Exome-wide association study to identify rare variants influencing COVID-19 outcomes: Results from the Host Genetics Initiative

Authors: Guillaume Butler-Laporte[1,2], Gundula Povysil[3], Jack Kosmicki[4], Elizabeth T Cirulli[5], Theodore Drivas[6–8], Simone Furini[9], Chadi Saad[10], Axel Schmidt[11], Pawel Olszewski[12], Mathieu Quinodoz[13–15], Elifnaz Çelik[13,15], Kousik Kundu[16,17], Klaudia Walter[17], Junghyung Jung[18,19], Amy D Stockwell[20], Laura G Sloofman[21], Alexander W Charney[22], Daniel Jordan[22], Noam Beckmann[22], Bartlomiej Przychodzen[23], Timothy Chang[24], Tess D Pottinger[3], Ning Shang[25], Fabian Brand[11], Francesca Fava[9,26,27], Francesca Mari[9,26,27], Karolina Chwialkowska[12,28], Magdalena Niemira[29], Urszula Korotko[12,28], Szymon Pula[12], J Kenneth Baillie[30–33], Alex Stuckey[34], Andrea Ganna[35,36], Konrad J Karczewski[37,38], Kumar Veerapen[37,38], Mathieu Bourgey[39,40], Guillaume Bourque[39–41], Robert JM Eveleigh[39,40], Vincenzo Forgetta[2], David Morrison[2], David Langlais[40,41], Mark Lathrop[40,41], Vincent Mooser[41], Tomoko Nakanishi[2,41–43], Robert Frithiof[44], Michael Hultström[1,2,44,45], Miklos Lipcsey[44,46], Yanara Marincevic-Zuniga[47], Jessica Nordlund[47], Kelly M Schiabor Barrett[5], Alexandre Bolze[5], Simon White[5], Stephen Riffle[5], Francisco Tanudjaja[5], Efren Sandoval[5], Iva Neveux[48], Shaun Dabe[49], Nicolas Casadei[50,51], Susanne Motameny[52,53], Manal Alaamery[54,55], Salam Massadeh[54,55], Nora Aljawini[54,55], Mansour S Almutairi[54,55], Yaseen M Arab[56], Saleh A Alqahtan[57,58], Fawz S Al Harthi[59], Amal Almutairi[59], Fatima Alqubaishi[59], Sarah Alotaibi[59], Albandari Binowayn[59], Ebtehal A Alsolm[59], Hadeel El Bardisy[59], Mohammad Fawzy[59], COVID-19 Host Genetics Initiative, DeCOI Host Genetics Group, GEN-COVID Multicenter Study, GenOMICC Consortium, Japan COVID-19 Task Force, Regeneron Genetics Center, Daniel H Geschwind[24], Stephanie Arteaga[24], Alexis Stephens[60], Manish J Butte[60,61], Paul C Boutros[62], Takafumi N Yamaguchi[62], Shu Tao[62], Stefan Eng[62], Timothy Sanders[63], Paul J Tung[63], Michael E Broudy[63], Yu Pan[63], Alfredo Gonzalez[62], Nikhil Chavan[63], Ruth Johnson[64], Bogdan Pasaniuc[62,65,66], Brian Yaspan[20], Sandra Smieszek[23], Carlo Rivolta[13–15], Stephanie Bibert[67], Pierre-Yves Bochud[67], Maciej Dabrowski[68], Pawel Zawadzki[68,69], Mateusz Sypniewski[68], Elżbieta Kaja[68,70], Pajaree Chariyavilaskul[71], Voraphoj Nilaratanakul[72], Nattiya Hirankarn[73], Vorasuk Shotelersuk[74], Monnat Pongpanich[75], Chureerat Phokaew[76], Wanna Chetruengchai[77], Yosuke Kawai[78], Takanori Hasegawa[79], Tatsuhiko Naito[80,81], Ho Namkoong[82], Ryuya Edahiro[80,83], Akinori Kimura[84], Seishi Ogawa[85–87], Takanori Kanai[88], Koichi Fukunaga[89], Yukinori Okada[80,81,90–92], Seiya Imoto[93], Satoru Miyano[79], Serghei Mangul[18], Malak S Abedalthagafi[59], Hugo Zeberg[94], Joseph J Grzymski[48], Nicole L Washington[5], Stephan Ossowski[50,51], Kerstin U Ludwig[11,95], Eva C Schulte[96–98], Olaf Riess[50,51], Marcin Moniuszko[99,100], Miroslaw Kwasniewski[12,28], Hamdi Mbarek[10], Said I Ismail[10], Anurag Verma[7,8,101], David B Goldstein[3,102], Krzysztof Kiryluk[3,25], Alessandra Renieri[9,26,27,103], Manuel Ferreira[4], J Brent Richards[1,2,41,104–106]

Affiliations:

1 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada
2 Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada
3 Institute for Genomic Medicine, Columbia University, New York, New York, USA
4 Regeneron Genetics Center, Tarrytown, NY, USA
5 Helix, 101 S Ellsworth Ave Suite 350, San Mateo, California 94401
6 Division of Human Genetics, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

7 Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
8 Department of Genetics, Perelman School of Medicine, University of Pennsylvania,, Philadelphia, PA, USA
9 Department of Medical Biotechnologies, Med Biotech Hub and Competence Center, University of Siena, Siena, Italy
10 Qatar Genome Program, Qatar Foundation Research, Development and Innovation, Qatar Foundation, Doha, Qatar
11 Institute of Human Genetics, School of Medicine and University Hospital Bonn, University of Bonn, Bonn, Germany
12 IMAGENE.ME SA, Bialystok, Poland
13 Institute of Molecular and Clinical Ophthalmology Basel (IOB), Basel, Switzerland
14 Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom
15 Department of Ophthalmology, University Hospital Basel, Basel, Switzerland
16 Department of Haematology, University of Cambridge, Cambridge CB1 8RN, UK
17 Department of Human Genetics, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK
18 Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, California, USA
19 Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA
20 Genentech Inc, South San Francisco, California, USA
21 Seaver Autism Center for Research and Treatment; Department of Psychiatry; Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
22 Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
23 Vanda Pharmaceuticals, Washington District of Columbia, USA
24 Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
25 Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, New York, USA
26 Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy
27 Medical Genetics, University of Siena, Siena, Italy
28 Centre for Bioinformatics and Data Analysis, Medical University of Bialystok, Bialystok, Poland
29 Centre for Clinical Research, Medical University of Bialystok, Bialystok, Poland
30 Roslin Institute, University of Edinburgh, Easter Bush, Edinburgh, EH25 9RG, UK
31 MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK
32 Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, 47 Little France Crescent, Edinburgh, UK
33 Intensive Care Unit, Royal Infirmary of Edinburgh, 54 Little France Drive, Edinburgh, EH16 5SA, UK
34 Genomics England, London UK
35 Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland
36 Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
37 Stanley Center for Psychiatric Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

38 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
39 Canadian Centre for Computational Genomics, McGill University, Montréal, Québec, Canada
40 McGill Genome Center, McGill University, Montréal, Québec, Canada
41 Department of Human Genetics, McGill University, Montréal, Québec, Canada
42 Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan
43 Research Fellow, Japan Society for the Promotion of Science
44 Anaesthesiology and Intensive Care Medicine, Department of Surgical Sciences, Uppsala University, Uppsala, Sweden
45 Integrative Physiology, Department of Medical Cell Biology, Uppsala University, Uppsala, Sweden
46 Hedenstierna Laboratory, CIRRUS, Anaesthesiology and Intensive Care Medicine, Department of Surgical Sciences, Uppsala University, Uppsala, Sweden
47 Department of Medical Sciences, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
48 Center for Genomic Medicine, Desert Research Institute, 2215 Raggio Pkwy, Reno, Nevada 89512
49 Renown Health, 1155 Mill St., Reno, Nevada 89502
50 Institute of Medical Genetics and Applied Genomics, University of Tuebingen, Tuebingen, Germany
51 NGS Competence Center Tuebingen, Institute of Medical Genetics and Applied Genomics, University of Tuebingen, Tuebingen, Germany
52 West German Genome Center, site Cologne, University of Cologne, Cologne, Germany
53 Cologne Center for Genomics, University of Cologne, Cologne, Germany
54 Developmental Medicine Department, King Abdullah International Medical Research Center, King Saud Bin Abdulaziz University for Health Sciences, King Abdulaziz Medical City, Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia
55 Saudi Human Genome Project at King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia
56 Ministry of the National Guard Health Affairs, King Abdullah International Medical Research Center and King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
57 Liver Transplant Unit, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia
58 Division of Gastroenterology and Hepatology, Johns Hopkins University, Baltimore, Maryland, USA
59 Genomics Research Department, Saudi Human Genome Project, King Fahad Medical City and King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia
60 Department of Pediatrics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
61 Department of Microbiology, Immunology, and Molecular Genetics (MIMG), David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
62 Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

63 Office of Health Informatics and Analytics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
64 Department of Computer Science, McGill University, Montréal, Québec, Canada
65 Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
66 Department of Pathology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
67 Infectious Diseases Service, Department of Medicine, University Hospital and University of Lausanne, Lausanne, Switzerland
68 MNM Bioscience Inc.,1 Broadway, Cambridge, MA 02142
69 Faculty of Physics, Adam Mickiewicz University, Uniwersytetu Poznańskiego 2, 61-614, Poznan, Poland
70 Department of Medical Chemistry and Laboratory Medicine, Poznań University of Medical Sciences, 60-806 Poznań, Poland
71 Clinical Pharmacokinetics and Pharmacogenomics Research Unit, Department of Pharmacology, Faculty of Medicine, Chulalongkorn University, Bangkok Thailand
72 Healthcare-associated Infection Research Group STAR (Special Task Force for Activating Research) and Division of Infectious Diseases, Department of Medicine,Chulalongkorn University, Bangkok, Thailand
73 Center of Excellence in Immunology and Immune-mediated Diseases, Department of Microbiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
74 Center of Excellence for Medical Genomics, Medical Genomics Cluster, and Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
75 Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand
76 Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
77 Center of Excellence for Medical Genomics, Medical Genomics Cluster, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
78 Genome Medical Science Project, National Center for Global Health and Medicine (NCGM), Tokyo, Japan
79 M&D Data Science Center, Tokyo Medical and Dental University, Tokyo, Japan
80 Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan
81 Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
82 Department of Infectious Diseases, Keio University School of Medicine, Tokyo, Japan
83 Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine, Suita, Japan
84 Institute of Research, Tokyo Medical and Dental University, Tokyo, Japan
85 Department of Pathology and Tumor Biology, Kyoto University, Kyoto, Japan
86 Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan
87 Department of Medicine, Center for Hematology and Regenerative Medicine, Karolinska Institute, Stockholm, Sweden
88 Division of Gastroenterology and Hepatology, Department of Medicine, Keio University School of Medicine, Tokyo, Japan
89 Division of Pulmonary Medicine, Department of Medicine, Keio University School

of Medicine, Tokyo, Japan
90 Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan
91 Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan
92 Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita, Japan
93 Division of Health Medical Intelligence, Human Genome Center, the Institute of Medical Science, the University of Tokyo, Tokyo, Japan
94 Department of Neuroscience, Karolinska Institutet, Stockholm, Sweden
95 West German Genome Center, site Bonn, University of Bonn, Bonn, Germany
96 Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, LMU Munich, 80336, Munich, Germany
97 Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, 80336, Munich, Germany
98 Institute of Virology, Technical University Munich/Helmholtz Zentrum München, Munich, Germany
99 Department of Regenerative Medicine and Immune Regulation, Medical University of Bialystok, Bialystok, Poland
100 Department of Allergology and Internal Medicine, Medical University of Bialystok, Bialystok, Poland
101 Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, USA
102 Department of Genetics & Development, Columbia University, New York, New York, USA
103 Medical Genetics Unit, University of Siena, Policlinico Le Scotte, Viale Bracci, 2, 53100, Siena, Italy
104 Infectious Diseases and Immunity in Global Health Program, Research Institute of the McGill University Health Centre, Montréal, Québec, Canada
105 Department of Twin Research, King's College London, London, United Kingdom
106 5 Prime Sciences Inc, Montreal, Quebec, Canada

Corresponding author:
Brent Richards, Professor of Medicine
McGill University
Senior Lecturer, King's College London (Honorary)
Contact:
Pavilion H-413, Jewish General Hospital
3755 Côte-Ste-Catherine
Montréal, Québec, Canada, H3T 1E2
T: +1 514 340 8222 x24362 F: +1 514 340 7529
E: brent.richards@mcgill.ca
www.mcgill.ca/genepi

Funding:
See **Suppl. Tables 2-4**.

Disclosures:
See **Supp. Tables 2-4**.

Article type: Article
Word count: 4190 (Introduction, Results, Discussion, and Methods)

Key words: COVID-19, SARS-CoV-2, Whole-Exome Sequencing, Whole-Genome Sequencing

Abstract

Host genetics is a key determinant of COVID-19 outcomes. Previously, the COVID-19 Host Genetics Initiative genome-wide association study used common variants to identify multiple loci associated with COVID-19 outcomes. However, variants with the largest impact on COVID-19 outcomes are expected to be rare in the population. Hence, studying rare variants may provide additional insights into disease susceptibility and pathogenesis, thereby informing therapeutics development. Here, we combined whole-exome and whole-genome sequencing from 21 cohorts across 12 countries and performed rare variant exome-wide burden analyses for COVID-19 outcomes. In an analysis of 5,048 severe disease cases and 571,009 controls, we observed that carrying a rare deleterious variant in the SARS-CoV-2 sensor toll-like receptor *TLR7* (on chromosome X) was associated with a 5.3-fold increase in severe disease (95% CI: 2.75-10.05, $p=5.41 \times 10^{-7}$). These results further support *TLR7* as a genetic determinant of severe disease and suggest that larger studies on rare variants influencing COVID-19 outcomes could provide additional insights.

Introduction

Despite successful vaccine programs, SARS-CoV-2 is still a major cause of mortality and widespread societal disruption[1,2]. While disease severity has correlated with well established epidemiological and clinical risk factors (e.g., advanced age, obesity, immunosuppression), these do not explain the wide range of COVID-19 presentations[3]. Hence, individuals without one of these known risk factors may have a genetic predisposition to severe COVID-19[4]. These genetic determinants to severe disease can, in turn, inform about the pathophysiology underlying COVID-19 severity and accelerate therapeutics development[5,6].

Previous work on COVID-19 host genetics using genome-wide association studies (GWASs) revealed 23 statistically robust genetic loci associated with either COVID-19 severity or susceptibility[7–11]. Given that most GWASs use genetic data obtained from genome-wide genotyping followed by imputation to measure the association between a phenotype and genetic variation, their reliability and statistical power declines as a variant's frequency decreases, especially at allele frequencies of less than 1%[12]. Ascertainment of rare genetic variation can be improved with sequencing technology[13]. Rare variants are expected to be enriched for larger effect sizes, due to evolutionary pressure on highly deleterious variants, and may therefore provide unique insights into genetic predisposition to COVID-19 severity. Identifying such genes may highlight critical control points in the host response to SARS-CoV-2 infection.

Measuring the effect of rare genetic variants on a given phenotype (here COVID-19) is difficult. Specifically, while variants of large effect on COVID-19 are more likely to be rare, the converse is not true, and most rare variants are not expected to strongly impact COVID-19 severity[14]. Therefore, unless large sample sizes and careful statistical adjustments are used, most rare variant genetic associations studies risk being underpowered, and are at higher risk of false or inflated effect estimates if significant associations are found between COVID-19 and genetic loci. This is exemplified by the fact that several rare variant associations reported for COVID-19 have not been replicated in independent cohorts[15–17].

Here, we investigated the association of rare genetic variants on the risk of COVID-19 by combining gene burden test results from whole exome and whole genome sequencing. To our knowledge, this is the first rare genetic variant burden test meta-analysis ever performed on a worldwide scale, including 21 cohorts, in 12 countries, including all main continental genetic ancestries.

Results

Study population and outcome

The final analysis included up to 28,159 individuals infected with SARS-CoV-2, and up to 596,189 controls from 21 cohorts in 12 countries (**Figure 1**). Most participants were of European genetic ancestry (n=576,389), but the consortium also included participants of Admixed American (n=4,529), African (n=25,465), East Asian (n=4,058), Middle Eastern (n=4,977) and South Asian ancestries (n=9,943). Participating cohorts enrolled patients based on local protocols, and both retrospective and prospective designs were used. Genetic sequencing was also performed locally, and cohorts were provided with a specific framework for quality control analyses, but each were allowed to deviate based on individual needs. Both exome (n = 11 cohorts) and genome sequencing (n = 10 cohorts) were included in the meta-analyses. The mean age of participants was 55.7, and 55.9% were females.

We studied three separate outcome phenotypes, as previously described by the COVID-19 Host Genetics Initiative (COVID-19 HGI)[8]. Briefly, the outcome cases were defined according to three standard COVID-19 HGI outcomes: A) severe disease: individuals with SARS-CoV-2 infection who died or required invasive respiratory support (extracorporeal membrane oxygenation, intubation with mechanical ventilation, high-flow oxygen support, or new bilevel or continuous positive airway pressure ventilation), B) hospitalisation: individuals with SARS-CoV-2 who died or required hospitalisation, and C) susceptibility to infection: any individual with SARS-CoV-2 infection. These are also referred to as A2, B2, and C2, respectively, in the COVID-19 HGI meta-analyses[8]. For all three phenotypes, controls were all individuals not classified as cases (including population controls with unknown COVID-19 status). The final meta-analyses included up to 5,048 cases and 571,009 controls for the severe disease outcome, 12,267 cases and 589,175 controls for the hospitalisation outcome, and 28,159 cases and 596,189 controls for the susceptibility outcome.

Burden test definition

Given the expected paucity of large-effect size rare deleterious variants, strategies have been devised to increase statistical power to test associations between rare variants and biomedically-relevant outcomes. One such strategy is to use burden tests[18], where each variant is collapsed into larger sets of variants, and association is tested between groups of variants and an outcome. Here, we collapsed deleterious variants in each gene and devised the following burden test: for each gene, an individual received a score of 0 if they do not carry any deleterious variant, a score of 1 if they carry at least one non-homozygous deleterious variant, and a score of 2 if they carry at least 1 homozygous deleterious variant. Similar to previous studies on burden testing of rare variants[17,19] deleterious variants were chosen using three masks: 1) “pLoF" which uses only predicted loss of function variants, 2) “coding5” which uses all variants in pLoF, as well as indels of moderate consequence as predicted by Ensembl[20], and missense variants classified as deleterious in 5 *in-silico* algorithms (see Methods), and 3) “coding1”, which uses all variants in coding5 and coding5, and also adds all missense variants classified as deleterious in at least 1 of the *in-silico* algorithms. The analyses were performed for variants with minor allele frequency (MAF) between < 1% and < 0.1%. MAFs were obtained from a combination of gnomAD[21] and cohort-specific common variant exclusion lists. These common variant lists included variants that achieved a MAF of >1 % or > 0.1 % in at least one study population within the consortium. To reduce the effect of fluctuations due to sampling, a minor allele count (MAC) ≥ 6 in the corresponding study was required for inclusion in the common variant list. Such “blacklists” have been shown to increase statistical power by removing variants at lower risk of being highly deleterious, and it reduces the risk of having cohort-specific false-positive variants being retained on the overall analysis[22].

The resulting score (either 0, 1, or 2) for each mask was then regressed on each of our three phenotypes using logistic regression, controlling for age, age^2, sex, sex*age, sex*age^2, and 10 common variant (MAF > 1%) genetic principal components (the same covariates as for COVID-19 HGI GWASs[7,8]). Additionally, given that population genetic structure and its confounding effect on phenotypes is different at the rare variant level[23], we also used the first 20 genetic principal components from rare variants (MAF<1%) as covariates in all our analyses. Analyses were performed separately by each cohort and each ancestry using Firth regression as applied in the Regenie software[24]. Firth regression is a penalized likelihood regression method that provides unbiased effect

estimates even in highly unbalanced case-control analyses, as expected with rare variants[25]. The summary statistics were then meta-analyzed with a fixed effect inverse-variance weighted model within each ancestry, and then with a DerSimonian-Laird random effect model across ancestries.

Main analysis results

Our meta-analysis included a total of 18,883 protein-coding genes, and all burden test genetic inflation factors, for all masks, were less than 1, suggesting that our results were not biased by population stratification and that Firth regression adequately adjusted for unbalanced case-control counts. Using an exome-wide significance p-value threshold of 0.05/20,000 = 2.5×10^{-6}, we found 3 genes associated with one of the COVID-19 phenotypes in at least one mask in our meta-analyses (**Table 1**). Of specific interest, we observed that carrying a predicted loss of function (pLoF) or missense variant (mask coding5) in the toll-like receptor 7 (*TLR7*) gene was associated with a 5.3-fold increase (95% CI: 2.7-10.1, $p=5.41 \times 10^{-7}$) in odds of severe COVID-19. *TLR7* is an important part of the innate viral immunity, encoding a protein that recognizes coronaviruses and other single-stranded RNA viruses, leading to upregulation of the type-1 and type-2 interferon pathway[26]. Results from the severe COVID-19 outcome analyses of *TLR7* with other masks also nearly reach our statistical significance threshold, with larger effects found in the pLoF mask (OR: 13.6, 95% CI: 4.41-44.3, $p=1.64 \times 10^{-5}$) and smaller effect in the coding1 mask (OR: 3.12, 95% CI: 1.91-5.10, $p=5.30 \times 10^{-6}$), though the latter was balanced by smaller standard errors due to the larger number of cases (3275 cases in coding1 vs 1577 in pLoF), as expected. These findings further support previous reports of *TLR7* errors of immunity underlying severe COVID-19 presentations[17,27–30].

In the meta-analyses, we also found that pLoFs in *MARK1* were associated with a 23.9-fold increase in the odds of severe COVID-19 (95% CI: 6.5-88.2, $p=1.89 \times 10^{-6}$), and a 12.3-fold increase in the odds of hospitalisation due to COVID-19 (95% CI: 4.8-31.2, $p=1.43 \times 10^{-7}$). While the number of *MARK1* pLoFs found in severe and hospitalized cases was small (MAC=4 and MAC=8, respectively), the signal was consistent in our three largest cohorts: UK Biobank, Penn Medicine, and Geisinger Health Services. *MARK1* is a member of the microtubule affinity-regulating kinase family, and is involved in multiple biological processes, chief among which is the promotion of microtubule dynamics[31]. *MARK1* has previously been shown to interact with the SARS-CoV-2 ORF9b protein[32], further supporting its potential role in COVID-19. Lastly, our meta-analyses also found marginal evidence for an association between severe COVID-19 and pLoFs in *RILPL1* (OR: 20.2, 95% CI: 5.8-70.7, $p=2.42 \times 10^{-6}$), a gene that, like *MARK1,* is associated with microtubule formation and ciliopathy[33].

When we meta-analyzed p-values using the aggregated Cauchy association test[34] (ACAT), the association between *TLR7* and severe COVID-19 ($p=1.58 \times 10^{-6}$), and between *MARK1* and hospitalisation ($p=4.30 \times 10^{-7}$) remained exome-significant (**Figure 2**). Full summary statistics are available in **Supp. Table 6**.

Rare variants in interferon-related genes and at previously reported genome-wide significant loci

Despite a 7.7-fold increase in number of cases, and a 1,069-fold increase in number of controls, the previously reported associations of genes in the interferon pathway with COVID-19 outcomes[15,16] could not be replicated with either our exome-wide significance threshold (**Supp. Table 7**) or a more liberal one of p=0.05/10=0.005 (based on

Bonferroni correction by the number of genes in the interferon pathway defined in a previous study[15]).

We also tested for rare variant associations between GWAS candidate genes from genome-wide significant loci in the COVID-19 HGI GWAS meta-analyses, but observed no exome-wide significant associations (**Supp Table 8**). However, at a more liberal Bonferroni threshold of p=0.05/46=0.001 (correcting for the 46 genes in the COVID-19 HGI GWAS associated loci), we observed an increased burden of pLOF or missense variants (coding5 mask) in *ABO* gene among those susceptible to SARS-CoV-2 infection. For example, individuals carrying a pLoF with MAF<0.1% in *ABO* were at a 2.34-fold higher risk of having a positive SARS-CoV-2 infection (95% CI: 1.50-3.64, $p=1.6x10^{-4}$). Note that deleterious variants in *ABO* often lead to blood groups A and B[35,36], which is consistent with the epidemiological association that non-type-O individuals are at higher risk of COVID-19[37]. However, more work is required to better understand the genetics of this locus as it relates to COVID-19 outcomes. Lastly, missense variants in *NSF* (mask coding1, MAF<1%) were also associated with higher susceptibility to SARS-CoV-2 (OR: 1.48, 95% CI: 1.21-1.82, $p=1.4x10^{-4}$), but this association was not present in other masks (**Supp Table 8**).

Replication in GenOMICC

Data for the pLoF mask for *TLR7* and *MARK1* in the severe COVID-19 phenotype was then replicated with the GenOMICC cohort[11], a prospective study enrolling critically ill individuals with COVID-19, with controls selected from the 100,000 genomes cohort[38]. Results are shown in **Table 2**. For *TLR7*, European ancestry individuals with a pLoF had a 4.70-fold increase in odds of severe disease (95% CI: 1.58 to 14.0, p=0.005). In the sample of South Asian ancestry individuals, a pLoF was associated with a 1.90-fold increase in odds of severe disease, but the 95% confidence interval crossed the null (0.23 to 15.6, p=0.55), which was likely due to a much smaller sample size than in the European ancestry subgroup (1,202 vs 10,645). Of interest, in both Europeans and South Asians, no pLoFs were observed in either of the control groups.

On the other hand, we could not replicate an effect from *MARK1*, which demonstrated an OR of 1.21 in European ancestry participants (95% CI 0.075 to 19.7, p=0.89) and an OR of 4.21 in South Asian ancestry individuals (95% CI 0.058 to 307, p=0.51).

Single-variant analysis

We performed an exome-wide association study using single variants with MAF higher than 0.1% and allele count of 6 or more, with the same analysis design used in the COVID-19 HGI GWAS[8], and the same outcome phenotypes used in the burden testing above. The previously described Neanderthal chromosome 3 locus associated with COVID-19 outcomes was also found in all three phenotypes, with lead variants in the *FYCO1* gene for the severe COVID-19 and hospitalisation phenotypes (rs13059238 and rs41289622, respectively), and for the *LIMD1* gene in the susceptibility phenotype (rs141045534). One other variant was found in the hospitalisation phenotype in *SRRM1* (rs1479489847); this association was found in only two of our smaller cohorts (Genentech EUR and AMR ancestries, and Vanda EUR ancestry), and was not replicated in the larger ones (p=0.30 in the UK Biobank EUR ancestry). Summary statistics for genome-wide significant variants can be found in **Supp. Table 9**, and Manhattan plots can be found in the **Supp. Figures**.

Discussion

Whole genome and whole exome sequencing can provide unique insights into genetic determinants of COVID-19, by uncovering associations between rare genetic variants and COVID-19. Specifically, burden tests can be particularly helpful, because they test for coding variants, thereby pointing directly to a causal gene and often suggesting a direction of effect. However, such studies require careful control for population stratification and an adapted analysis method such as burden testing, in order to have enough statistical power to find those associations. In our study, we observed that individuals with rare deleterious variants at *TLR7* are at increased risk of severe COVID-19 (up to 13.1-fold increase in odds in those with pLoFs). Although this association was suggested by previous studies[27–29], our study provides the most definitive evidence for the role of TLR7 in COVID-19 pathogenesis, with exome-wide significance for this gene in the discovery phase followed by strong replication in a large independent cohort. *TLR7* is a well-studied part of the antiviral immunity cascade and stimulates the interferon pathway after recognizing viral pathogen-associated molecular patterns.

We also uncovered a potential role for cellular microtubule disruption in the pathogenesis of COVID-19 and the microtubule network is known to be exploited by other viruses during infections[39]. Indeed, the MARK1 protein has been shown to interact with SARS-CoV-2 in previous *in-vitro* experiments[32]. Nevertheless, these findings at *MARK1* were not replicated in the GenOMICC cohort and will need to be tested in larger cohorts, especially given the small number of highly deleterious variants that we found in our consortium.

To our knowledge, this is the first time a rare variant burden test meta-analysis has been attempted on such a large scale. Our framework allowed for easy and interpretable summary statistics results, while at the same time preventing participant de-identification or any breach of confidentiality that stems from sharing results of rare genetic variant analyses[40]. It also provides important insights into how these endeavours should be planned in the future. First, our burden test operated under the assumption that the effect of any of the deleterious variants on the phenotype would be in the same direction and did not account for compound deleterious variant heterozygosity. This allowed for easier meta-analysis across cohorts but may have decreased statistical power. Other methods may be needed in future analysis to soften this assumption, though some of these cannot be easily meta-analyzed across multiple cohorts directly from summary statistics (e.g., SKAT-O[41]). Similarly, methods that combine both rare and common variants might also provide additional insights into disease outcomes[30,42]. Second, our results highlight the importance at looking at different categories of variants through different masks to increase sensitivity and specificity of our burden tests. Lastly, work remains to be done to standardize sequencing and annotation pipelines to allow comparisons of results easily across studies and cohorts. Here, we provided a pipeline framework to evert participating cohorts, but there remains room for process harmonization in the future.

Our study had limitations. First, even if this is one of the world's largest consortia using sequencing technologies for the study of rare variants, we remain limited by a relatively small sample size. For example, in a recent analyses of UK Biobank exomes, many of the phenotypes for which multiple genes were found using burden tests had a much higher number of cases than in our analyses (e.g. blonde hair colour, with 48,595 cases)[19]. Further, rare variant signals were commonly found in regions enriched in common variants found in GWASs. The fact that *ABO* and *NSF* were the only genes from the COVID-19 HGI GWAS that were also identified in our burden test (albeit using

a more liberal significance threshold), also suggests a lack of statistical power. Similarly, GenOMICC, a cohort of similar size, was also unable to find rare variant associations using burden tests[11]. However, their analysis methods were different from ours, making further comparisons difficult. Nevertheless, this provides clear guidance that smaller studies looking at the effect of rare variants across the genome are at considerable risk of finding both false positive and false negative associations. Second, many cohorts used population controls, which may have decreased statistical power given that some controls may have been misclassified. However, given that COVID-19 critical illness remains a rare phenomenon[43], our severe disease phenotype results are unlikely to be strongly affected by this. Further, the use of population control is a long-established strategy in GWAS burden tests[7,8,11,19,44], and the statistical power gain from increasing our sample size is likely to have counter-balanced the misclassification bias.

In summary, we reproduced an exome-wide significant association with severe COVID-19 outcomes in carriers of rare deleterious variants at *TLR7*. Our results also suggest an association between the cellular microtubule network and severe disease, which requires further validation. More importantly, our results underline the fact that future genome-wide studies of rare variants will require considerably larger sample size, but our work provides a roadmap for such collaborative efforts.

Methods

COVID-19 outcome phenotypes

For all analyses, we used three case-control definitions: A) Severe COVID-19, where cases were those who died, or required either mechanical ventilation (including extracorporeal membrane oxygenation), high-flow oxygen supplementation, new continuous positive airway pressure ventilation, or new bilevel positive airway pressure ventilation, B) Hospitalized COVID-19, where cases were all those who died or were admitted with COVID-19, and C) Susceptibility to COVID-19, where cases are anyone who tested positive for COVID-19, self-reported an infection to SARS-CoV-2, or had a mention of COVID-19 in their medical record. For all three, controls were individuals who did not match case definitions, including population controls for which case status was unknown (given that most patients are neither admitted with COVID-19, nor develop severe disease[45]). These three analyses are also referred to as analyses A2, B2, and C2 by the COVID-19 Host Genetics Initiative[8], respectively.

Cohort inclusion criteria and genetic sequencing

Any cohort with access to genetic sequencing data and the associated patient level phenotypes were allowed in this study. Specifically, both whole-genome and whole-exome sequencing was allowed, and there were no limitations in the platform used. There were no minimal number of cases or controls necessary for inclusion. However, the first step of Regenie, which was used to perform all tests (see below), uses a polygenic risk score which implicitly requires that a certain sample size threshold be reached (which depends on the phenotype and the observed genetic variation). Hence, cohorts were included if they were able to perform this step. All cohorts obtained approval from their respective institutional review boards, and informed consent was obtained from all participants. More details on each cohort's study design and ethics approval can be found in the **Supp Tables 1-2**.

Variant calling and quality control

Variant calling was performed locally by each cohort, with the pre-requisite that variants not be joint-called separately between cases and controls. Quality control was also

performed individually by each cohort according to individual needs. However, a general quality control framework was made available using the Hail software[46]. This included variant normalization and left alignment to a reference genome, removal of samples with call rate less than 97% or mean depth less than 20. Genotypes were set to unknown if they had genotype quality less than 20, depth less than 10, or poor allele balance (more than 0.1 for homozygous reference calls, less than 0.9 for homozygous alternative calls, and either below 0.25 or above 0.75 for heterozygous calls. Finally, variants were removed from if the mean genotype quality was less than 11, mean depth was less than 6, mean call rate less than or equal to 0.8, and Hardy-Weinberg equilibrium p-value less than or equal to $5x10^{-8}$ (10^{-16} for single variant association tests). Details on variant calling and quality control is described for each cohort in the **Supp. Table 1**.

Variant exclusion list

For the burden tests, we also compiled a list of variants that had a MAF > 1 % or > 0.1 % in any of the participating cohorts. This list was used to filter out variants that were less likely to have a true deleterious effect on COVID-19, even if they were considered rare in other cohorts, or in reference panels[22]. We created two such variant exclusion lists: one to be used in our burden test with variants of MAF less than 1%, and the other for the analysis with MAF less than 0.1%. In any cohort, if a variant had a minor allele count of 6 or more, and a MAF of more than 1% (or 0.1%), this variant was added to our exclusion list. This list was then shared with all participating cohorts, and all variants contained were removed from our burden tests.

Gene burden tests

The following analyses generally followed the methods used by recent literature on large-scale whole-exome sequencing[19] and the COVID-19 HGI[8].

The burden tests were performed by pooling variants in three different variant sets (called masks): “pLoF” which included loss of functions as defined by high impact variants in the Ensembl database[20] (i.e. transcript ablation, splice acceptor variant, splice donor variant, stop gained, frameshift variant, stop lost, start lost, transcript amplification), “coding5” which included all variants in pLoF as well as moderate impact indels and any missense variants that was predicted to be deleterious based on all of the *in-silico* pathogenicity prediction scores used, and “coding1” which included all variants in coding5 as well as all missense variants that were predicted to be deleterious in at least one of the *in-silico* pathogenicity prediction scores used. For *in-silico* prediction, we used the following five tools: SIFT[47], LRT[48], MutationTaster[49], PolyPhen2[50] with the HDIV database, and PolyPhen2 with the HVAR database. Protein coding variants were collapsed on canonical gene transcripts. Masks pLoF, coding5, and coding1 are equivalent to masks M1, M3, and M4, respectively, from the recent UK Biobank whole-exome sequencing paper by Backman *et al.*[19] and Kosmicki *et al.*[17].

Once variants were collapsed into genes in each participant, for each mask, genes were given a score of 0 if the participant had no variants in the mask, a score of 1 if the participant had one or more heterozygous variant in this mask, and a score of 2 if the participant had one or more homozygous variant in this mask. These scores were used as regressors in logistic regression models for the three COVID-19 outcomes above. These regressions were also adjusted for age, age*age, sex, age*sex, age*age*sex, 10 genetic principal components obtained from common genetic variants (MAF>1%), and 20 genetic principal components obtained from rare genetic variants (MAF<1%). The Regenie software[24] was used to perform all burden tests, and generate the scores

above. Regenie uses Firth penalized likelihood to adjust for rare or unbalanced events, providing unbiased effect estimates.

All analyses were performed separately for each of 6 genetic ancestries (African, Admixed American, East Asian, European, Middle Eastern, and South Asian). Summary statistics were then meta-analyzed using a fixed effect model within each ancestry and using a DerSimonian-Laird random effect model across ancestries with the Metal package[51] and its random effect extension[52]. Participant assignment to genetic ancestry was done locally by each cohort, more details on the methods can be found in the **Supp. Table 1**.

Lastly, we used ACAT[34] to meta-analyze p-values across masks, within each phenotype separately. ACAT is not affected by lack of independence between tests. These values were used to draw Manhattan and QQ plots in **Figure 2**.

Single variant association tests

We performed single variant association tests using a GWAS additive model framework. We used the same COVID-19 outcomes and covariates as above, except for the addition of the 20 rare genetic variant principal components. Once again, each cohort performed their analyses separately for each genetic ancestry, but also restricted their variants to those with MAF>0.01% and MAC>6. Summary statistics were meta-analyzed as above. Lastly, given that multiple technologies were used for sequencing, and that whole-exome sequencing can provide variant calls of worse quality in its off-target regions[53], we used the UKB, GHS, and Penn Medicine whole-exome sequencing variants as our "reference panel" for whole-exome sequencing. Hence, only variants reported in at least one of these biobanks were used in the final analyses.

Code availability

Code guidance is available at https://github.com/DrGBL/WES.WGS.

Data availability

The exome-wide burden test summary statistics are available in the Supplements. The single variant association studies summary statistics will be made available openly on the GWAS Catalog[54].

References

1 Cai Y, Kwek S, Tang SSL, *et al.* Impact of the COVID-19 pandemic on a tertiary care public hospital in Singapore: Resources and economic costs. *J Hosp Infect* 2021; published online Dec 14. DOI:10.1016/j.jhin.2021.12.007.

2 Mulholland RH, Wood R, Stagg HR, *et al.* Impact of COVID-19 on accident and emergency attendances and emergency and planned hospital admissions in Scotland: an interrupted time-series analysis. *J R Soc Med* 2020; **113**: 444–53.

3 Huang C, Wang Y, Li X, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020; **395**: 497–506.

4 Nakanishi T, Pigazzini S, Degenhardt F, *et al.* Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. *J Clin Invest* 2021; **131**: e152386.

5 Zhou S, Butler-Laporte G, Nakanishi T, *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat Med* 2021; **27**: 659–67.

6 Gaziano L, Giambartolomei C, Pereira AC, *et al.* Actionable druggable genome-

wide Mendelian randomization identifies repurposing opportunities for COVID-19. *Nat Med* 2021; **27**: 668–76.
7 COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19: an update. *medRxiv* 2021; : 2021.11.08.21265944.
8 COVID-19 Host Genetics Initiative, Niemi MEK, Karjalainen J, *et al.* Mapping the human genetic architecture of COVID-19. *Nature* 2021. DOI:10.1038/s41586-021-03767-x.
9 Ellinghaus D, Degenhardt F, Bujanda L, *et al.* Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med* 2020; **383**: 1522–34.
10 Pairo-Castineira E, Clohisey S, Klaric L, *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* 2021; **591**: 92–8.
11 Kousathanas A, Pairo-Castineira E, Rawlik K, *et al.* Whole genome sequencing reveals host factors underlying critical Covid-19. *Nature* 2022. DOI:10.1038/s41586-022-04576-6.
12 Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019; **20**: 467–84.
13 Taliun D, Harris DN, Kessler MD, *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021; **590**: 290–9.
14 Ganna A, Satterstrom FK, Zekavat SM, *et al.* Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *Am J Hum Genet* 2018; **102**: 1204–11.
15 Qian Z, Paul B, Zhiyong L, *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science (80-)* 2020; **370**: eabd4570.
16 Povysil G, Butler-Laporte G, Shang N, *et al.* Rare loss-of-function variants in type I IFN immunity genes are not associated with severe COVID-19. *J Clin Invest* 2021; **131**. DOI:10.1172/JCI147834.
17 Kosmicki JA, Horowitz JE, Banerjee N, *et al.* Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. *Am J Hum Genet* 2021; **108**: 1350–5.
18 Cirulli ET. The Increasing Importance of Gene-Based Analyses. *PLOS Genet* 2016; **12**: e1005852.
19 Backman JD, Li AH, Marcketta A, *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 2021; **599**: 628–34.
20 Howe KL, Achuthan P, Allen J, *et al.* Ensembl 2021. *Nucleic Acids Res* 2021; **49**: D884–91.
21 Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020; **581**: 434–43.
22 Maffucci P, Bigio B, Rapaport F, *et al.* Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis. *Proc Natl Acad Sci* 2019; **116**: 950 LP – 959.
23 Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012; **44**: 243–6.
24 Mbatchou J, Barnard L, Backman J, *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* 2021; **53**: 1097–103.
25 Wang X. Firth logistic regression for rare variant association tests . Front. Genet. . 2014; **5**. https://www.frontiersin.org/article/10.3389/fgene.2014.00187.
26 Petes C, Odoardi N, Gee K. The Toll for Trafficking: Toll-Like Receptor 7 Delivery to the Endosome . Front. Immunol. . 2017; **8**: 1075.
27 van der Made CI, Simons A, Schuurs-Hoeijmakers J, *et al.* Presence of Genetic Variants Among Young Men With Severe COVID-19. *JAMA* 2020; **324**: 663–73.
28 Fallerini C, Daga S, Mantovani S, *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* 2021; **10**: e67569.
29 Asano T, Boisson B, Onodi F, *et al.* X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19. *Sci Immunol* 2021; **6**. DOI:10.1126/sciimmunol.abl4348.
30 Mantovani S, Daga S, Fallerini C, *et al.* Rare variants in Toll-like receptor 7 results in functional impairment and downregulation of cytokine-mediated signaling in COVID-19 patients. *Genes Immun* 2022; **23**: 51–6.
31 Drewes G, Ebneth A, Preuss U, Mandelkow EM, Mandelkow E. MARK, a novel family of protein kinases that phosphorylate microtubule-associated proteins and trigger microtubule disruption. *Cell* 1997; **89**: 297–308.
32 Gordon DE, Jang GM, Bouhaddou M, *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020; **583**: 459–68.
33 Schaub JR, Stearns T. The Rilp-like proteins Rilpl1 and Rilpl2 regulate ciliary membrane content. *Mol Biol Cell* 2013; **24**: 453–64.
34 Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* 2019; **104**: 410–21.
35 Ying Y, Hong X, Xu X, *et al.* Molecular Basis of ABO Variants Including Identification of 16 Novel ***ABO*** Subgroup Alleles in Chinese Han Population. *Transfus Med Hemotherapy* 2020; **47**: 160–6.
36 Hult AK, Yazer MH, Jørgensen R, *et al.* Weak A phenotypes associated with novel ABO alleles carrying the A2-related 1061C deletion and various missense substitutions. *Transfusion* 2010; **50**: 1471–86.
37 Zietz M, Zucker J, Tatonetti NP. Associations between blood type and COVID-19 infection, intubation, and death. *Nat Commun* 2020; **11**: 5761.
38 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *N Engl J Med* 2021; **385**: 1868–80.
39 Simpson C, Yamauchi Y. Microtubules in Influenza Virus Entry and Egress. Viruses . 2020; **12**. DOI:10.3390/v12010117.
40 Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. *Nat Genet* 2009; **41**: 965–7.
41 Lee S, Emond MJ, Bamshad MJ, *et al.* Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am J Hum Genet* 2012; **91**: 224–37.
42 Wainschtein P, Jain D, Zheng Z, *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* 2022. DOI:10.1038/s41588-021-00997-7.
43 Williamson EJ, Walker AJ, Bhaskaran K, *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020; **584**: 430–6.
44 Cirulli ET, White S, Read RW, *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat Commun* 2020; **11**: 542.
45 Mutambudzi M, Niedzwiedz C, Macdonald EB, *et al.* Occupation and risk of severe COVID-19: prospective cohort study of 120 075 UK Biobank participants. *Occup Environ Med* 2021; **78**: 307 LP – 314.
46 Hail Team. Hail 0.2. 2021. https://github.com/hail-is/hail.
47 Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc* 2016; **11**: 1–9.
48 Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009; **19**: 1553–61.

49 Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; **7**: 575–6.
50 Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr Protoc Hum Genet* 2013; **76**: 7.20.1-7.20.41.
51 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**: 2190–1.
52 Hemani G. random-metal. GitHub Repos. 2017. https://github.com/explodecomputer/random-metal (accessed March 15, 2022).
53 Guo Y, Long J, He J, *et al.* Exome sequencing generates high quality data in non-target regions. *BMC Genomics* 2012; **13**: 194.
54 Buniello A, Macarthur JAL, Cerezo M, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019. DOI:10.1093/nar/gky1120.

Supplements

Supplementary Table: Contains **Supp. Tables 1-9**.
Supplementary Figures: contains single variant exome-wide association studies Manhattan plots.

Acknowledgement

We thank the patients who volunteered to all participating cohorts, and the researchers and clinicians who enrolled them into the respective studies. A full list of acknowledgements can be found in **Supp. Table 2**.

Author contributions

Conceptualization and methodology: GBL, GP, JK, ETC, TD, SF, CS, ASchmidt, PO, MQ, EÇ, KK, KV, TN, MSA, HZ, SO, ECS, KUL, HM, DBG, KK, AR, MF, JBR
Formal analyses: GBL, GP, JK, ETC, TD, SF, CS, ASchmidt, PO, MQ, EÇ, KK, KW, JJ, ADS, LGS, BP, TC, TDP, AStuckey, YK, YO, AR
Investigation: all authors.
Resources: ETC, KW, AWC, FM, JKB, MB, GB, RJME, VF, DM, MLathrop, VM, MH, RF, MLipcsey, BP, BY, SS, CR, SB, PYB, MS, EK, PL, YK, YO, SM, MSA, HZ, SO, KUL, ECS, OR, MK, HM, SII, AV, DBG, KK, AR, MF, JBR
Data curation: GBL, GP, JK, ETC, TD, SF, CS, ASchmidt, PO, MQ, EÇ, KK, KW, JJ, ADS, KUL, LGS, BP, TC, TDP, YK, YO
Writing – original draft: GBL
Writing – review & editing: all authors.
Visualization: GBL
Supervision: ETC, KW, AWC, FM, JKB, MH, VM, MH, RF, MLipcsey, BP, BY, SS, CR, SB, PYB, MS, EK, PL, YK, YO, SM, MSA, HZ, SO, KUL, ECS, OR, MK, HM, SII, AV, DBG, KK, AR, MF, JBR
Project administration: GBL, FF, FM, AG, DBG, KK, AR MF, JBR

Competing interests

See **Supp. Table 2**.

Materials & Correspondence

J Brent Richards is the corresponding author (brent.richards@mcgill.ca).

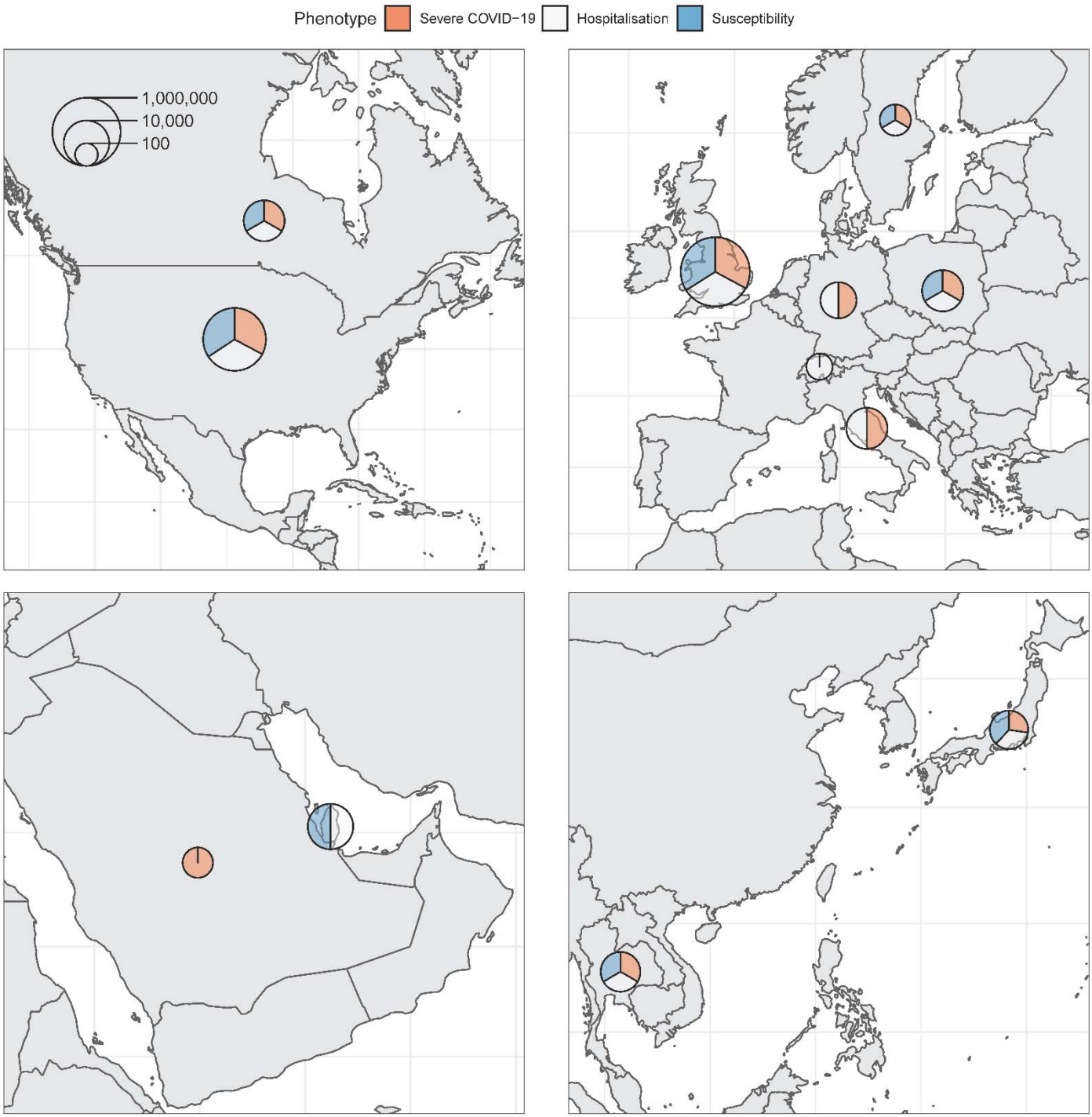

Figure 1: Maps of countries contributing data to the consortium. Sample sizes (cases and controls) for each phenotype were added and represented on the logarithmic scale by each circle. Relative contribution to each phenotype is represented by the three colors.

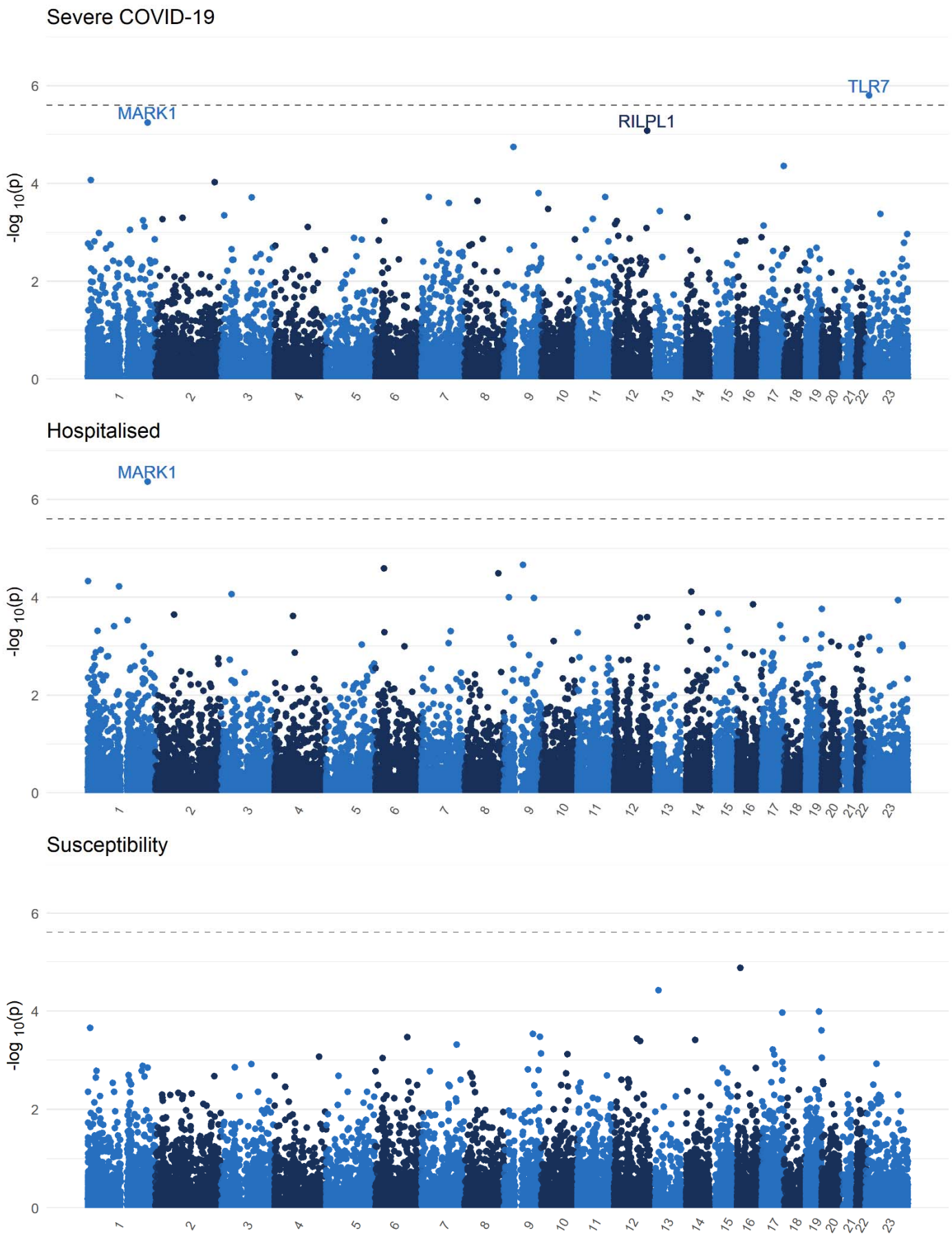

Figure 2: Exome burden test ACAT p-value meta-analysis Manhattan plots and QQ plots.

Gene	Mask	Phenotype	MAF	Beta	Standard Error	Odds Ratio	95% Confidence Interval	P-value	Heterogeneity p-value	N Cases 0|1|2 Burden Test	N Controls 0|1|2 Burden Test
Meta-Analysis Across Ancestries											
MARK1	pLoF	Severe COVID-19	<0.1%	3.17	0.67	23.9	6.5-88.2	$1.89x10^{-6}$	0.883	1935|4|0	540031|92|0
MARK1	pLoF	Severe COVID-19	<1%	3.17	0.67	23.9	6.5-88.2	$1.89x10^{-6}$	0.883	1935|4|0	540031|92|0
MARK1	pLoF	Hospitalisation	<0.1%	2.51	0.48	12.3	4.8-31.2	$1.43x10^{-7}$	0.893	6132|8|0	547943|93|0
MARK1	pLoF	Hospitalisation	<1%	2.51	0.48	12.3	4.8-31.2	$1.43x10^{-7}$	0.893	6132|8|0	547943|93|0
RILPL1	pLoF	Severe COVID-19	<0.1%	3.01	0.64	20.2	5.8-70.7	$2.42x10^{-6}$	0.941	1745|4|0	558448|121|0
TLR7	coding5	Severe COVID-19	<0.1%	1.66	0.33	5.25	2.75-10.05	$5.41x10^{-7}$	0.755	3101|2|5	519047|83|47
TLR7	coding5	Severe COVID-19	<1%	1.63	0.33	5.10	2.67-9.72	$7.48x10^{-7}$	0.760	3275|2|5	519834|85|47
Other TLR7 results for severe phenotype											
TLR7	pLoF	Severe COVID-19	<0.1%	2.61	0.60	13.6	4.14-44.4	1.64x10-5	0.820	1577|0|2	508987|13|11
TLR7	pLoF	Severe COVID-19	<1%	2.61	0.60	13.6	4.14-44.4	1.64x10-5	0.820	1577|0|2	508987|13|11
TLR7	coding1	Severe COVID-19	<0.1%	1.14	0.25	3.12	1.91-5.10	$5.30x10^{-6}$	0.854	3275|3|7	519616|210|139
TLR7	coding1	Severe COVID-19	<1%	1.11	0.24	3.03	1.90-4.85	$3.43x10^{-6}$	0.956	3273|5|8	521166|221|144

Table 1: Exome-wide significant findings, as well as other *TLR7* results (for the severe phenotype only). Note that for Masks pLoF, all deleterious variants had a MAF<0.1%, and hence both burden tests (MAF<1% and 0.1%) gave the same results. Full results available in **Supp. Table 6**.

Gene	Ancestry	Beta	Standard Error	Odds Ratio	95% Confidence Interval	P-value	N Cases 0|1|2 Burden Test	N Controls 0|1|2 Burden Test
MARK1	EUR	0.195	1.42	1.21	0.075-19.7	0.891	5988|1|0	4655|1|0
MARK1	SAS	1.44	2.19	4.21	0.058-307.6	0.511	787|1|0	414|0|0
TLR7	EUR	1.55	0.555	4.70	1.58-14.0	0.005	5980|1|8	4566|0|0
TLR7	SAS	0.640	1.08	1.90	0.230-15.6	0.552	786|0|2	414|0|0

Table 2: Replication of pLoF mask, severe COVID-19, *MARK1* and *TLR7* results in the GenOMICC cohort. Note that the same variants were included in both the MAF<1% and MAF<0.1% replication, and the same results were obtained (shown here).

Gene	Mask	Phenotype	MAF	Beta	Standard Error	Odds Ratio	95% Confidence Interval	P-value	Heterogeneity p-value	N Cases 0|1|2 Burden Test	N Controls 0|1|2 Burden Test
NSF	coding1	Susceptibility	<1%	0.395	0.104	1.484	1.21-1.82	$1.44x10^{-4}$	0.866	25752|127|2	585642|1907|5
ABO	pLoF	Susceptibility	<0.1%	0.851	0.226	2.341	1.50-3.65	$1.68x10^{-4}$	0.498	22778|27|0	572310|296|0
ABO	pLoF	Susceptibility	<1%	0.784	0.209	2.19	1.45-3.30	$1.75x10^{-4}$	0.826	23460|34|0	574608|364|0
ABO	coding5	Susceptibility	<1%	0.729	0.195	2.073	1.41-3.04	$1.89x10^{-4}$	0.869	24455|42|0	575051|434|0
ABO	pLoF	Hospitalisation	<0.1%	1.33	0.395	3.78	1.74-8.20	$7.56x10^{-4}$	0.542	7859|12|0	561642|291|0
ABO	pLoF	Susceptibility	<0.1%	0.736	0.222	2.088	1.35-3.23	$9.35x10^{-4}$	0.512	23779|29|0	572799|320|0

Table 3: Results of burden tests at genes identified from common variants GWAS in the COVID-19 HGI. Only genes with $p<0.05/46$ are shown here. Full results available in **Supp. Table 8**.