge NHS Digital, Public Health England and the Intensive Care National Audit and Research Centre who provided life course longitudinal clinical data on the participants. This work forms part of the portfolio of research of the National Institute for Health Research Barts Biomedical Research Centre. Mark Caulfield is an NIHR Senior Investigator. This study owes a great deal to the National Institute for Healthcare Research Clinical Research Network (NIHR CRN) and the Chief Scientist's Office (Scotland), who facilitate recruitment into research studies in NHS hospitals, and to the global ISARIC and InFACT consortia. Additional replication was conducted using the UK Biobank Resource (Project 26041). The Penn Medicine BioBank is funded by a gift from the Smilow family, the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA Award Number UL1TR001878, and the Perelman School of Medicine at the University of Pennsylvania. We thank the AncestryDNA customers who voluntarily contributed information in the Covid-19 survey. HRS (dbGaP accession: phs000428.v1.p1): HRS was supported by the National Institute on Aging (NIA U01AG009740). The genotyping was funded separately by the National Institute on Aging (RC2 AG036495, RC4 AG039029). Genotyping was conducted by the NIH Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotyping quality control and final preparation of the data were performed by the Genetics Coordinating Center at the University of Washington. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on August 22nd, 2021 (GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)). We would like to thank the research participants and employees of 23andMe for making this work possible. A full list of contributors who have provided data that was collated in the HGI project, including previous iterations, is available at <https://www.covid19hg.org/acknowledgements>. The views expressed

are those of the authors and not necessarily those of the DHSC, the NHS, DID, NIHR, MRC, Wellcome Trust or PHE.

Author contributions AK, EP-C, KR, AS, CAO, SW, TM, YW, XS, KSE, BW, DR, LK, MZ, NP, JAK, JEH, AB, GRA, MARF, AJ, TMI, MO, DJR, MDR, AV, JY, ADB, SCH, LMo, AL and JKB contributed to data analysis. AK, EP-C, KR, AS, CAO, SW, CDR, JM, AR, SCH, LMo and AL contributed to bioinformatics. AK, EP-C, KR, CDR, JM, DM, AN, MGS, SCH, LMo, MJC and JKB contributed to writing and reviewing the manuscript. EP-C, KR, KM, SK, AF, LM, KRo, CPP, VV, JFW, SCH, AL, MJC and JKB contributed to design. SW, FG, WO, PG and SD contributed to project management. FG, WO, KM, SK, PG, SD, DM, AN, MGS, SS, JK, TAF, MS-H, CS, CH, PH, LL, DMc, HM, PJO, PE, TW, AT, AF, LM, KRo, CPP, RHS, SCH and AL contributed to oversight. FG, WO, FM-C and JKB contributed to ethics and governance. KM, ASi, AF and LM contributed to sample handling and sequencing. and ASi contributed to data collection. and TZ contributed to sample handling. TZ, GE, CP, DB and CK contributed to sequencing. and LT contributed to recruitment of controls. GC, PA, KRo and AL contributed to clinical data management. KRo, CPP, SCH and JKB contributed to conception. KRo, CPP, VV and JFW contributed to reviewing the manuscript. MJC and JKB contributed to scientific leadership.

Competing interests J.A.K., J.E.H., A.B., G.R.A., and M.A.R.F. are current employees and/or stockholders of Regeneron Genetics Center or Regeneron Pharmaceuticals. Genomics England Ltd is a wholly owned Department of Health and Social Care company created in 2013 to work with the NHS to introduce advanced genomic technologies and analytics into healthcare. All Genomics England affiliated authors are, or were, salaried by Genomics England during this programme. All other authors declare that they have no conflicts of interest relating to this work.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-022-04576-6>.

Correspondence and requests for materials should be addressed to Mark J. Caulfield or J. Kenneth Baillie.

Peer review information Nature thanks Jacques Fellay and the other, anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Extended Data Fig. 1 | Diagram showing the analysis workflow for genome-wide association study (GWAS) and aggregate variant testing (AVT) analyses of this study. The cohorts displayed in yellow and green in the top box were processed with Genomics England Pipeline 2.0 and Illumina NSV4, respectively (see Methods on WGS Alignment and variant calling for details on differences between pipelines). We used individuals that were processed with either pipeline for the GWAS analyses and individuals

processed only with Genomics England Pipeline 2.0 for the aggregate variant burden testing (AVT) analyses. The definition of the cases and controls was the same for GWAS and AVT, cases were the Covid-19 severe individuals for both, and controls included individuals from the 100,000 genomes project (100K-Genomes) and also Covid-19 positive individuals that were recruited for this study and experienced only mild symptoms (covid-mild).

Extended Data Fig. 2 | Regional detail showing fine-mapping to identify three adjacent independent signals on Chromosome 1. Top two panels: variants in linkage disequilibrium with the lead variants shown. The variants that are included in two independent credible sets are displayed with black

outline circles. r^2 values in the legend denote upper limits, 0.2=[0.2,0.2], 0.4=[0.2,0.4], 0.6=[0.4,0.6], 0.8=[0.6,0.8], 1=[0.8,1]. Bottom panel: locations of protein-coding genes, coloured by TWAS P -value. The red dashed line shows the Bonferroni-corrected P -value= 2.2×10^{-8} for Europeans.

Extended Data Fig. 3 | Regional detail showing fine-mapping to identify two adjacent independent signals on Chromosome 19. Top two panels: variants in linkage disequilibrium with the lead variants shown. The variants that are included in two independent credible sets are displayed with black

outline circles. r^2 values in the legend denote upper limits, 0.2=[0,0.2], 0.4=[0.2,0.4], 0.6=[0.4,0.6], 0.8=[0.6,0.8], 1=[0.8,1]. Bottom panel: locations of protein-coding genes, coloured by TWAS P -value. The red dashed line shows the Bonferroni-corrected P -value= 2.2×10^{-8} for Europeans.

Extended Data Fig. 4 | Regional detail showing fine-mapping to identify three adjacent independent signals on Chromosome 21. Top three panels: variants in linkage disequilibrium with the lead variants shown. The variants that are included in three independent credible sets are displayed with black

outline circles. r^2 values in the legend denote upper limits, $0.2=[0,0.2]$, $0.4=[0.2,0.4]$, $0.6=[0.4,0.6]$, $0.8=[0.6,0.8]$, $1=[0.8,1]$. Bottom panel: locations of protein-coding genes, coloured by TWAS P -value. The red dashed line shows the Bonferroni-corrected P -value $=2.2 \times 10^{-8}$ for Europeans.

Extended Data Fig. 5 | Predicted structural consequences of lead variants at⁷⁰ and⁷¹. (a) Crystal structure of PLSCR1 nuclear localization signal (orange, Gly257–Ile266, numbering correspond to UniProt entry O15162) in complex with Importin α (blue), Protein Data Bank (PDB) ID 1Y2A. Side chains of PLSCR1 are shown as connected spheres with carbon atoms coloured in orange, nitrogens in blue and oxygens in red. Hydrogen atoms were not determined at this resolution (2.20) and are not shown. (b) Close-up view showing side chains of PLSCR1 Ser260, His262 and Importin Glu107 as sticks. Distance (in) between selected atoms (PLSCR1 His262 Ne2 and Importin Glu107 carboxyl O) is indicated. A hydrogen bond between PLSCR1 His262 and Importin Glu107 is indicated with a dashed line. The risk variant is predicted to eliminate this bond, disrupting nuclear import, an essential step for effect on antiviral signalling²⁹ and neutrophil maturation⁷². (c) Since there is very strong

sequence conservation between IFNA10 and the gene encoding IFNω, we used existing crystal structure data (Protein Data Bank ID 3SE4) for IFNω (cyan) to display a ternary complex with interferon α/β receptor IFNAR1 (blue), IFNAR2 (red). The side chain of Trp164 is shown as spheres and indicated with a black line. (d) The hydrophobic core of IFNω with Trp164 shielded from the solvent in the center. Trp164-surrounding residues of IFNω are numbered and correspond to UniProt entry P05000. Trp164 and surrounding residues are conserved in IFNA10 (UniProt ID P01566) and share the same numbering as in IFNω (P05000). Side chains of four residues are shown as sticks. Carbon and nitrogen atoms coloured in cyan and blue, respectively. The critical Covid-19-associated mutation, Trp164Cys, would replace an evolutionarily conserved, bulky side chain in the hydrophobic core of IFNA10 with a smaller one, which may destabilise IFNA10.

Overlaid GWAS and HLA associations: conditioned on DRB1*04:01
(Europeans vs vs_mld_aggV2)

Extended Data Fig. 6 | Manhattan plot of HLA and GWAS signal across the extended MHC region for the EUR cohort. Grey circles mark the GWAS (small variant) associations and diamonds represent the HLA each allele association, coloured by locus. The lead variant from the GWAS and lead allele from HLA are

labelled. The left-panel shows the raw association $-\log_{10}(p\text{-values})$ per variant prior to conditional analysis. The right-panel shows the $-\log_{10}(p\text{-values})$ per variant following conditioning on. The dashed red line shows the Bonferroni-corrected genome-wide significance threshold for Europeans.

Extended Data Fig. 7 | Effect-effect plots for Mendelian randomisation analyses to assess causal evidence for circulating proteins in critical Covid-19. Each plot shows effect size (β) of variants associated with protein

concentration (x-axis) and critical Covid-19 (y-axis). A full list of instruments is found in Supplementary Table 13.

Extended Data Table 1 | Fine-mapping results for lead variants and worst consequence variant in each credible set

Lead variant	Pop	Focal CS	nCS	Worst variant	Worst variant Pval	Lead variant CADD	Worst variant CADD	Worst Consequence	Worst gene
chr1:155066988:C:T	EUR	chr1:155197995:A:G	9	chr1:155066988:C:T	6.8×10^{-10}	10.3	10.3	synonymous	<i>EFNA4</i>
chr1:155175305:G:A	META	chr1:155197995:A:G	5	chr1:155134292:T:C	1.17×10^{-07}	8.34	1.37	3' UTR	<i>EFNA1</i>
chr1:155197995:A:G	EUR	chr1:155197995:A:G	3	chr1:155202934:T:C	3.15×10^{-12}	2.1	21.2	missense	<i>THBS3</i>
chr3:45796521:G:T	EUR	chr3:45859597:C:T	1	chr3:45796521:G:T	9.9×10^{-17}	9.19	9.19	5' UTR	<i>SLC6A20</i>
chr3:45859597:C:T	EUR	chr3:45859597:C:T	9	chr3:45825948:A:G	6.1×10^{-132}	0.143	7.96	3' UTR	<i>LZTFL1</i>
chr3:146517122:G:A	EUR	chr3:146517122:G:A	9	chr3:146517122:G:A	4.94×10^{-09}	22.6	22.6	missense	<i>PLSCR1</i>
chr5:131995059:C:T	EUR	chr5:131995059:C:T	32	chr5:132075767:T:C	1.48×10^{-09}	0.206	6.09	missense	<i>CSF2</i>
chr6:32623820:T:C	EUR	chr6:32623820:T:C	33	chr6:32467073:G:C	6.65×10^{-08}	10.1	8.12	intron	<i>HLA-DRB9</i>
chr6:41515007:A:C	META	chr6:41515652:G:C	8	chr6:41515652:G:C	5.17×10^{-08}	4.11	4.17	intron	<i>LINC01276</i>
chr9:21206606:C:G	EUR	chr9:21206606:C:G	3	chr9:21206606:C:G	1.93×10^{-09}	23.9	23.9	missense	<i>IFNA10</i>
chr11:34482745:G:A	EUR	chr11:34482745:G:A	4	chr11:34479140:G:A	2.56×10^{-10}	0.073	1.32	3' UTR	<i>ELF5</i>
chr12:132489230:G:C	EUR	chr12:132489230:G:C	25	chr12:132565387:T:C	1.42×10^{-07}	4.91	4.64	non coding transcript exon	-
chr13:112889041:C:T	EUR	chr13:112889041:C:T	4	chr13:112886111:C:T	5.36×10^{-11}	0.676	5.5	3' UTR	<i>ATP11A</i>
chr15:93046840:T:A	EUR	chr15:93046840:T:A	2	chr15:93046840:T:A	8.61×10^{-13}	4.45	4.45	intron	<i>RGMA</i>
chr16:89196249:G:A	EUR	chr16:89196249:G:A	4	chr16:89196249:G:A	4.4×10^{-09}	22.8	22.8	missense	<i>SLC22A31</i>
chr17:46152620:T:C	EUR	chr17:46152620:T:C	1430	chr17:45830530:T:C	1.14×10^{-07}	5.27	3.96	stop lost	<i>CRHR1</i>
chr17:49863260:C:A	EUR	chr17:49863260:C:A	5	chr17:49880589:C:T	1.91×10^{-09}	5.38	7.22	TF binding site	-
chr19:4717660:A:G	EUR	chr19:4717660:A:G	1	chr19:4717660:A:G	3.91×10^{-36}	16.3	16.3	intron	<i>DPP9</i>
chr19:10305768:G:A	EUR	chr19:10352442:G:C	3	chr19:10380329:C:G	7.93×10^{-11}	0.422	7.91	intron	<i>TYK2</i>
chr19:10352442:G:C	EUR	chr19:10352442:G:C	1	chr19:10352442:G:C	6.98×10^{-17}	25.1	25.1	missense	<i>TYK2</i>
chr19:48697960:C:T	EUR	chr19:48697960:C:T	10	chr19:48703346:C:T	6.75×10^{-10}	2.44	7.02	synonymous	<i>FUT2</i>
chr21:33230000:C:A	EUR	chr21:33230000:C:A	16	chr21:33262573:G:A	5.35×10^{-21}	10.1	3.43	missense	<i>IFNAR2</i>
chr21:33287378:C:T	EUR	chr21:33230000:C:A	33	chr21:33288868:T:G	1.59×10^{-07}	3.63	5.84	intron	<i>IL10RB</i>
chr21:33959602:T:TAC	EUR	chr21:33230000:C:A	23	chr21:33972178:G:A	1.32×10^{-08}	0.246	0.282	non coding transcript exon	<i>LINC00649</i>

Fine-mapping was performed in EUR for all variants except chr6:41515007:A:C which was fine-mapped in the SAS population for which the signal was strongest among the per-population analyses. Lead variant chr2:60480453:A:G (rs1123573) that was discovered in multi-ancestry meta-analysis is not included in the table as fine-mapping did not generate any credible sets with the required posterior inclusion probability of >0.95 for any of the populations. Focal CS is the index SNP that was used for fine-mapping with SusieR, 1.5 Mb on each side. nCS indicates the number of variants included in each credible set. Consequence annotation for all variants across credible sets was generated using VEPv104 and the worst consequence across GENCODE basic transcripts was chosen. All variants were ranked according to their consequence type and chr:pos_{hg38}:ref_{hg38}:alt, P-value and CADD score are provided for the variant with the worst consequence across all variants in each credible set.

Extended Data Table 2 | Identification of 16 proteins by the GSMR analysis for Covid-19 severity at FDR < 0.05

Gene	<i>BETA</i>	<i>SE</i>	<i>P</i>	$BETA_{hgib2.23m}$	$SE_{hgib2.23m}$	$P_{hgib2.23m}$
ICAM5	-0.15	0.025	2.82×10^{-9}	-0.07	0.013	7.65×10^{-8} *
GOLM1	0.22	0.037	2.92×10^{-9}	0.20	0.021	1.04×10^{-21} *
ICAM1	0.10	0.017	6.33×10^{-9}	0.013	0.009	0.14
ICAM5	-0.19	0.033	1.58×10^{-8}	-0.048	0.017	0.0054
FAM3D	0.13	0.024	2.12×10^{-8}	0.12	0.013	3.12×10^{-18} *
PDGFRL	0.10	0.021	1.85×10^{-6}	0.021	0.010	0.041
CD209	0.11	0.024	6.58×10^{-6}	0.11	0.014	1.88×10^{-15} *
ABO	0.064	0.017	0.00012	0.084	0.0088	7.76×10^{-22} *
C1GALT1C1	0.13	0.037	0.00026	0.055	0.030	0.063
CCL25	0.15	0.040	0.00026	0.035	0.023	0.13
F8	0.14	0.042	0.0011	0.16	0.020	1.46×10^{-14} *
TLR4:LY96	-0.12	0.038	0.0014	-	-	-
IL3RA	-0.087	0.028	0.0019	-0.065	0.014	4.33×10^{-6} *
SELE	-0.069	0.022	0.0019	-0.095	0.013	3.76×10^{-14} *
CAMK1	-0.064	0.021	0.00205	0.0047	0.0110	0.664
IL27RA	-0.084	0.028	0.00229	0.0020	0.0150	0.892

We report the effect size *BETA*, the standard error *SE* and the *P*-value *P* for the GenOMICC analysis and the replication with HGI B2 and 23andme meta-analysis. An asterisk (*) next to the replication *P*-value ($P_{hgib2.23m}$) indicates that the protein result is replicated with concordant direction of effect. We considered as replicated those results that passed a Bonferroni correction of the *p*-values of the replicated outcome Mendelian randomisation.

Reporting Summary

Nature Portfolio 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 Portfolio 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ 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 <i>Give P values as exact values whenever suitable.</i>
☐	☒ 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	Nucleon Kit (Cytiva), Chemagic 360 platform using Chemagic DNA blood kit (Perkin Elmer) . T Qubit and normalised Illumina TruSeq DNA PCR-Free High Throughput Sample Preparation kit, Illumina HiSeq X instrument (for 100,000 Genomes Project samples), NovaSeq instrument (for the COVID-19 critical and mild cohorts).
Data analysis	VerifyBAMID, DRAGEN(v3.2.22), Illumina North Star Version 4 Whole Genome Sequencing Workflow (NSV4, version 2.6.53.23), ISAAC Aligner (version 03.16.02.19), Starling Small Variant Caller (version 2.4.7), GVGCFGenotyper (GG) v3.8.1, vt v0.57721, gvcfgenotyper v2019.02.26, bcftools v1.10.2, plink 1.9, HiSeq+NSV4, HiSeq+Pipeline 2.0, NovaSeq+Pipeline 2.0, KING 2.1, plink2, GCTA v1.93.1_beta, Strelka2, SAIGE v0.44.5, GTCA 1.9.3, SusieR v0.11.42, VEPv104, metal 2018-08-28, MetaSubtract package (v1.60), R(v4.0.2), SAIGE-GENE v0.44.5, LOFTEE, VEPv99, MetaXCan (v0.6.5), coloc R package 5.1.0, R 3.6.3, GSNR, Heidi, HIBAG R package 1.8.3, XGR package (20-Apr-2020), LDSC (v1.0.1), HDL(v.1.4.0), REGENIEv2.2

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 Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Full summary data in support of the findings of this study will be available for download from <https://genomicc.org/data> concurrently with publication. Individual-level data can be analysed by qualified researchers in the UK Outbreak Data Analysis Platform at the University of Edinburgh by application at <https://genomicc.org/data>. Genomicc data for 1000,000 genomes participants and cases are available through the Genomics England research environment. The full GWAS summary statistics for the 23andMe discovery data set will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit <https://research.23andme.com/dataset-access/> for more information and to apply to access the data.

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](https://www.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	Cases: n=7,491, controls n=48,400 (mild cases n=1630, 100k controls n=46,770). European ancestry ncases=5,989 ncontrols=41,891 South Asian ancestry ncases=788, ncontrols=3,793 African ancestry ncases=440, ncontrols=1,350 East Asian ancestry ncases=274, ncontrols=366
Data exclusions	no exclusions
Replication	Data from the Host Genetics Initiative (HGI) data freeze 6 B2 analysis (hospitalised cases) were combined in a meta-analysis with data shared by 23andMe, Inc. We removed signals in HGI derived from GenOMICC cases for independence. 22 of the 25 independent GWAS signals were replicated. Replication of rs28368148 and rs4424872 was attempted using a trans-ancestry meta-analysis UKB, AncestryDNA, Penn Medicine Biobank (PMBB), and Geisinger Health Systems (GHS) totaling 9937 hospitalized COVID-19 cases and 1,059,390 controls (COVID-19 negative or unknown). 1 extra loci (rs28368148) was replicated using this method. The loci not replicated correspond to the lead snps: rs9271609 in HLA region and rs4424872 next to RGMA.
Randomization	Not relevant to the study. There wasn't any allocation to experimental groups
Blinding	Not relevant to the study.

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

Severe COVID-19 (n=7491). Significant comorbidities: 1,605. Died(60 days) 2154, Invasive Ventilation 4028. mean age=60, mean BMI=29.9
 European ancestry (n=5,989): males 4,062 females 1,927.
 South Asian ancestry (n=788): males:586, females:202
 East Asian ancestry(n=274) males:162, females:112
 African ancestry(n=440) males:286, females:154
 100K controls: 18,915 unaffected family members of rare diseases participants. 14,701 affected rare diseases participants, 1,005 not assessed for disease status, 12,149 cancer participants.
 mean age: 51
 mean BMI: 26.1
 European ancestry (n=41,384): males 18,971 females 22,413
 South Asian ancestry (n=3698): males:1802, females:1896
 East Asian ancestry(n=352) males:138, females:224
 African ancestry(n=1,236) males:632, females:704
 Mild COVID-19 cohort. mean age: 46.
 European ancestry (n=1507): males:410, females:1,097
 South Asian ancestry (n=95): males:43, females:52
 East Asian ancestry(n=14) males:8, females:6
 African ancestry(n=14) males:6, females:8

Recruitment

Critically ill patients recruited to the GenOMICC study (genomicc.org) had confirmed COVID-19 according to local clinical testing and were deemed, in the view of the treating clinician, to require continuous cardiorespiratory monitoring. In UK practice this kind of monitoring is undertaken in high dependency or intensive care units. Patients were recruited from 224 ICU across the UK.
 Participants were recruited to the mild COVID-19 cohort on the basis of having experienced mild (non-hospitalised) or asymptomatic COVID-19. Participants volunteered to take part in the study via a microsite and were required to self-report the details of a positive COVID-19 test. Volunteers were prioritised for genome sequencing based on demographic matching with the critical COVID-19 cohort considering self-reported ancestry, sex, age and location within the UK.
 Participants were enrolled in the 100,000 Genomes Project from families with a broad range of rare diseases, cancers and infection by 13 regional NHS Genomic Medicine Centres across England and in Northern Ireland, Scotland and Wales. For this analysis, participants for whom a positive SARS-CoV-2 test had been recorded as of March, 2021 were not included due to uncertainty in the severity of COVID-19 symptoms. Only participants for whom genome sequencing was performed from blood derived DNA were included and participants with haematological malignancies were excluded to avoid potential tumour contamination.

Ethics oversight

Research ethics committees (Scotland 15/SS/0110, England, Wales and Northern Ireland: 19/WM/0247. Current and previous versions of the study protocol are available at genomicc.org/protocol.
 UKBiobank Study: ethical approval for the UK Biobank was previously obtained from the North West Centre for Research Ethics Committee(11/NW/0382). The work described herein was approved by UK Biobank under application number 26041.
 GHS study: approval for DiscovEHR analyses was provided by the Geisinger Health System Institutional Review Board under project number 2006-0258.
 AncestryDNA study: all data for this research project was from subjects who provided prior informed consent to participate in AncestryDNA's Human Diversity Project, as reviewed and approved by our external institutional review board, Advarra (formerlyQuorum). All data was de-identified prior to use.
 PMBBstudy: appropriate consent was obtained from each participant regarding storage of biological specimens, genetic sequencing and genotyping, and access to all available EHR data. This study was approved by the Institutional Review Board of the University of Pennsylvania and complied with the principles set out in the Declaration of Helsinki. Informed consent was obtained for all study participants.

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