REPORT

Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals

Jack A. Kosmicki,[1,27] Julie E. Horowitz,[1,27] Nilanjana Banerjee,[1] Rouel Lanche,[1] Anthony Marcketta,[1] Evan Maxwell,[1] Xiaodong Bai,[1] Dylan Sun,[1] Joshua D. Backman,[1] Deepika Sharma,[1] Fabricio S.P. Kury,[1] Hyun M. Kang,[1] Colm O'Dushlaine,[1] Ashish Yadav,[1] Adam J. Mansfield,[1] Alexander H. Li,[1] Kyoko Watanabe,[1] Lauren Gurski,[1] Shane E. McCarthy,[1] Adam E. Locke,[1] Shareef Khalid,[1] Sean O'Keeffe,[1] Joelle Mbatchou,[1] Olympe Chazara,[2] Yunfeng Huang,[3] Erika Kvikstad,[5] Amanda O'Neill,[2] Paul Nioi,[4] Meg M. Parker,[4] Slavé Petrovski,[2] Heiko Runz,[3] Joseph D. Szustakowski,[5] Quanli Wang,[2] Emily Wong,[6] Aldo Cordova-Palomera,[6] Erin N. Smith,[6] Sandor Szalma,[6] Xiuwen Zheng,[7] Sahar Esmaeeli,[7] Justin W. Davis,[7] Yi-Pin Lai,[8] Xing Chen,[8] Anne E. Justice,[9] Joseph B. Leader,[9] Tooraj Mirshahi,[9] David J. Carey,[9] Anurag Verma,[10] Giorgio Sirugo,[10] Marylyn D. Ritchie,[10] Daniel J. Rader,[10] Gundula Povysil,[11] David B. Goldstein,[11,12] Krzysztof Kiryluk,[11,13] Erola Pairo-Castineira,[14,15] Konrad Rawlik,[14] Dorota Pasko,[16] Susan Walker,[16] Alison Meynert,[15] Athanasios Kousathanas,[16] Loukas Moutsianas,[16] Albert Tenesa,[14,15,17] Mark Caulfield,[16,18] Richard Scott,[16,19] James F. Wilson,[15,17] J. Kenneth Baillie,[14,15,20] Guillaume Butler-Laporte,[21,22] Tomoko Nakanishi,[21,23,24] Mark Lathrop,[23,25] J. Brent Richards,[21,22,23,26] Regeneron Genetics Center, UKB Exome Sequencing Consortium, Marcus Jones,[1] Suganthi Balasubramanian,[1] William Salerno,[1] Alan R. Shuldiner,[1] Jonathan Marchini,[1] John D. Overton,[1] Lukas Habegger,[1] Michael N. Cantor,[1] Jeffrey G. Reid,[1] Aris Baras,[1,28] Goncalo R. Abecasis,[1,28,*] and Manuel A.R. Ferreira[1,28,*]

Summary

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) causes coronavirus disease 2019 (COVID-19), a respiratory illness that can result in hospitalization or death. We used exome sequence data to investigate associations between rare genetic variants and seven COVID-19 outcomes in 586,157 individuals, including 20,952 with COVID-19. After accounting for multiple testing, we did not identify any clear associations with rare variants either exome wide or when specifically focusing on (1) 13 interferon pathway genes in which rare deleterious variants have been reported in individuals with severe COVID-19, (2) 281 genes located in susceptibility loci identified by the COVID-19 Host Genetics Initiative, or (3) 32 additional genes of immunologic relevance and/or therapeutic potential. Our analyses indicate there are no significant associations with rare protein-coding variants with detectable effect sizes at our current sample sizes. Analyses will be updated as additional data become available, and results are publicly available through the Regeneron Genetics Center COVID-19 Results Browser.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1] causes coronavirus disease 2019 (COVID-19).[2] COVID-19 ranges in clinical presentation from asymptomatic infection to flu-like illness with respiratory failure, hyperactive immune responses, and death.[3–5] Known risk factors for severe disease include male sex, older age, ancestry, obesity, and underlying cardiovascular, renal, and respiratory diseases,[6–9] among others. Since the start

[1]Regeneron Genetics Center, 777 Old Saw Mill River Road, Tarrytown, NY 10591, USA; [2]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge CB2 0AA, UK; [3]Biogen, 300 Binney Street, Cambridge, MA 02142, USA; [4]Alnylam Pharmaceuticals, 675 West Kendall Street, Cambridge, MA 02142, USA; [5]Bristol Myers Squibb, Route 206 and Province Line Road, Princeton, NJ 08543, USA; [6]Takeda California, Inc., 9625 Towne Centre Drive, San Diego, CA 92121, USA; [7]AbbVie, Inc., 1 N. Waukegan Road, North Chicago, IL 60064, USA; [8]Pfizer, Inc., 1 Portland Street, Cambridge, MA 02139, USA; [9]Geisinger, Danville, PA 17822, USA; [10]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; [11]Institute for Genomic Medicine, Columbia University Irving Medical Center, New York, NY 10032, USA; [12]Department of Genetics and Development, Columbia University, New York, NY 10032, USA; [13]Division of Nephrology, Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA; [14]Roslin Institute, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, UK; [15]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK; [16]Genomics England, London EC1M 6BQ, UK; [17]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, Teviot Place, Edinburgh EH8 9AG, UK; [18]William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK; [19]Great Ormond Street Hospital for Children NHS Foundation Trust, London WC1N 3JH, UK; [20]Intensive Care Unit, Royal Infirmary of Edinburgh, 54 Little France Drive, Edinburgh EH16 5SA, UK; [21]Lady Davis Institute, Jewish General Hospital, Montréal, QC H3T 1E2, Canada; [22]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC H3A 0G4, Canada; [23]Department of Human Genetics, McGill University, Montréal, QC H3A 0G4, Canada; [24]Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan; [25]Canadian Centre for Computational Genomics, McGill University, Montréal, QC H3A 0G4, Canada; [26]Department of Twins Research, King's College London, London WC2R 2LS, UK
[27]These authors contributed equally
[28]Senior author
*Correspondence: manuel.ferreira@regeneron.com (M.A.R.F.), goncalo.abecasis@regeneron.com (G.R.A.)
https://doi.org/10.1016/j.ajhg.2021.05.017.

Table 1. Top associations between COVID-19 outcomes and protein-coding rare variants ($p < 5E{-}8$)

Gene	Variant[a]	Variant effect	Odds ratio (95% CI)	p value	N affected individuals with 0|1|2 copies of effect allele	N control individuals with 0|1|2 copies of effect allele	Effect allele frequency	Heterogeneity p value
COVID-19 positive versus COVID-19 negative or unknown								
ZC3HAV1	rs769102632	missense	26.72 (8.37, 85.38)	2.95E−8	13,950|7|0	401,218|8|0	0.00002	0.9517
FLNB	rs1256764500	missense	26.6 (8.25, 85.77)	3.97E−8	18,616|7|0	500,616|8|0	0.00001	0.4354
COVID-19 positive versus COVID-19 negative								
DISP3	burden	pLoF and deleterious missense with MAF $< 10^{-3}$	1.88 (1.51, 2.34)	2.26E−8	20,727|145|0	74,172|301|0	0.00234	0.9972
COVID-19 hospitalized versus COVID-19 negative or unknown								
WDR78	rs754119466	splice region	49.21 (13.61, 177.85)	2.81E−9	3,619|6|0	392,658|24|0	0.00004	1
TES	rs761377603	missense	38.91 (10.75, 140.9)	2.44E−8	4,555|5|0	511,328|23|0	0.00003	0.6601
MARK1	burden	pLoF variants with MAC = 1	40.19 (10.9, 148.1)	2.86E−8	4,473|5|0	530,595|34|0	0.00004	0.4035
SHC2	rs2287960	stop gained	42.94 (11.17, 165.02)	4.42E−8	4,237|5|0	483,826|17|0	0.00002	0.6742
COVID-19 severe versus COVID-19 negative or unknown								
TLR7[b]	burden	pLoF and missense variants with MAF $< 10^{-5}$	4.53 (2.64, 7.77)	4.28E−8	1,266|1|7	517,523|383|123	0.00062	0.7188

MAF, minor allele frequency; MAC, minor allele count; CI, confidence interval.
[a]Effect allele for individual variants was rs769102632:A, rs1256764500:G, rs754119466:G, rs761377603:T, and rs2287960:T. For burden tests, individuals were considered to have 0 copies of the effect allele if they were homozygous for the reference allele for all variants included in the burden test, 1 copy of the effect allele if they were heterozygous for at least one variant, and 2 copies if they were homozygous for the alternate allele for at least one variant.
[b]*TLR7* is located on the X chromosome. Hemizygous males are included in the N of individuals with two copies of the effect allele.

of the SARS-CoV-2 pandemic, host genetic analysis of common genetic variation among SARS-CoV-2 patients have identified at least 15 genome-wide significant loci that modulate COVID-19 susceptibility, including variants in/near *LZTFL1*, *IFNAR2*, and *DPP9*.[10–14] However, to date, there has been no exome-wide assessment of the contribution of rare coding genetic variation to COVID-19 disease susceptibility or severity through large population-based exome-wide association analyses.

To identify rare variants (RVs, minor allele frequency [MAF] < 1%) associated with COVID-19 susceptibility and severity, we received approval from institutional review boards (supplemental methods) and analyzed exome-wide sequencing data for 586,157 consented individuals from three studies (Geisinger Health System [GHS], Penn Medicine BioBank [PMBB], and UK Biobank [UKB]) across five continental ancestries (African, Admixed American, European, East Asian, and South Asian; Table S1). Of these, 20,952 had COVID-19, and among those, 4,928 (23.5%) were hospitalized and 1,304 (6.2%) had severe disease (i.e., requiring ventilation or resulting in death; Table S2). Using these data, we tested the association between RVs and seven COVID-19 outcomes: five related to disease susceptibility and two related to disease severity among individuals with COVID-19 (Table S3). In a separate paper,[13] we used these same phenotypes to validate the association with common risk variants reported in previous COVID-19 genome-wide association studies (GWASs),[10–12,14] thus demonstrating that our phenotypes are calibrated with those used in other studies.

For each phenotype, exome-wide association analyses were performed separately in each study and ancestry via REGENIE,[15] testing individual RVs (~7 million) and a burden of RVs in 18,886 protein-coding genes. The genomic inflation factor (λ_{GC}) for RVs was often <1 in individual studies, caused by a large proportion of variants having a minor allele count (MAC) of 0 in affected individuals (Table S4). In meta-analyses across studies and ancestries, we found no RV associations at a conservative $p < 9.6E{-}10$, which corresponds to a Bonferroni correction for the number of variants and traits tested. At a less conservative significance threshold of $p < 5E{-}8$, we found eight genes with RV associations (Table 1), of which, we highlight two with an established role in anti-viral responses. First, we highlight an association between higher risk of severe COVID-19

and a burden of ultra-rare (MAF < 0.001%) predicted loss-of-function (pLoF) and missense variants in the toll-like receptor 7 gene (*TLR7*; p = 4E−8; OR = 4.53; 95% CI = 2.64–7.77), consistent with relatively small exome-sequencing studies of males with severe COVID-19.[16,17] *TLR7* encodes a single-stranded viral RNA sensor that recognizes coronaviruses, including SARS-CoV-1, MERS, and most likely SARS-CoV-2,[18] and that activates the type-1 interferon pathway in COVID-19.[16] Second, we highlight an association between higher risk of COVID-19 and an ultra-rare missense variant in *ZC3HAV1* (rs769102632:A, MAF = 0.002%; p = 3E−8; OR = 26.7; 95% CI 8.37–85.38; Figure S1), a gene that encodes a zinc finger antiviral protein[19,20] that inhibits SARS-CoV-2 replication,[21] potentially by upregulating type I interferon responses.[22] Given the potential significance of this finding, we attempted to replicate the *ZC3HAV1* rs769102632:A association in an additional 6,223 individuals with COVID-19 with exome or whole-genome sequence data generated as part of the GenOMICC (n = 4,851),[11] Columbia University COVID-19 Biobank (n = 1,152), and Biobanque Quebec (n = 220)[23] studies. We found no carriers for this variant in these additional COVID-19 cases (Table S5) when we expected about four given the observed allele frequency in cases in our study (three and one carriers expected in individuals of African and European ancestry, respectively). Given these findings, we conclude that it is unlikely that there is a true association between rs532051930 and COVID-19 risk. Similarly, the association with a promoter variant in *EEF2* that we reported in an earlier version of these analyses[24] was considerably attenuated (from p = 6E−9 to 3E−6), consistent with a false-positive association.

Next, we addressed the possibility that associations with protein-coding RVs might help pinpoint target genes of common risk variants identified in GWASs of COVID-19. To this end, we focused on 281 genes located within 500 kb of the 15 common risk variants identified by the COVID-19 Host Genetics Initiative (HGI)[14] and asked whether there was any evidence for association between our five COVID-19 susceptibility outcomes and a burden of RVs in any of these genes. We considered associations with pLoF variants alone (M1 burden test) or pLoF together with deleterious missense variants (M3 burden test). No associations surpassed the Bonferroni significance threshold of 3.5E−6, which accounts for the 14,050 gene burden tests performed (281 genes × two burden tests × five allele frequency cut-offs × five susceptibility phenotypes; Table S6). As such, at current sample sizes, RV associations do not point to potential effector genes underlying associations between common variants and COVID-19.

We then examined the association with 13 genes in the interferon pathway,[25] given a previous report that deleterious RVs in these genes may be implicated in severe clinical outcomes.[25] Specifically, we examined whether there was any evidence for association between the COVID-19 hospitalization phenotype (4,928 affected individuals versus 558,763 control individuals) and the burden of rare (MAF < 0.1%, as reported by Zhang et al.[25]) pLoF variants (M1 burden test) or pLoF plus deleterious missense variants (M3 burden test) in these 13 genes. There were no significant associations with any gene, either individually or on aggregate (all burden tests with p > 0.05; Table 2). Further, these results were unchanged when testing severe cases of COVID-19 (n = 1,304) or when restricting the burden tests to include variants with an MAF < 1% or singleton variants (Table S7). Therefore, in alignment with a similar report,[23] we also found no evidence for an association between RVs in these 13 interferon-signaling genes.

Lastly, we performed the same analysis for an additional 32 genes that are involved in the etiology of SARS-CoV-2 infection (*ACE2*, *TMPRSS2*), encode therapeutic targets for COVID-19 obtained through the ClinicalTrials database (see web resources) (e.g., *IL6R*, *JAK1*), or have been implicated in other immune or infectious diseases through GWASs (e.g., *IL33*). After correcting for 1,600 burden tests performed (32 genes × five traits × five allele frequency thresholds × two burden tests; Bonferroni significance threshold p < 3.1E−5), there were no significant associations with deleterious RVs for this group of therapeutic target genes for COVID-19 (Table S8).

There are caveats to be considered when interpreting results from this study. First, the five continental ancestry groups considered in our analysis included a small number of individuals with admixed ancestry (specifically, those with two continental ancestries with a likelihood > 0.3; see supplemental methods). For example, individuals with admixed African and European ancestry were included in our analysis of African ancestry. This was done to maximize the number and ancestral diversity of the samples included in our analysis and was adequately controlled for in the association analyses carried out with the whole-genome regression approach implemented in REGENIE (test statistics were not inflated). Second, the burden tests we performed were not designed to identify associations with genes that harbor both risk-increasing and risk-lowering rare variants and are expected to provide limited power in these instances. Other approaches have been developed for these situations, such as SKAT[26]/SKAT-O.[27] However, we have not tested the robustness of these alternative burden tests in the context of multi-ancestry meta-analyses, so we opted against applying them in this study. Third, we used a stringent Bonferroni correction to define significance thresholds that account for multiple testing, which are most likely conservative, given the high correlation between traits and burden tests performed.

In summary, we explored the role of rare coding variants on COVID-19 outcomes on the basis of exome-sequence data, capturing genetic variation not assayed by array genotyping or imputation. We did not find any convincing associations with current sample sizes but will continue to expand our analyses and update results periodically at

Table 2. Burden associations among interferon signaling genes

Variants included in burden test	Gene	Odds ratio (95% CI)	p value	N affected individuals with RR|RA|AA genotype[a]	N control individuals with RR|RA|AA genotype[a]	AAF	Heterogeneity p value
pLoF, MAF < 0.1%	*IFNAR1*	1.46 (0.51, 4.17)	0.4786	4,775|5|0	549,164|374|0	0.00034	0.9111
	IFNAR2	1.96 (0.91, 4.19)	0.0844	4,920|8|0	558,068|695|0	0.00062	0.0964
	IKBKG[b]	0.51 (0.04, 6.57)	0.6048	4,394|0|0	500,582|32|10	0.00005	0.9584
	IRF3	0.91 (0.39, 2.11)	0.8293	4,924|3|1	558,279|483|1	0.00043	0.6339
	IRF7	1.15 (0.57, 2.31)	0.6975	4,920|8|0	557,892|871|0	0.00078	0.5267
	IRF9	0.36 (0.02, 6.96)	0.5024	4,478|0|0	530,571|58|0	0.00005	0.9996
	STAT1	0.36 (0.01, 19.89)	0.6207	4,394|0|0	500,584|40|0	0.00004	0.9996
	STAT2	0.36 (0.07, 1.91)	0.2311	4,644|0|0	541,214|144|0	0.00013	1.0000
	TBK1	0.36 (0.04, 3.13)	0.3553	4,478|0|0	530,539|90|0	0.00008	0.9995
	TICAM1	0.81 (0.14, 4.73)	0.8160	4,477|1|0	530,454|175|0	0.00016	0.7587
	TLR3	1.56 (0.47, 5.13)	0.4656	4,924|4|0	558,457|306|0	0.00027	0.7039
	TRAF3	0.37 (0.0, 217.91)	0.7576	4,394|0|0	500,597|27|0	0.00003	1.0000
	UNC93B1	0.77 (0.28, 2.06)	0.5974	4,641|3|0	540,929|429|0	0.00040	0.9294
	all autosomal genes	0.81 (0.56, 1.18)	0.2709	4,655|23|0	514,810|3,219|0	0.00320	0.9492
pLoF and missense predicted deleterious, MAF < 0.1%	*IFNAR1*	1.51 (0.71, 3.18)	0.2831	4,918|10|0	557,991|772|0	0.00069	0.8283
	IFNAR2	1.87 (0.88, 3.97)	0.1021	4,920|8|0	558,045|718|0	0.00064	0.0862
	IKBKG[b]	1.48 (0.18, 12.34)	0.7184	4,393|1|0	500,544|70|10	0.00009	0.6366
	IRF3	0.9 (0.42, 1.92)	0.7778	4,923|4|1	558,128|634|1	0.00057	0.7436
	IRF7	1.15 (0.67, 1.96)	0.6102	4,914|14|0	557,238|1,525|0	0.00137	0.3523
	IRF9	0.36 (0.02, 6.96)	0.5024	4,478|0|0	530,571|58|0	0.00005	0.9996
	STAT1	0.35 (0.08, 1.49)	0.1563	4,762|0|0	547,803|231|0	0.00021	1.0000
	STAT2	1.26 (0.73, 2.2)	0.4089	4,909|19|0	557,153|1,609|1	0.00145	0.7935
	TBK1	1.0 (0.54, 1.85)	0.9951	4,917|11|0	557,567|1,195|1	0.00107	0.6983
	TICAM1	0.8 (0.14, 4.66)	0.8084	4,477|1|0	530,451|178|0	0.00017	0.7558
	TLR3	0.74 (0.49, 1.11)	0.1396	4,911|17|0	556,016|2,745|2	0.00245	0.8319
	TRAF3	1.7 (0.44, 6.62)	0.4431	4,778|2|0	549,284|254|0	0.00023	0.1923
	UNC93B1	0.92 (0.56, 1.5)	0.7309	4,913|15|0	557,079|1,684|0	0.00151	0.9180
	all autosomal genes	0.94 (0.76, 1.17)	0.5835	4,590|88|0	507,793|10,233|3	0.00990	0.5285

Association between the phenotype COVID-19 positive hospitalized versus COVID-19 negative or unknown and 13 genes (12 autosomal) related to interferon signaling that were recently reported to contain rare (MAF < 0.1%) deleterious variants in individuals with severe COVID-19.[25] AAF, alternative allele frequency; CI, confidence interval.

[a]RR, individuals who have genotype reference/reference for all variants included in burden test; RA, individuals who have genotype reference/alternate for at least one variant; AA, individuals who have genotype alternate/alternate for at least one variant.

[b]*IKBKG* is located on the X chromosome. Hemizygous males are included in the N of individuals with two copies of the effect allele.

the Regeneron Genetics Center COVID-19 Results Browser (web resources).

Data and code availability

All genotype-phenotype association results reported in this study are available for download and browsing via the RGC's COVID-19 Results Browser (https://rgc-covid19.regeneron.com). Data access and use is limited to research purposes in accordance with the Terms of Use (https://rgc-covid19.regeneron.com/terms-of-use).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.05.017.

Declaration of interests

J.A.K., J.E.H., A.D., D.S., N.B., A.Y., A.M., R.L., E.M., X.B., D.S., F.S.P.K., J.D.B., C.O'D., A.J.M., D.A.T., A.H.L., J.M., K.W., L.G., S.E.M., H.M.K., L.D., E.S., M.J., S.B., K.S.M., W.J.S., A.R.S., A.E.L., J.M., J.O., L.H., M.N.C., J.G.R., A.B., G.R.A., and M.A.F. are current employees and/or stockholders of Regeneron Genetics Center or Regeneron Pharmaceuticals. X.Z., S.E., and J.W.D. are employees of AbbVie and may hold stock in AbbVie. Financial support for this research was provided by AbbVie through the UKB Exome Sequencing Consortium. AbbVie participated in the interpretation of data, review, and approval of the publication. P.N. and M.M.P. are employees and stockholders of Alnylam Pharmaceuticals. J.B.R. has served as an advisor to GlaxoSmithKline and Deerfield Capital and these agencies had no role in the design, implementation, or interpretation of this study. S.S., E.W., A.C.P., and E.N.S. are employed by Takeda. S.S. holds shares in Takeda and Janssen. All other authors declare no competing interests.

Received: February 18, 2021
Accepted: May 24, 2021
Published: June 3, 2021

Web resources

BWA software (v.0.7.17), http://bio-bwa.sourceforge.net
ClinicalTrials database, clinicaltrials.gov
METAL software, https://github.com/statgen/METAL
PLINK (v.1.90b6.21), https://www.cog-genomics.org/plink2/
Picard software (v.1.141), https://broadinstitute.github.io/picard/
Regeneron Genetics Center COVID-19 Results Browser, https://rgc-covid19.regeneron.com
REGENIE software, https://github.com/rgcgithub/regenie
Samtools (v.1.7), http://www.htslib.org
WeCall software (v.1.1.2), https://github.com/Genomicsplc/wecall

References

1. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. N. Engl. J. Med. *382*, 727–733.
2. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. *5*, 536–544.
3. Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., Liu, L., Shan, H., Lei, C.L., Hui, D.S.C., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. N. Engl. J. Med. *382*, 1708–1720.
4. Kimball, A., Hatfield, K.M., Arons, M., James, A., Taylor, J., Spicer, K., Bardossy, A.C., Oakley, L.P., Tanwar, S., Chisty, Z., et al. (2020). Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility - King County, Washington, March 2020. MMWR Morb. Mortal. Wkly. Rep. *69*, 377–381.
5. Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.Y., Chen, L., and Wang, M. (2020). Presumed Asymptomatic Carrier Transmission of COVID-19. JAMA *323*, 1406–1407.
6. Richardson, S., Hirsch, J.S., Narasimhan, M., Crawford, J.M., McGinn, T., Davidson, K.W., Barnaby, D.P., Becker, L.B., Chelico, J.D., Cohen, S.L., et al. (2020). Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA *323*, 2052–2059.
7. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet *395*, 1054–1062.
8. Cummings, M.J., Baldwin, M.R., Abrams, D., Jacobson, S.D., Meyer, B.J., Balough, E.M., Aaron, J.G., Claassen, J., Rabbani, L.E., Hastie, J., et al. (2020). Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. Lancet *395*, 1763–1770.
9. Atkins, J.L., Masoli, J.A.H., Delgado, J., Pilling, L.C., Kuo, C.L., Kuchel, G.A., and Melzer, D. (2020). Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort. J. Gerontol. A Biol. Sci. Med. Sci. *75*, 2224–2230.
10. Shelton, J.F., Shastri, A.J., Ye, C., Weldon, C.H., Filshtein-Sonmez, T., Coker, D., Symons, A., Esparza-Gordillo, J., Aslibekyan, S., Auton, A.; and 23andMe COVID-19 Team (2021). Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. Nat. Genet. https://doi.org/10.1038/s41588-021-00854-7.
11. Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A.D., Rawlik, K., Pasko, D., Walker, S., Parkinson, N., Fourman, M.H., Russell, C.D., Furniss, J., et al. (2020). Genetic mechanisms of critical illness in Covid-19. Nature *591*, 92–98.
12. Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J., Prati, D., Baselli, G., Asselta, R., et al. (2020). Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N. Engl. J. Med. *383*, 1522–1534.
13. Horowitz, J.E., Kosmicki, J.A., Damask, A., Sharma, D., Roberts, G.H.L., Justice, A.E., Banerjee, N., Coignet, M.V., Yadav, A., Leader, J.B., et al. (2020). Common genetic variants identify therapeutic targets for COVID-19 and individuals at high risk of severe disease. medRxiv, 2020.12.14.20248176.
14. Ganna, A.; and The COVID-19 Host Genetics Initiative (2021). Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. medRxiv, 2021.03.10.21252820.
15. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber,

M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. Published online May 20, 2021. https://doi.org/10.1038/s41588-021-00870-7.

16. van der Made, C.I., Simons, A., Schuurs-Hoeijmakers, J., van den Heuvel, G., Mantere, T., Kersten, S., van Deuren, R.C., Steehouwer, M., van Reijmersdal, S.V., Jaeger, M., et al. (2020). Presence of Genetic Variants Among Young Men With Severe COVID-19. JAMA *324*, 663–673.
17. Fallerini, C., Daga, S., Mantovani, S., Benetti, E., Picchiotti, N., Francisci, D., Paciosi, F., Schiaroli, E., Baldassarri, M., Fava, F., et al. (2021). Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: findings from a nested case-control study. eLife *10*, e67569.
18. Moreno-Eutimio, M.A., López-Macías, C., and Pastelin-Palacios, R. (2020). Bioinformatic analysis and identification of single-stranded RNA sequences recognized by TLR7/8 in the SARS-CoV-2, SARS-CoV, and MERS-CoV genomes. Microbes Infect. *22*, 226–229.
19. Gao, G., Guo, X., and Goff, S.P. (2002). Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. Science *297*, 1703–1706.
20. Zhu, Y., Chen, G., Lv, F., Wang, X., Ji, X., Xu, Y., Sun, J., Wu, L., Zheng, Y.T., and Gao, G. (2011). Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. Proc. Natl. Acad. Sci. USA *108*, 15834–15839.
21. Nchioua, R., Kmiec, D., Müller, J.A., Conzelmann, C., Groß, R., Swanson, C.M., Neil, S.J.D., Stenger, S., Sauter, D., Münch, J., et al. (2020). SARS-CoV-2 Is Restricted by Zinc Finger Antiviral Protein despite Preadaptation to the Low-CpG Environment in Humans. MBio *11*, e01930-20.
22. Zhang, B., Goraya, M.U., Chen, N., Xu, L., Hong, Y., Zhu, M., and Chen, J.L. (2020). Zinc Finger CCCH-Type Antiviral Protein 1 Restricts the Viral Replication by Positively Regulating Type I Interferon Response. Front. Microbiol. *11*, 1912.
23. Povysil, G., Butler-Laporte, G., Shang, N., Weng, C., Khan, A., Alaamery, M., Nakanishi, T., Zhou, S., Forgetta, V., Eveleigh, R., et al. (2021). Failure to replicate the association of rare loss-of-function variants in type I IFN immunity genes with severe COVID-19. J. Clin. Invest. 2020.12.18.20248226. https://doi.org/10.1172/JCI147834.
24. Kosmicki, J.A., Horowitz, J.E., Banerjee, N., Lanche, R., Marcketta, A., Maxwell, E., Bai, X., Sun, D., Backman, J.D., Sharma, D., et al. (2021). A catalog of associations between rare coding variants and COVID-19 outcomes. medRxiv, 2020.10.28.20221804.
25. Zhang, Q., Bastard, P., Liu, Z., Le Pen, J., Moncada-Velez, M., Chen, J., Ogishi, M., Sabli, I.K.D., Hodeib, S., Korol, C., et al. (2020). Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science *370*, eabd4570.
26. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
27. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., Lin, X.; and NHLBI GO Exome Sequencing Project—ESP Lung Project Team (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.

The American Journal of Human Genetics, Volume *108*

Supplemental information

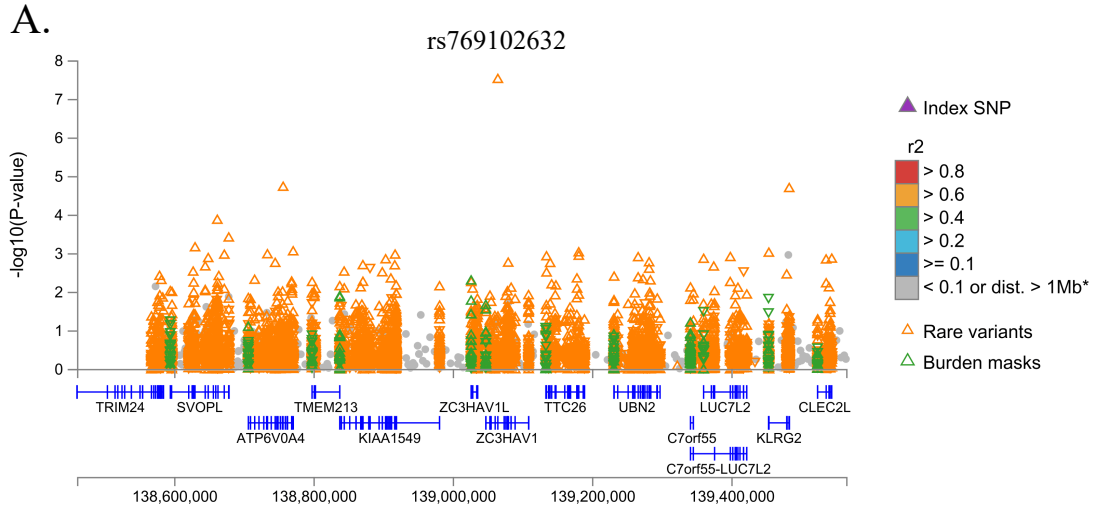
Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals

Jack A. Kosmicki, Julie E. Horowitz, Nilanjana Banerjee, Rouel Lanche, Anthony Marcketta, Evan Maxwell, Xiaodong Bai, Dylan Sun, Joshua D. Backman, Deepika Sharma, Fabricio S.P. Kury, Hyun M. Kang, Colm O'Dushlaine, Ashish Yadav, Adam J. Mansfield, Alexander H. Li, Kyoko Watanabe, Lauren Gurski, Shane E. McCarthy, Adam E. Locke, Shareef Khalid, Sean O'Keeffe, Joelle Mbatchou, Olympe Chazara, Yunfeng Huang, Erika Kvikstad, Amanda O'Neill, Paul Nioi, Meg M. Parker, Slavé Petrovski, Heiko Runz, Joseph D. Szustakowski, Quanli Wang, Emily Wong, Aldo Cordova-Palomera, Erin N. Smith, Sandor Szalma, Xiuwen Zheng, Sahar Esmaeeli, Justin W. Davis, Yi-Pin Lai, Xing Chen, Anne E. Justice, Joseph B. Leader, Tooraj Mirshahi, David J. Carey, Anurag Verma, Giorgio Sirugo, Marylyn D. Ritchie, Daniel J. Rader, Gundula Povysil, David B. Goldstein, Krzysztof Kiryluk, Erola Pairo-Castineira, Konrad Rawlik, Dorota Pasko, Susan Walker, Alison Meynert, Athanasios Kousathanas, Loukas Moutsianas, Albert Tenesa, Mark Caulfield, Richard Scott, James F. Wilson, J. Kenneth Baillie, Guillaume Butler-Laporte, Tomoko Nakanishi, Mark Lathrop, J. Brent Richards, Regeneron Genetics Center, UKB Exome Sequencing Consortium, Marcus Jones, Suganthi Balasubramanian, William Salerno, Alan R. Shuldiner, Jonathan Marchini, John D. Overton, Lukas Habegger, Michael N. Cantor, Jeffrey G. Reid, Aris Baras, Goncalo R. Abecasis, and Manuel A.R. Ferreira

SUPPLEMENTARY FIGURES

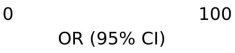
Figure S1: Association of a rare missense variant in *ZC3HAV1* and COVID-19.

A.



B.

STUDY	N CASES RR|RA|AA	N CONTROLS RR|RA|AA	OR [95% CI]	PVALUE	AAF
UKB_Freeze_450_AFR	578|3|0	8178|2|0	27.928 [4.444,175.515]	3.8e-04	0.00029
UKB_Freeze_450_EUR	13372|4|0	393040|6|0	25.956 [5.8,116.159]	2.1e-05	1e-05
META	**13950|7|0**	**401218|8|0**	**26.725 [8.366,85.377]**	**2.9e-08**	**2e-05**



Association between an ultra-rare missense variant in *ZC3HAV1* (rs769102632:A) and higher risk of COVID-19. (A) Regional association plot centered on rs769102632. Orange triangles: individual rare variants (MAF<0.5%). Green squares: burden tests. Grey circles: individual common variants (MAF>0.5%). (B) Forest plot showing association in the two individual datasets included in the meta-analysis of this variant.

SUPPLEMENTARY TABLES

Tables S1 to S8 are provided in a separate Excel document.

Table S1. Demographics and clinical characteristics of study participants.

Table S2. Breakdown of COVID-19 status across the four studies included in the analysis.

Table S3. Definitions used for the seven COVID-19 phenotypes analyzed.

Table S4. Genomic inflation factor (λ_{GC}) observed in the analysis of exome sequence variants for each of the eight phenotypes tested.

Table S5. No carriers of the rare rare missense variant rs769102632 in *ZC3HAV1* were observed in an additional 6,223 individuals with COVID-19.

Table S6. Nominally-significant associations (P<0.05) among 14,050 burden tests performed across 281 genes located in 15 susceptibility loci identified by the COVID-19 Host Genetics Initiative.

Table S7. Results from burden association tests for 13 genes related to interferon signaling and recently reported to contain rare (MAF<0.1%), deleterious variants in patients with severe COVID-19.

Table S8. Results from burden association tests for an additional 32 genes that are involved in the etiology of SARS-CoV-2, encode therapeutic targets or have been implicated in other immune or infectious diseases through GWAS.

SUPPLEMENTARY METHODS

Participating Studies

Geisinger Health System (GHS). The GHS MyCode Community Health Initiative study has been described previously [1]. Briefly, the GHS study is a health system-based cohort from central and eastern Pennsylvania (USA) with ongoing recruitment since 2006. A subset of 144,182 MyCode participants sequenced as part of the GHS-Regeneron Genetics Center DiscovEHR partnership were included in this study. All subjects consented to participation and the analysis was approved by the Geisinger Institutional Review Board under project number 2006-0258. Information on COVID-19 outcomes were obtained through GHS's COVID-19 registry. Patients were identified as eligible for the registry based on relevant lab results and ICD-10 diagnosis codes; patient charts were then reviewed to confirm COVID-19 diagnoses. The registry contains data on outcomes, comorbidities, medications, supplemental oxygen use and ICU admissions.

Penn Medicine BioBank (PMBB) study. PMBB study participants are recruited through the University of Pennsylvania Health System, which enrolls participants during hospital or clinic visits. After providing consent, participants donate blood or tissue and allow access to EHR information[2]. The PMBB COVID-19 registry consists of patients who have positive qPCR testing for SARS-COV-2. We then used electronic health records to classify COVID-19 patients into hospitalized and severe (ventilation or death) categories and the study was approved by the University of Pennsylvania Institutional Review Board (protocol #813913).

UK Biobank (UKB) study. We studied the host genetics of SARS-CoV-2 infection in participants of the UK Biobank study, which took place between 2006 and 2010 and includes approximately 500,000 adults aged 40-69 at recruitment[3]. In collaboration with UK health authorities, the UK Biobank has made available regular updates on COVID-19 status for all participants, including results from four main data types: qPCR test for SARS-CoV-2, anonymized electronic health records, primary care and death registry data. We report results based on the 8 March 2021 data refresh and excluded from the analysis 28,547 individuals with a death registry event prior to 2020. The study was approved by the research ethics committee under approval number 11/NW/0382.

COVID-19 phenotypes used for genetic association analyses

We grouped participants from each study into three broad COVID-19 disease categories (**Table S2**): (i) positive – those with a positive qPCR or serology test for SARS-CoV-2, or a COVID-19-related ICD10 code (U07), hospitalization or death; (ii) negative – those with only negative qPCR or serology test results for SARS-CoV-2 and no COVID-19-related ICD10 code (U07), hospitalization or death; and (iii) unknown – those with no qPCR or serology test results and no COVID-19-related ICD10 code (U07), hospitalization or death. We then used these broad COVID-19 disease categories, in addition to hospitalization and disease severity information, to create seven COVID-19-related phenotypes for genetic association analyses, as detailed in **Table S3**.

Array genotyping

Genotyping was performed on one of four SNP array types: Illumina OmniExpress Exome array (OMNI; 59345 samples from GHS), Illumina Global Screening Array (GSA; PMBB and 82,527 samples from GHS), Applied Biosystems UK BiLEVE Axiom Array (49,950 samples from UKB), or Applied Biosystems UK Biobank Axiom Array (438,427 samples from UKB). We retained variants with a minor allele frequency (MAF) >1%, <10% missingness, Hardy-Weinberg equilibrium test *P*-value>10^{-15}. Array data were then used: (i) to define ancestry subsets; and (ii) as part of the exome-wide association analyses carried out in REGENIE (see below).

Exome sequencing

Sample Preparation and Sequencing. Genomic DNA samples normalized to approximately 16 ng/ul were transferred to the Regeneron Genetics Center from the UK Biobank in 0.5ml 2D matrix tubes (Thermo Fisher Scientific) and stored in an automated sample biobank (LiCONiC Instruments) at -80°C prior to sample preparation. Exome capture was completed using a high-throughput, fully-automated approach developed at the Regeneron Genetics Center. Briefly, DNA libraries were created by enzymatically shearing 100ng of genomic DNA to a mean fragment size of 200 base pairs using a custom NEBNext Ultra II FS DNA library prep kit (New England Biolabs) and a common Y-shaped adapter (Integrated DNA Technologies [IDT]) was ligated to all DNA libraries. Unique, asymmetric 10 base pair barcodes were added to the DNA fragment during library amplification with KAPA HiFi polymerase (KAPA Biosystems) to facilitate multiplexed exome capture and sequencing. Equal amounts of sample were pooled prior to

overnight exome capture, approximately 16 hours, with either (i) a slightly modified version of IDT's xGen probe library (for UKB, PMBB and 81,620 samples of GHS); or (ii) NimbleGen VCRome (58,856 samples of GHS). Captured fragments were bound to streptavidin-coupled Dynabeads (Thermo Fisher Scientific) and non-specific DNA fragments removed through a series of stringent washes using the xGen Hybridization and Wash kit according to the manufacturer's recommended protocol (Integrated DNA Technologies). The captured DNA was PCR amplified with KAPA HiFi and quantified by qPCR with a KAPA Library Quantification Kit (KAPA Biosystems). The multiplexed samples were pooled and then sequenced using: (i) for UKB samples – 75 bp paired-end reads with two 10 base pair index reads on the Illumina NovaSeq 6000 platform using S2 or S4 flow cells; (ii) for GHS samples captured with VCRome – 75 bp paired-end reads with two 8 bp index reads on the Illumina HiSeq 2500; (iii) for GHS captured with IDT – two 8 bp index reads on the Illumina HiSeq 2500 or two 10 bp index reads on the Illumina NovaSeq 6000 on S4 flow cells; (iv) for UPENN-PMBB – two 10 bp index reads on the Illumina NovaSeq 6000 on S4 flow cells.

Variant calling and quality control. Sample read mapping and variant calling, aggregation and quality control were performed via the SPB protocol described in Van Hout et al. [4]. Briefly, for each sample, NovaSeq WES reads are mapped with BWA MEM to the hg38 reference genome. Small variants are identified with WeCall and reported as per-sample gVCFs. These gVCFs are aggregated with GLnexus into a joint-genotyped, multi-sample VCF (pVCF). SNV genotypes with read depth (DP) less than seven and indel genotypes with read depth less than ten are changed to no-call genotypes. After the application of the DP genotype filter, a variant-level allele balance filter is applied, retaining only variants that meet either of the following criteria: (i) at least one homozygous variant carrier or (ii) at least one heterozygous variant carrier with an allele balance (AB) greater than the cutoff (AB ³ 0.15 for SNVs and AB ³ 0.20 for indels).

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

from the analysis of variants with a MAF<1%; (v) transmitted singletons. The model was then trained on up to 30 available WeCall/GLnexus site quality metrics, including, for example, allele balance and depth of coverage. We split the data into training (80%) and test (20%) sets. We performed a grid search with 5-fold cross-validation on the training set to identify the hyperparameters that return the highest accuracy during cross-validation, which are then applied to the test set to confirm accuracy. This approach identified as low-quality a total of 7 million variants in the UKB study (86% in the buffer region), 7.2 million across the two GHS datasets (IDT and VCRome; 84% in the buffer region) and 1.1 million in the PMBB study (88% in the buffer region). These variants were removed from analysis in the respective studies.

Gene burden masks. Briefly, for each gene region as defined by Ensembl [5], genotype information from multiple rare coding variants was collapsed into a single burden genotype, such that individuals who were: (i) homozygous reference (Ref) for all variants in that gene were considered homozygous (RefRef); (ii) heterozygous for at least one variant in that gene were considered heterozygous (RefAlt); (iii) and only individuals that carried two copies of the alternative allele (Alt) of the same variant were considered homozygous for the alternative allele (AltAlt). We did not phase rare variants; compound heterozygotes, if present, were considered heterozygous (RefAlt). We did this separately for four classes of variants: (i) predicted loss of function (pLoF), which we refer to as an "M1" burden mask; (ii) pLoF or missense ("M2"); (iii) pLoF or missense variants predicted to be deleterious by 5/5 prediction algorithms ("M3"); (iv) pLoF or missense variants predicted to be deleterious by 1/5 prediction algorithms ("M4"). Variants were annotated using SnpEff 4.3[6] and the most severe consequence for each variant was chosen, considering complete protein-coding transcripts for each gene. The following variants were considered to be pLoF variants: frameshift-causing indels, variants affecting splice acceptor and donor sites, variants leading to stop gain, stop loss and start loss. The five missense deleterious algorithms used were SIFT [7], PolyPhen2 (HDIV), PolyPhen2 (HVAR) [8], LRT [9], and MutationTaster [10]. For each gene, and for each of these four groups, we considered five separate burden masks, based on the frequency of the alternative allele of the variants that were screened in that group: <1%, <0.1%, <0.01%, <0.001% and singletons only. Each burden mask was then tested for association with the same approach used for individual variants (see below).

Genetic association analyses

Association analyses in each study were performed using the genome-wide Firth logistic regression test implemented in REGENIE [11]. In this implementation, Firth's approach is applied when the p-value from standard logistic regression score test is below 0.05. As the Firth penalty (*i.e.,* Jeffrey's invariant prior) corresponds to a data augmentation procedure where each observation is split into a case and a control with different weights, it can handle variants with no minor alleles among cases. With no covariates, this corresponds to adding 0.5 in every cell of a 2x2 table of allele counts versus case-control status.

In the UKB study, we included in step 1 of REGENIE (*i.e.* prediction of individual trait values based on the genetic data) array variants with a minor allele frequency (MAF) >1%, <10% missingness, Hardy-Weinberg equilibrium test *P*-value>10^{-15} and linkage-disequilibrium (LD) pruning (1000 variant windows, 100 variant sliding windows and r^2<0.9). In the GHS and PMBB studies we instead used exome (not array) variants in step 1. We did this in the GHS study because two different exome capture technologies (IDT and VCRome) were used to sequence the GHS samples, and so it was important to capture in step 1 of REGENIE any differences in exome sequencing performance between IDT and VCRome. For the PMBB study, array data were not yet available for about 40K samples, and so we used exome data for step 1 to maximize the sample size available for analysis. We excluded from step 1 any SNPs with high inter-chromosomal LD, in the major histo-compatibility (MHC) region, or in regions of low complexity.

The association model used in step 2 of REGENIE included as covariates (i) age, age^2, sex, age-by-sex and age^2-by-sex; (ii) 10 ancestry-informative principal components (PCs) derived from the analysis of a 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; exome for PMBB) data generated separately for each ancestry; (iii) an indicator for exome sequencing batch (GHS: two IDT batches, one VCRome batch; UKB: six IDT batches); and (iv) 20 PCs derived from the analysis of exome variants with a MAF between 2.6x10^{-5} (roughly corresponding to a minor allele count [MAC] of 20) and 1% also generated separately for each ancestry. We corrected for PCs built from rare variants because previous studies demonstrated PCs derived from common variants do not adequately correct for fine-scale population structure [12, 13].

Within each study, association analyses were performed separately for different continental ancestries defined based on the array data: African (AFR), Admixed American (AMR), European (EUR) and South Asian (SAS). We determined continental ancestries by projecting each sample onto reference principal components calculated from the HapMap3 reference panel. Briefly, we merged our samples with HapMap3 samples and kept only SNPs in common between the two datasets. We further excluded SNPs with MAF<10%, genotype missingness >5% or Hardy-Weinberg Equilibrium test p-value < 10^{-5}. We calculated PCs for the HapMap3 samples and projected each of our samples onto those PCs. To assign a continental ancestry group to each non-HapMap3 sample, we trained a kernel density estimator (KDE) using the HapMap3 PCs and used the KDEs to calculate the likelihood of a given sample belonging to each of the five continental ancestry groups. When the likelihood for a given ancestry group was >0.3, the sample was assigned to that ancestry group. When two ancestry groups had a likelihood >0.3, we arbitrarily assigned AFR over EUR (N_{GHS} = 36 [0.9%], N_{UKB} = 56 [0.6%], $N_{UPENN\text{-}PMBB}$ = 7 [0.1%]), AMR over EUR (N_{GHS} = 455 [22.5%], N_{UKB} = 436 [47.8%], $N_{UPENN\text{-}PMBB}$ = 138 [23.5%]), AMR over EAS (N_{GHS} = 2 [0.05%], N_{UKB} = 2 [0.2%], $N_{UPENN\text{-}PMBB}$ = 1 [0.2%]), SAS over EUR (N_{GHS} = 32 [7.8%], N_{UKB} = 592 [9.6%], $N_{UPENN\text{-}PMBB}$ = 36 [6.3%]), and AMR over AFR (N_{GHS} = 192 [9.5%], N_{UKB} = 51 [5.6%], $N_{UPENN\text{-}PMBB}$ = 77 [13.1%]). Samples were excluded from analysis if no ancestry likelihoods were >0.3, or if more than three ancestry likelihoods were > 0.3 (N_{GHS} = 821, N_{UKB} = 1205, $N_{UPENN\text{-}PMBB}$ = 384).

Results were subsequently meta-analyzed across studies and ancestries using an inverse variance-weighed fixed-effects meta-analysis.

Frequency of *ZC3HAV1* rare missense variant in COVID-19 cases from independent studies

To help understand if the association between COVID-19 risk and rs769102632 in *ZC3HAV1* was likely to be a true-positive association, we determine its frequency in 6,223 cases from three additional studies.

GenOMICC (n=4,851). Individuals with severe COVID-19 were ascertained as described previously[14]. DNA samples were then whole-genome sequenced on the Illumina NovaSeq 6000 platform, aligned to the human reference genome hg38 and variant called to GVCF stage on the

DRAGEN pipeline (software v01.011.269.3.2.22, hardware v01.011.269) at Genomics England. rs769102632 +/-50bp was genotyped with the GATK GenotypeGVCFs tool v4.1.8.1 and filtered to minimum depth 8X. Ancestry for individuals with array genotyping (n=2,048) was inferred using ADMIXTURE[15] populations defined in 1000 Genomes[16]. When one individual had a probability > 80% of pertaining to one ancestry, then the individual was assigned to this ancestry (n=1,837), otherwise the individual was considered to be of admixed ancestry (n=211), as performed in the Million veteran program [17]. Of the remaining samples (n=3,014), Somalier v0.2.12[18] was used to estimate ancestry from the whole-genome sequencing data: 2,606 samples could be confidently (³92.5% probability) assigned to a population, while the remaining 408 were assigned to admixed ancestry.

Columbia University COVID-19 biobank (n=1,152). This cohort has previously been described in detail[19]. Briefly, 1,152 COVID-19 patients that were treated for COVID-19 at the Columbia University Irving Medical Center were recruited to the Columbia University COVID-19 Biobank between March and May 2020. All patients had PCR-confirmed SARS-CoV-2 infection and the vast majority had severe COVID-19 requiring hospitalization. For all cases, exomes were captured with the IDT xGen Exome Research Panel V1.0 and sequenced on Illumina's NovaSeq 6000 platform with 150 bp paired-end reads according to standard protocols. All cases were processed with the same bioinformatic pipeline for variant calling. In brief, reads were aligned to human reference GRCh37 using DRAGEN and duplicates were marked with Picard. Variants were called according to the Genome Analysis Toolkit (GATK) Best Practices recommendations v3.66[20]. Finally, variants were annotated with ClinEff[6] and the IGM's in-house tool ATAV[21]. A centralized database was used to store variant and per site coverage data for all samples enabling well controlled analyses without the need of generating jointly called VCF files (see Ren et al. 2021 for details[21]). For each patient, we performed ancestry classification into one of the six major ancestry groups (European, African, Latin, East Asian, South Asian and Middle Eastern) using a neural network trained on a set of samples with known ancestry labels. We used a 50% probability cut-off to assign an ancestry label to each sample and labeled samples that did not reach 50% for any of the ancestral groups as "Admixed". We only included samples that had at least 90% of the consensus coding sequence (CCDS release 20[22]) covered at ≥ 10x and ≤ 3% contamination levels according to VerifyBamID[23]. Additionally, we removed samples with a discordance between self-declared and sequence-derived gender and samples with an

inferred relationship of second-degree or closer according to KING[24]. All cases had at least 10x coverage at the position of rs769102632.

Biobanque Québec Covid-19 (n=220). The Biobanque Québec COVID-19 (www.BQC19.ca) is a provincial biobank prospectively enrolling patients with suspected COVID-19, or COVID-19 confirmed through SARS-CoV-2 PCR testing and was previously described[19]. For this study, we used results from patients with available WGS data and who were recruited at the Jewish General Hospital (JGH) in Montreal. The JGH is a university affiliated hospital serving a large multi-ethnic adult population and the Québec government designated the JGH as the primary COVID-19 reference center early in the pandemic. In total, Biobanque Quebec contained 533 participants with WGS, including 62 cases of COVID-19 who required invasive ventilatory support (BiPAP, high flow oxygen, or endotracheal intubation) or died, 128 COVID-19 patients who were hospitalized but did not require invasive ventilatory support, 30 individuals with COVID-19 did not require hospitalization, and 313 SARS-CoV-2 PCR-negative participants. Using genetic PCAs derived from genome-wide genotyping, 76% of participants were of European ancestry, 9% were of African ancestry, 7% were of east Asian ancestry, and 5% were of south Asian ancestry. We performed WGS at a mean depth of 30x on all individuals using Illumina's Novaseq 6000 platform (Illumina, San Diego, CA, USA). Sequencing results were analyzed using the McGill Genome Center bioinformatics pipelines[25], in accordance with Genome Analysis Toolkit (GATK) best practices recommendations[20]. Reads were aligned to the GRCh38 reference genome. Variant quality control was performed using the variantRecalibrator and applyVQSR functions from GATK.

Acknowledgements

This research has been 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. Whole genome sequencing of the Biobanque Québec Covid-19 cohort was funded by the CanCOGeN HostSeq project. The Richards research group is supported by the Canadian Institutes of Health Research (CIHR), the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK and the

Fonds de Recherche Québec Santé (FRQS). G.B.L. is supported by a joint research fellowship from Quebec's ministry of health and social services, and the FRQS. T.N. is supported by Research Fellowships of Japan Society for the Promotion of Science (JSPS) for Young Scientists and JSPS Overseas Challenge Program for Young Researchers. J.B.R. is supported by a FRQS Clinical Research Scholarship. The Columbia University Biobank was supported by Columbia University and the National Center for Advancing Translational Sciences, NIH, through Grant Number UL1TR001873. Columbia University COVID-19 Biobank members that additionally contributed to this work include Muredach P. Reilly, Wendy Chung, Eldad Hod, Soumitra Sengupta, Danielle Pendrick, Nitin Bhardwaj, Ning Shang, Atlas Khan, Chen Wang, Sheila M. O'Byrne, Renu Nandakumar, Amritha Menon, Yat S. So, Richard Mayeux, Ali G. Gharavi, Iuliana Ionita-Laza, Andrea Califano, Christine K. Garcia, Peter Sims, and Anne-Catrin Uhlemann. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or Columbia University. GenOMICC was funded by Sepsis Research (the Fiona Elizabeth Agnew Trust), the Intensive Care Society, a Wellcome-Beit Prize award to J. K. Baillie (Wellcome Trust 103258/Z/13/A), a BBSRC Institute Program Support Grant to the Roslin Institute (BBS/E/D/20002172, BBS/E/D/10002070 and BBS/E/D/30002275), the Medical Research Council [grant MC_PC_19059]. Research performed at the Human Genetics Unit was funded by the MRC (MC_UU_00007/10, MC_UU_00007/15). Whole-genome sequencing was done in partnership with Genomics England and was funded by UK Department of Health and Social Care, UKRI and LifeArc. 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. M Caulfield is an NIHR Senior Investigator. This work is part of the portfolio of translational research at the NIHR Biomedical Research Centre at Barts and Cambridge. LK was supported by an RCUK Innovation Fellowship from the National Productivity Investment Fund (MR/R026408/1). We acknowledge support from the MRC Human Genetics Unit programme grant, "Quantitative traits in health and disease" (U. MC_UU_00007/10). A. Tenesa acknowledges funding from MRC research grant MR/P015514/1, and HDR-UK award HDR-9004 and HDR-9003. Recruitment to GenOMICC was enabled by the National Institute of Healthcare Research Clinical Research Network (NIHR CRN) and the Chief Scientist Office (Scotland), who facilitate recruitment into research studies

in NHS hospitals, and to the global ISARIC and InFACT consortia. 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.

SUPPLEMENTARY TEXT

Regeneron Genetics Center (RGC) Research Team and Contribution Statements

All authors/contributors are listed in alphabetical order.

RGC Management and Leadership Team

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

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

Sequencing and Lab Operations

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

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

Clinical Informatics

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

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

Genome Informatics

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

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

Analytical Genetics

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

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

Immune, Respiratory, and Infectious Disease Therapeutic Area Genetics

Julie E. Horowitz, PhD.

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

Research Program Management

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

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

UK Biobank Exome Sequencing Consortium Research Team

[1]Bristol Myers Squibb

Oleg Moiseyenko, Carlos Rios, Saurabh Saha

[2]Regeneron Pharmaceuticals Inc.

Listed in pages 38 to 40.

[3]Biogen Inc.

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

[4]Alnylam Pharmaceuticals

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

[5]AstraZeneca

Olympe Chazara, Sri VV. Deevi, Xiao Jiang, Amanda O'Neill, Slavé Petrovski, Katherine Smith, Quanli Wang

[6]Takeda California Inc

Jason Tetrault, Dorothee Diogo, Aldo Cordova Palomera, Emily Wong, Rajesh Mikkilineni, David Merberg, Sunita Badola, Erin N. Smith, Sandor Szalma

[7]Pfizer, Inc

Yi-Pin Lai, Xing Chen, Xinli Hu, Melissa R. Miller

[8]Abbvie

Xiuwen Zheng, Bridget Riley-Gillis, Jason Grundstad, Sahar Esmaeeli, Jeff Waring, J. Wade Davis

[1]Bristol Myers Squibb, Route 206 and Province Line Road, Princeton, NJ 08543, USA
[2]Regeneron Pharmaceuticals Inc., 777 Old Saw Mill River Road, Tarrytown, New York 10591, USA
[3]Biogen Inc., 225 Binney Street, Cambridge, MA 02139, USA
[4]Alnylam Pharmaceuticals, 675 West Kendall St, Cambridge, MA 02142, USA
[5]AstraZeneca Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, Cambridge, UK
[6]Takeda California Inc., 9625 Towne Centre Dr, San Diego, CA 92121, USA
[7] Pfizer, Inc., 1 Portland Street, Cambridge MA 02139, USA
[8] AbbVie, Inc., 1 N. Waukegan Rd, North Chicago, IL 60064, USA

GenOMICC Consortium

Sara Clohisey[1], Fiona Griffiths[1], James Furniss[1], James Furniss[1], Trevor Paterson[1], Tony Wackett[1], Ruth Armstrong[1], Wilna Oosthuyzen[1], Nick Parkinson[1], Max Head Fourman[1], Andrew Law[1], Veronique Vitart[2], Lucija Klaric[2], Anne Richmond[2], Chris P. Ponting[2], Andrew D. Bretherick[2], Charles Hinds[3], Timothy Walsh[4], Sean Keating[4], Clark D Russell[1,5], Malcolm G. Semple[6,7], Kathy Rowan[8], Elvina Gountouna[9], Nicola Wrobel[10], Lee Murphy[10], Angie Fawkes[10], Richard Clark[10], Audrey Coutts[10], Lorna Donnelly[10], Tammy Gilchrist[10], Katarzyna Hafezi[10], Louise Macgillivray[10], Alan Maclean[10], Sarah McCafferty[10], Kirstie Morrice[10], , Angie Fawkes[10], Julian Knight[11], Charlotte Summers[12], Manu Shankar-Hari[13,14], Peter Horby[15], Alistair Nichol[16,17,18], David Maslove[19], Lowell Ling[20], Danny McAuley[21,22], Hugh Montgomery[23], Peter J.M. Openshaw[24,25].

[1]Roslin Institute, University of Edinburgh, Easter Bush, Edinburgh, EH25 9RG, UK
[2]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK
[3]William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK

[4]Intensive Care Unit, Royal Infirmary of Edinburgh, 54 Little France Drive, Edinburgh, EH16 5SA, UK
[5]Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, 47 Little France Crescent, Edinburgh, UK
[6]NIHR Health Protection Research Unit for Emerging and Zoonotic Infections, Institute of Infection, Veterinary and Ecological Sciences University of Liverpool, Liverpool, L69 7BE, UK
[7]Respiratory Medicine, Alder Hey Children's Hospital, Institute in The Park, University of Liverpool, Alder Hey Children's Hospital, Liverpool, UK
[8]Intensive Care National Audit & Research Centre, London, UK
[9]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK
[10]Edinburgh Clinical Research Facility, Western General Hospital, University of Edinburgh, EH4 2XU, UK
[11]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK
[12]Department of Medicine, University of Cambridge, Cambridge, UK
[13]Department of Intensive Care Medicine, Guy's and St. Thomas NHS Foundation Trust, London, UK
[14]School of Immunology and Microbial Sciences, King's College London, UK
[15]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Roosevelt Drive, Oxford, OX3 7FZ, UK
[16]Clinical Research Centre at St Vincent's University Hospital, University College Dublin, Dublin, Ireland
[17]Australian and New Zealand Intensive Care Research Centre, Monash University, Melbourne, Australia
[18]Intensive Care Unit, Alfred Hospital, Melbourne, Australia
[19]Department of Critical Care Medicine, Queen's University and Kingston Health Sciences Centre, Kingston, ON, Canada
[20]Department of Anaesthesia and Intensive Care, The Chinese University of Hong Kong, Prince of Wales Hospital, Hong Kong, China
[21]Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, Northern Ireland, UK

[22]Department of Intensive Care Medicine, Royal Victoria Hospital, Belfast, Northern Ireland, UK

[23]UCL Centre for Human Health and Performance, London, W1T 7HA, UK

[24]National Heart and Lung Institute, Imperial College London, London, UK

[25]Imperial College Healthcare NHS Trust: London, London, UK

1. Dewey, F.E., et al., *Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study.* Science, 2016. **354**(6319).
2. Park, J., et al., *A genome-first approach to aggregating rare genetic variants in LMNA for association with electronic health record phenotypes.* Genet Med, 2020. **22**(1): p. 102-111.
3. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
4. Van Hout, C.V., et al., *Exome sequencing and characterization of 49,960 individuals in the UK Biobank.* Nature, 2020.
5. Zerbino, D.R., et al., *Ensembl 2018.* Nucleic Acids Research, 2017. **46**(D1): p. D754-D761.
6. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
7. Vaser, R., et al., *SIFT missense predictions for genomes.* Nat Protoc, 2016. **11**(1): p. 1-9.
8. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **7**(1): p. 7.20.1-7.20.41.
9. Chun, S. and J.C. Fay, *Identification of deleterious mutations within three human genomes.* Genome research, 2009. **19**(9): p. 1553-1561.
10. Schwarz, J.M., et al., *MutationTaster evaluates disease-causing potential of sequence alterations.* Nat Methods, 2010. **7**(8): p. 575-6.
11. Mbatchou, J., et al., *Computationally efficient whole genome regression for quantitative and binary traits.* bioRxiv, 2020: p. 2020.06.19.162354.
12. Mathieson, I. and G. McVean, *Differential confounding of rare and common variants in spatially structured populations.* Nature Genetics, 2012. **44**(3): p. 243-246.
13. Zaidi, A.A. and I. Mathieson, *Demographic history mediates the effect of stratification on polygenic scores.* eLife, 2020. **9**: p. e61548.
14. Pairo-Castineira, E., et al., *Genetic mechanisms of critical illness in Covid-19.* Nature, 2020.
15. Alexander, D.H. and K. Lange, *Enhancements to the ADMIXTURE algorithm for individual ancestry estimation.* BMC Bioinformatics, 2011. **12**: p. 246.
16. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
17. Gaziano, J.M., et al., *Million Veteran Program: A mega-biobank to study genetic influences on health and disease.* J Clin Epidemiol, 2016. **70**: p. 214-23.

18. Pedersen, B.S., et al., *Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches.* Genome Med, 2020. **12**(1): p. 62.
19. Povysil, G., et al., *Failure to replicate the association of rare loss-of-function variants in type I IFN immunity genes with severe COVID-19.* medRxiv, 2020: p. 2020.12.18.20248226.
20. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**(1110): p. 11.10.1-11.10.33.
21. Ren, Z., et al., *ATAV: a comprehensive platform for population-scale genomic analyses.* BMC Bioinformatics, 2021. **22**(1): p. 149.
22. Pruitt, K.D., et al., *The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.* Genome Res, 2009. **19**(7): p. 1316-23.
23. Jun, G., et al., *Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.* Am J Hum Genet, 2012. **91**(5): p. 839-48.
24. Manichaikul, A., et al., *Robust relationship inference in genome-wide association studies.* Bioinformatics, 2010. **26**(22): p. 2867-73.
25. Bourgey, M., et al., *GenPipes: an open-source framework for distributed and scalable genomic analyses.* Gigascience, 2019. **8**(6).