

Novel risk loci for COVID-19 hospitalization among admixed American populations

Reviewed Preprint

v2 • September 5, 2024

Revised by authors

Reviewed Preprint

v1 • February 14, 2024

Silvia Diz-de Almeida, Raquel Cruz, Andre D Luchessi, José M Lorenzo-Salazar, Miguel López de Heredia, Inés Quintela, Rafaela González-Montelongo, Vivian N Silbiger, Marta Sevilla Porras, Jair Antonio Tenorio Castaño, Julian Nevado, Jose María Aguado, Carlos Aguilar, Sergio Aguilera-Albesa, Virginia Almadana, Berta Almoguera, Nuria Alvarez, Álvaro Andreu-Bernabeu, Eunáte Arana-Arri, Celso Arango, María J Arranz, Maria-Jesus Artiga, Raúl C Baptista-Rosas, María Barreda- Sánchez, Moncef Belhassen-Garcia, Joao F Bezerra, Marcos AC Bezerra, Lucía Boix-Palop, María Brion, Ramón Brugada, Matilde Bustos, Enrique J Calderón, Cristina Carbonell, Luis Castano, Jose E Castela, Rosa Conde-Vicente, M Lourdes Cordero-Lorenzana, Jose L Cortes-Sanchez, Marta Corton, M Teresa Darnaude, Alba De Martino-Rodríguez, Victor del Campo-Pérez, Aranzazu Diaz de Bustamante, Elena Domínguez-Garrido, Rocío Eirós, María Carmen Fariñas, María J Fernandez-Nestosa, Uxía Fernández-Robelo, Amanda Fernández-Rodríguez, Tania Fernández-Villa, Manuela Gago-Domínguez, Belén Gil-Fournier, Javier Gómez-Arrue, Beatriz González Álvarez, Fernan Gonzalez Bernaldo de Quirós, Anna González-Neira, Javier González-Peñas, Juan F Gutiérrez-Bautista, María José Herrero, Antonio Herrero-Gonzalez, María A Jimenez-Sousa, María Claudia Lattig, Anabel Liger Borja, Rosario Lopez-Rodriguez, Esther Mancebo, Caridad Martín-López, Vicente Martín, Oscar Martinez-Nieto, Iciar Martinez-Lopez, Michel F Martinez-Resendez, Ángel Martinez-Perez, Juliana F Mazzeu, Eleuterio Merayo Macías, Pablo Minguez, Victor Moreno Cuerda, Silviene F Oliveira, Eva Ortega-Paino, Mara Parellada, Estela Paz-Artal, Ney PC Santos, Patricia Pérez-Matute, Patricia Perez, M Elena Pérez-Tomás, Teresa Perucho, Mel Lina Pinsach-Abuin, Guillermo Pita, Ericka N Pompa-Mera, Gloria L Porras-Hurtado, Aurora Pujol, Soraya Ramiro León, Salvador Resino, Marianne R Fernandes, Emilio Rodríguez-Ruiz, Fernando Rodriguez-Artalejo, José A Rodriguez-Garcia, Francisco Ruiz-Cabello, Javier Ruiz-Hornillos, Pablo Ryan, José Manuel Soria, Juan Carlos Souto, Eduardo Tamayo, Alvaro Tamayo-Velasco, Juan Carlos Taracido-Fernandez, Alejandro Teper, Lilian Torres-Tobar, Miguel Urioste, Juan Valencia-Ramos, Zuleima Yáñez, Ruth Zarate, Itziar de Rojas, Agustín Ruiz, Pascual Sánchez, Luis Miguel Real, SCOURGE Cohort Group, Encarna Guillen-Navarro, Carmen Ayuso, Esteban Parra, José A Riancho, Augusto Rojas-Martinez, Carlos Flores ✉, Pablo Lapunzina, Ángel Carracedo ✉

Centro Singular de Investigación en Medicina Molecular y Enfermedades Crónicas (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain • CIBERER, ISCIII, Madrid, Spain • Pediatric Neurology Unit, Department of Pediatrics, Navarra Health Service Hospital, Pamplona, Spain • ERN-ITHACA-European Reference Network • Universidade Federal do Rio Grande do Norte, Departamento de Analises Clinicas e Toxicologicas, Natal, Brazil • Genomics Division, Instituto Tecnológico y de Energías Renovables, Santa Cruz de Tenerife, Spain • Fundación Pública Galega de Medicina Xenómica, Sistema Galego de Saúde (SERGAS) Santiago de Compostela, Spain • Instituto de Genética Médica y Molecular (INGEMM), Hospital Universitario La Paz-IDIPAZ, Madrid, Spain • Unit of Infectious Diseases, Hospital Universitario 12 de Octubre, Instituto de Investigación Sanitaria Hospital 12 de Octubre (imas12), Madrid, Spain • Spanish Network for Research in Infectious Diseases (REIPI RD16/0016/0002), Instituto de Salud Carlos III, Madrid, Spain • School of Medicine, Universidad Complutense, Madrid, Spain • CIBERINFEC, ISCIII, Madrid, Spain • Hospital General Santa Bárbara de Soria, Soria, Spain • Navarra Health Service, NavarraBioMed Research Group, Pamplona, Spain • Hospital Universitario Virgen Macarena, Neumología, Seville, Spain • Department of Genetics & Genomics, Instituto de Investigación Sanitaria-Fundación Jiménez Díaz University Hospital - Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain • Spanish National Cancer Research Centre, Human Genotyping-CEGEN Unit, Madrid, Spain • Department of Child and Adolescent Psychiatry, Institute of Psychiatry and Mental Health, Hospital General Universitario Gregorio

Marañón (IISGM), Madrid, Spain • Biocruces Bizkaia HRI, Bizkaia, Spain • Cruces University Hospital, Osakidetza, Bizkaia, Spain • Centre for Biomedical Network Research on Mental Health (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain • Fundació Docència I Recerca Mutua Terrassa, Barcelona, Spain • Spanish National Cancer Research Center, CNIO Biobank, Madrid, Spain • Hospital General de Occidente, Zapopan Jalisco, Mexico • Centro Universitario de Tonalá, Universidad de Guadalajara, Tonalá Jalisco, Mexico • Centro de Investigación Multidisciplinario en Salud, Universidad de Guadalajara, Tonalá Jalisco, Mexico • Instituto Murciano de Investigación Biosanitaria (IMIB-Arrixaca), Murcia, Spain • Universidad Católica San Antonio de Murcia (UCAM), Murcia, Spain • Hospital Universitario de Salamanca-IBSAL, Servicio de Medicina Interna-Unidad de Enfermedades Infecciosas, Salamanca, Spain • Universidad de Salamanca, Salamanca, Spain • Escola Tecnica de Saúde, Laboratório de Vigilância Molecular Aplicada, Brazil • Federal University of Pernambuco, Genetics Postgraduate Program, Recife, PE, Brazil • Hospital Universitario Mutua Terrassa, Barcelona, Spain • Instituto de Investigación Sanitaria de Santiago (IDIS), Xenética Cardiovascular, Santiago de Compostela, Spain • CIBERCV, ISCIII, Madrid, Spain • Cardiovascular Genetics Center, Institut d'Investigació Biomèdica Girona (IDIBGI), Girona, Spain • Medical Science Department, School of Medicine, University of Girona, Girona, Spain • Hospital Josep Trueta, Cardiology Service, Girona, Spain • Institute of Biomedicine of Seville (IBiS), Consejo Superior de Investigaciones Científicas (CSIC)- University of Seville- Virgen del Rocío University Hospital, Seville, Spain • Departamento de Medicina, Hospital Universitario Virgen del Rocío, Universidad de Sevilla, Seville, Spain • CIBERESP, ISCIII, Madrid, Spain • Hospital Universitario de Salamanca-IBSAL, Servicio de Medicina Interna, Salamanca, Spain • Osakidetza, Cruces University Hospital, Bizkaia, Spain • Centre for Biomedical Network Research on Diabetes and Metabolic Associated Diseases (CIBERDEM), Instituto de Salud Carlos III, Madrid, Spain • University of Pais Vasco, UPV/EHU, Bizkaia, Spain • Oncology and Genetics Unit, Instituto de Investigación Sanitaria Galicia Sur, Xerencia de Xestión Integrada de Vigo-Servizo Galego de Saúde, Vigo, Spain • Hospital Universitario Río Hortega, Valladolid, Spain • Servicio de Medicina intensiva, Complejo Hospitalario Universitario de A Coruña (CHUAC), Sistema Galego de Saúde (SERGAS), A Coruña, Spain • Tecnológico de Monterrey, Monterrey, Mexico • Otto von Guericke University, Department of Microgravity and Translational Regenerative Medicine, Magdeburg, Germany • Hospital Universitario Mostoles, Unidad de Genética, Madrid, Spain • Instituto Aragonés de Ciencias de la Salud (IACS), Zaragoza, Spain • Instituto Investigación Sanitaria Aragón (IIS-Aragon), Zaragoza, Spain • Preventive Medicine Department, Instituto de Investigación Sanitaria Galicia Sur, Xerencia de Xestión Integrada de Vigo-Servizo Galego de Saúde, Vigo, Spain • Unidad Diagnóstico Molecular. Fundación Rioja Salud, La Rioja, Spain • Hospital Universitario de Salamanca-IBSAL, Servicio de Cardiología, Salamanca, Spain • IDIVAL, Cantabria, Spain • Universidad de Cantabria, Cantabria, Spain • Hospital U M Valdecilla, Cantabria, Spain • Universidad Nacional de Asunción, Facultad de Politécnica, Paraguay • Urgencias Hospitalarias, Complejo Hospitalario Universitario de A Coruña (CHUAC), Sistema Galego de Saúde (SERGAS), A Coruña, Spain • Unidad de Infección Viral e Inmunidad, Centro Nacional de Microbiología (CNM), Instituto de Salud Carlos III (ISCIII), Madrid, Spain • Grupo de Investigación en Interacciones Gen-Ambiente y Salud (GIIGAS) - Instituto de Biomedicina (IBIOMED), Universidad de León, León, Spain • IDIS • Hospital Universitario de Getafe, Servicio de Genética, Madrid, Spain • Ministerio de Salud Ciudad de Buenos Aires, Buenos Aires, Argentina • Hospital Universitario Virgen de las Nieves, Servicio de Análisis Clínicos e Inmunología, Granada, Spain • IIS La Fe, Plataforma de Farmacogenética, Valencia, Spain • Universidad de Valencia, Departamento de Farmacología, Valencia, Spain • Data Analysis Department, Instituto de Investigación Sanitaria-Fundación Jiménez Díaz University Hospital - Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain • Universidad de los Andes, Facultad de Ciencias, Bogotá, Colombia • SIGEN Alianza Universidad de los Andes - Fundación Santa Fe de Bogotá, Bogotá, Colombia • Hospital General de Segovia, Medicina Intensiva, Segovia, Spain • Facultad de Farmacia, Universidad San Pablo-CEU, CEU Universities, Urbanización Montepríncipe, Boadilla del Monte, España • Hospital Universitario 12 de Octubre, Department of Immunology, Madrid, Spain • Instituto de Investigación Sanitaria Hospital 12 de Octubre (imas12), Transplant Immunology and Immunodeficiencies Group, Madrid, Spain • Fundación Santa Fe de Bogota, Departamento Patología y Laboratorios, Bogotá, Colombia • Unidad de Genética y Genómica Islas Baleares, Islas Baleares, Spain • Hospital Universitario Son Espases, Unidad de Diagnóstico Molecular y Genética Clínica, Islas Baleares, Spain • Genomics of Complex Diseases Unit, Research Institute of Hospital de la Santa Creu i Sant Pau, IIB Sant Pau, Barcelona, Spain • Universidade de Brasília, Faculdade de Medicina, Brazil • Programa de Pós-Graduação em

Ciências Médicas (UnB), Brazil • Programa de Pós-Graduação em Ciências da Saúde (UnB), Brazil • Hospital El Bierzo, Unidad Cuidados Intensivos, León, Spain • Hospital Universitario Mostoles, Medicina Interna, Madrid, Spain • Universidad Francisco de Vitoria, Madrid, Spain • Departamento de Genética e Morfología, Instituto de Ciencias Biológicas, Universidade de Brasília, Brazil • Programa de Pós-Graduação em Biologia Animal (UnB), Brazil • Programa de Pós-Graduação Profissional em Ensino de Biologia (UnB), Brazil • Universidad Complutense de Madrid, Department of Immunology, Ophthalmology and ENT, Madrid, Spain • Universidade Federal do Pará, Núcleo de Pesquisas em Oncologia, Belém, Pará, Brazil • Infectious Diseases, Microbiota and Metabolism Unit, CSIC Associated Unit, Center for Biomedical Research of La Rioja (CIBIR), Logroño, Spain • Inditex, A Coruña, Spain • GENYCA, Madrid, Spain • Instituto Mexicano del Seguro Social (IMSS), Centro Médico Nacional Siglo XXI, Unidad de Investigación Médica en Enfermedades Infecciosas y Parasitarias, Mexico City, Mexico • Instituto Mexicano del Seguro Social (IMSS), Centro Médico Nacional La Raza. Hospital de Infectología, Mexico City, Mexico • Clínica Comfamiliar Risaralda, Pereira, Colombia • Bellvitge Biomedical Research Institute (IDIBELL), Neurometabolic Diseases Laboratory, L'Hospitalet de Llobregat, Spain • Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain • Hospital Ophir Loyola, Departamento de Ensino e Pesquisa, Belém, Pará, Brazil • Unidad de Cuidados Intensivos, Hospital Clínico Universitario de Santiago (CHUS), Sistema Galego de Saúde (SERGAS), Santiago de Compostela, Spain • Department of Preventive Medicine and Public Health, School of Medicine, Universidad Autónoma de Madrid, Madrid, Spain • IdiPaz (Instituto de Investigación Sanitaria Hospital Universitario La Paz), Madrid, Spain • IMDEA-Food Institute, CEI UAM+CSIC, Madrid, Spain • Complejo Asistencial Universitario de León, León, Spain • Instituto de Investigación Biosanitaria de Granada (ibs GRANADA), Granada, Spain • Universidad de Granada, Departamento Bioquímica, Biología Molecular e Inmunología III, Granada, Spain • Hospital Infanta Elena, Allergy Unit, Valdemoro, Madrid, Spain • Instituto de Investigación Sanitaria-Fundación Jiménez Díaz University Hospital - Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain • Faculty of Medicine, Universidad Francisco de Vitoria, Madrid, Spain • Hospital Universitario Infanta Leonor, Madrid, Spain • Complutense University of Madrid, Madrid, Spain • Gregorio Marañón Health Research Institute (IiSGM), Madrid, Spain • Haemostasis and Thrombosis Unit, Hospital de la Santa Creu i Sant Pau, IIB Sant Pau, Barcelona, Spain • Hospital Clínico Universitario de Valladolid, Servicio de Anestesiología y Reanimación, Valladolid, Spain • Universidad de Valladolid, Departamento de Cirugía, Valladolid, Spain • Hospital Clínico Universitario de Valladolid, Servicio de Hematología y Hemoterapia, Valladolid, Spain • Hospital de Niños Ricardo Gutierrez, Buenos Aires, Argentina • Fundación Universitaria de Ciencias de la Salud • Spanish National Cancer Research Centre, Familial Cancer Clinical Unit, Madrid, Spain • University Hospital of Burgos, Burgos, Spain • Universidad Simón Bolívar, Facultad de Ciencias de la Salud, Barranquilla, Colombia • Centro para el Desarrollo de la Investigación Científica, Paraguay • Research Center and Memory clinic, ACE Alzheimer Center Barcelona, Universitat Internacional de Catalunya, Spain • Centre for Biomedical Network Research on Neurodegenerative Diseases (CIBERNED), Instituto de Salud Carlos III, Madrid, Spain • CIEN Foundation/Queen Sofia Foundation Alzheimer Center, Madrid, Spain • Hospital Universitario de Valme, Unidad Clínica de Enfermedades Infecciosas y Microbiología, Sevilla, Spain • Sección Genética Médica - Servicio de Pediatría, Hospital Clínico Universitario Virgen de la Arrixaca, Servicio Murciano de Salud, Murcia, Spain • Departamento Cirugía, Pediatría, Obstetricia y Ginecología, Facultad de Medicina, Universidad de Murcia (UMU), Murcia, Spain • Grupo Clínico Vinculado, Centre for Biomedical Network Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III, Madrid, Spain • Department of Anthropology, University of Toronto at Mississauga, Mississauga, Ontario, Canada • Tecnológico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Monterrey, Mexico • Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Instituto de Investigación Sanitaria de Canarias, Santa Cruz de Tenerife, Spain • Centre for Biomedical Network Research on Respiratory Diseases (CIBERES), Instituto de Salud Carlos III, Madrid, Spain • Department of Clinical Sciences, University Fernando Pessoa Canarias, Las Palmas de Gran Canaria, Spain

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

The genetic basis of severe COVID-19 has been thoroughly studied, and many genetic risk factors shared between populations have been identified. However, reduced sample sizes from non-European groups have limited the discovery of population-specific common risk loci. In this second study nested in the SCOURGE consortium, we conducted a GWAS for COVID-19 hospitalization in admixed Americans, comprising a total of 4,702 hospitalized cases recruited by SCOURGE and seven other participating studies in the COVID-19 Host Genetic Initiative. We identified four genome-wide significant associations, two of which constitute novel loci and were first discovered in Latin American populations (*BAZZ2B* and *DDIAS*). A trans-ethnic meta-analysis revealed another novel cross-population risk locus in *CREBBP*. Finally, we assessed the performance of a cross-ancestry polygenic risk score in the SCOURGE admixed American cohort. This study constitutes the largest GWAS for COVID-19 hospitalization in admixed Latin Americans conducted to date. This allowed to reveal novel risk loci and emphasize the need of considering the diversity of populations in genomic research.

eLife assessment

The authors conduct a **valuable** GWAS meta-analysis for COVID-19 hospitalization in admixed American populations and prioritized risk variants and genes. The evidence supporting the claims of the authors is **solid**. The work will be of interest to scientists studying the genetic basis of COVID pathogenesis.

<https://doi.org/10.7554/eLife.93666.2.sa4>

Introduction

To date, more than 50 loci associated with COVID-19 susceptibility, hospitalization, and severity have been identified using genome-wide association studies (GWAS)^{1,2}. The COVID-19 Host Genetics Initiative (HGI) has made significant efforts³ to augment the power to identify disease loci by recruiting individuals from diverse populations and conducting a trans-ancestry meta-analysis. Despite this, the lack of genetic diversity and a focus on cases of European ancestries still predominate in the studies^{4,5}. In addition, while trans-ancestry meta-analyses are a powerful approach for discovering shared genetic risk variants with similar effects across populations⁶, they may fail to identify risk variants that have larger effects on particular underrepresented populations. Genetic disease risk has been shaped by the particular evolutionary history of populations and environmental exposures⁷. Their action is particularly important for infectious diseases due to the selective constraints that are imposed by host–pathogen interactions^{8,9}. Literature examples of this in COVID-19 severity include a *DOCK2* gene variant in East Asians¹⁰ and frequent loss-of-function variants in *IFNAR1* and *IFNAR2* genes in Polynesian and Inuit populations, respectively^{11,12}.

Including diverse populations in case–control GWAS studies with unrelated participants usually requires a prior classification of individuals in genetically homogeneous groups, which are typically analyzed separately to control the population stratification effects¹³. Populations with recent admixture impose an additional challenge to GWASs due to their complex genetic diversity and linkage disequilibrium (LD) patterns, requiring the development of alternative approaches

and a careful inspection of results to reduce false positives due to population structure⁷. In fact, there are benefits in study power from modeling the admixed ancestries either locally, at the regional scale in the chromosomes, or globally, across the genome, depending on factors such as the heterogeneity of the risk variant in frequencies or the effects among the ancestry strata¹⁴. Despite the development of novel methods specifically tailored for the analysis of admixed populations¹⁵, the lack of a standardized analysis framework and the difficulties in confidently clustering admixed individuals into particular genetic groups often leads to their exclusion from GWAS.

The Spanish Coalition to Unlock Research on Host Genetics on COVID-19 (SCOURGE) recruited COVID-19 patients between March and December 2020 from hospitals across Spain and from March 2020 to July 2021 in Latin America (<https://www.scourge-covid.org>). A first GWAS of COVID-19 severity among Spanish patients of European descent revealed novel disease loci and explored age- and sex-varying effects of genetic factors¹⁶. Here, we present the findings of a GWAS meta-analysis in admixed Latin American (AMR) populations, comprising individuals from the SCOURGE Latin-American cohort and the HGI studies, which allowed us to identify two novel severe COVID-19 loci, *BAZ2B* and *DDIAS*. Further analyses modeling the admixture from three genetic ancestral components and performing a trans-ethnic meta-analysis led to the identification of an additional risk locus near *CREBBP*. We finally assessed a cross-ancestry polygenic risk score model with variants associated with critical COVID-19.

Results

Meta-analysis of COVID-19 hospitalization in admixed Americans

Study cohorts

Within the SCOURGE consortium, we included 1,608 hospitalized cases and 1,887 controls (not hospitalized COVID-19 patients) from Latin American countries and from recruitments of individuals of Latin American descent conducted in Spain (Supplementary Table 1). Quality control details and estimation of global genetic inferred ancestry (GIA) (Supplementary Figure 1) are described in Methods, whereas clinical and demographic characteristics of patients included in the analysis are shown in **Table 1**. Summary statistics from the SCOURGE cohort were obtained under a logistic mixed model with the SAIGE model (Methods). Another seven studies participating in the COVID-19 HGI consortium were included in the meta-analysis of COVID-19 hospitalization in admixed Americans (**Figure 1**).

GWAS meta-analysis

We performed a fixed-effects GWAS meta-analysis using the inverse of the variance as weights for the overlapping markers. The combined GWAS sample size consisted of 4,702 admixed AMR hospitalized cases and 68,573 controls.

This GWAS meta-analysis revealed genome-wide significant associations at four risk loci (**Table 2**, **Figure 2**), two of which (*BAZ2B* and *DDIAS*) were novel discoveries. Four lead variants were identified, linked to other 310 variants (Supplementary Tables 2-3). A gene-based association test revealed a significant association in *BAZ2B* and in previously known COVID-19 risk loci: *LZTFL1*, *XCR1*, *FYCO1*, *CCR9*, and *IFNAR2* (Supplementary Table 4).

Located within the *BAZ2B* gene, the sentinel variant rs13003835 is an intronic variant associated with an increased risk of COVID-19 hospitalization (Odds Ratio [OR]=1.20, 95% Confidence Interval [CI]=1.12-1.27, $p=3.66 \times 10^{-8}$). This association was not previously reported in any GWAS of COVID-

Table 1.

Demographic characteristics of the SCOURGE Latin American cohort.

Variable	Non Hospitalized N = 1,887	Hospitalized N = 1,625
Age – mean years $\pm$ SD	39.1 $\pm$ 11.9	54.1 $\pm$ 14.5
Sex - N (%)		
Female (%)	1253 (66.4)	668 (41.1)
GIA* – % mean $\pm$ SD		
European	54.4 $\pm$ 16.2	39.4 $\pm$ 20.7
African	15.3 $\pm$ 12.7	9.1 $\pm$ 11.6
Native American	30.3 $\pm$ 19.8	51.3 $\pm$ 26.5
Comorbidities - N (%)		
Vascular/endocrinological	488 (25.9)	888 (64.5)
Cardiac	60 (3.2)	151 (9.3)
Nervous	15 (0.8)	61 (3.8)
Digestive	14 (0.7)	33 (2.0)
Onco-hematological	21 (1.1)	48 (3.00)
Respiratory	76 (4.0)	118 (7.3)

*Global genetic inferred ancestry.

Figure 1.

Flow chart of this study.

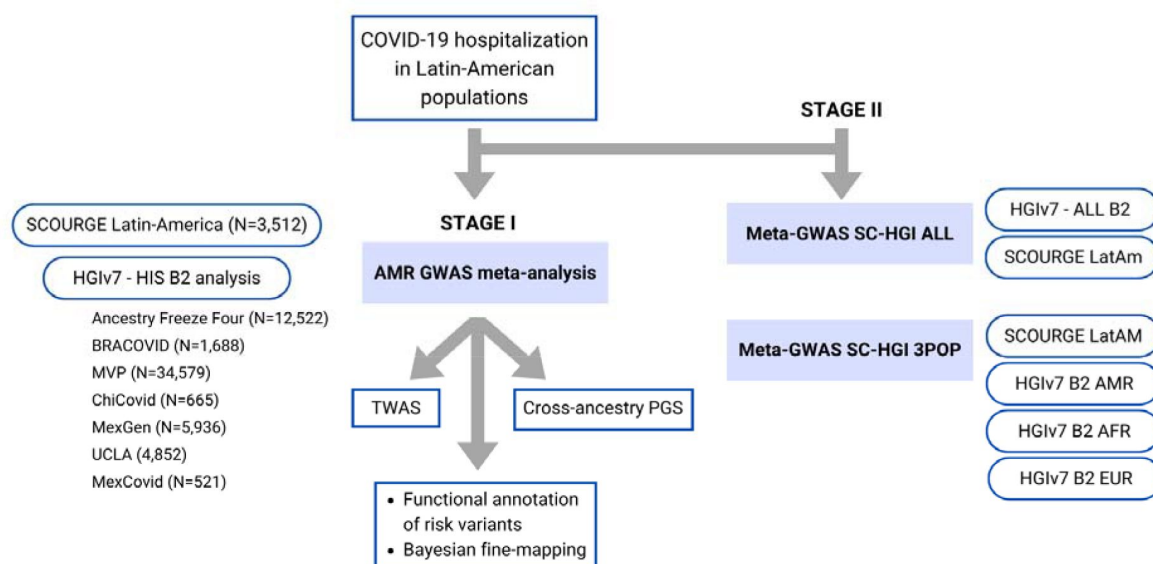


Table 2.

Lead independent variants in the admixed AMR GWAS meta-analysis.

<i>SNP rsID</i>	<i>chr:pos</i>	<i>EA</i>	<i>NEA</i>	<i>OR (95% CI)</i>	<i>P-value</i>	<i>EAF</i> <i>cases</i>	<i>EAF</i> <i>controls</i>	<i>Nearest gene</i>	<i>Mamba</i> <i>PIP</i>
<i>rs13003835</i>	2:159407982	T	C	1.20 (1.12-1.27)	3.66E-08	0.563	0.429	<i>BAZ2B</i>	0.30
<i>rs35731912</i>	3:45848457	T	C	1.65 (1.47-1.85)	6.30E-17	0.087	0.056	<i>LZTFL1</i>	0.95
<i>rs2477820</i>	6:41535254	A	T	0.84 (0.79-0.89)	1.89E-08	0.453	0.517	<i>FOXP4-AS1</i>	0.18
<i>rs77599934</i>	11:82906875	G	A	2.27 (1.7-3.04)	2.26E-08	0.016	0.011	<i>DDIAS</i>	0.95

EA: effect allele; NEA: noneffect allele; EAF: effect allele frequency in the SCOURGE study.

Table 3.

Novel variants in the SC-HGI_{ALL} and SC-HGI_{3POP} meta-analyses (with respect to HGIv7).

Independent signals after LD clumping.

<i>SNP rsID</i>	<i>chr:pos</i>	<i>EA</i>	<i>NEA</i>	<i>OR (95% CI)</i>	<i>P-value</i>	<i>Nearest gene</i>	<i>Analysis</i>
<i>rs76564172</i>	16:3892266	T	G	1.31 (1.19-1.44)	9.64E-09	<i>CREBBP</i>	SC-HGI _{3POP}
<i>rs66833742</i>	19:4063488	T	C	0.94 (0.92-0.96)	1.89E-08	<i>ZBTB7A</i>	SC-HGI _{3POP}
<i>rs66833742</i>	19:4063488	T	C	0.94 (0.92-0.96)	2.50E-08	<i>ZBTB7A</i>	SC-HGI _{ALL}
<i>rs2876034</i>	20:6492834	A	T	0.95 (0.93-0.97)	2.83E-08	<i>CASC20</i>	SC-HGI _{ALL}

EA: effect allele; NEA: non-effect allele.

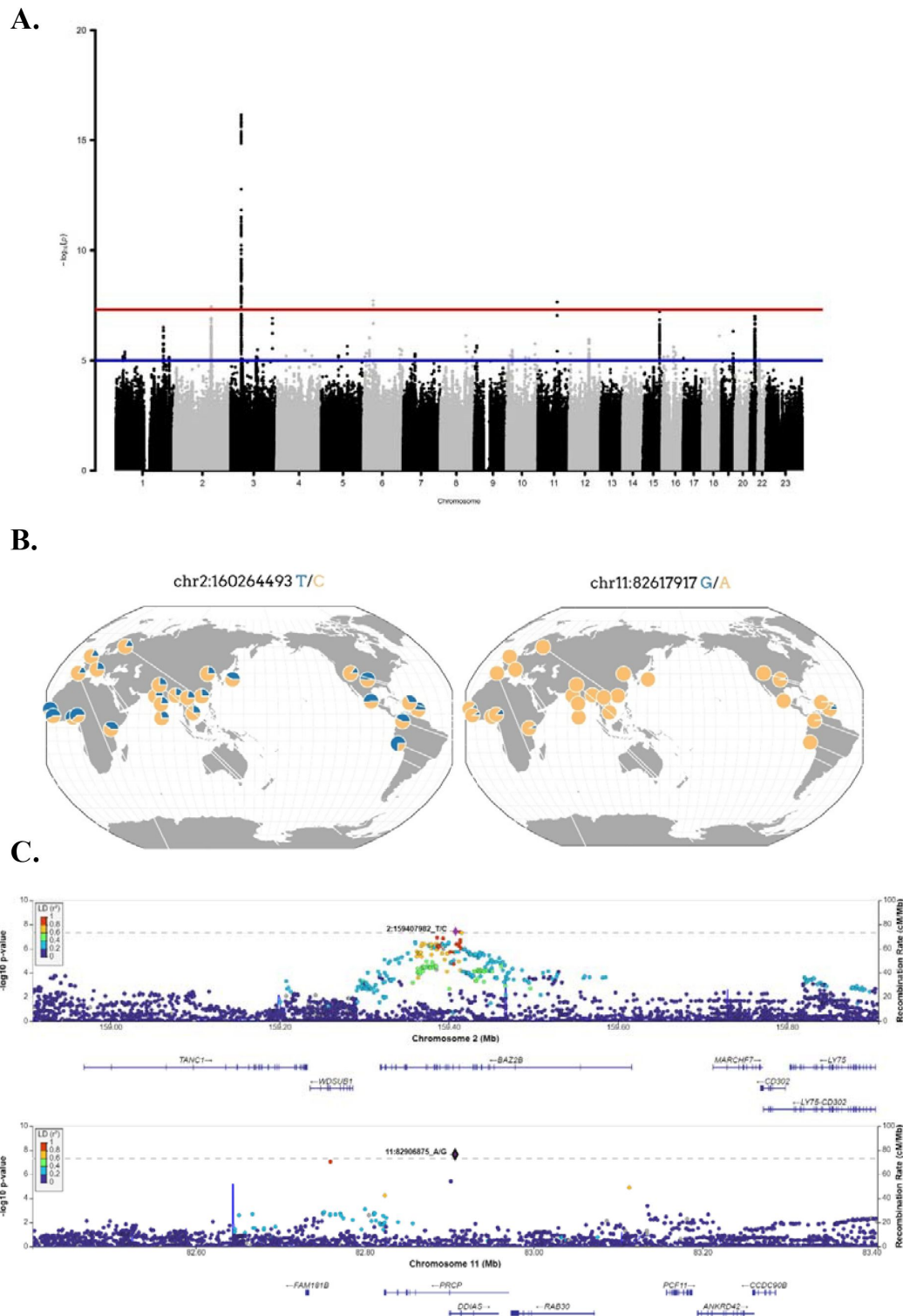


Figure 2.

A) Manhattan plot for the admixed AMR GWAS meta-analysis. Probability thresholds at $p=5 \times 10^{-8}$ and $p=5 \times 10^{-5}$ are indicated by the horizontal lines. Genome-wide significant associations with COVID-19 hospitalizations were found on chromosome 2 (within *BAZ2B*), chromosome 3 (within *LZTFL1*), chromosome 6 (within *FOXP4*), and chromosome 11 (within *DDIAS*). A Quantile–Quantile plot is shown in supplementary Figure 2. **B)** Regional association plots for rs1003835 at chromosome 2 and rs77599934 at chromosome 11; **C)** Allele frequency distribution across the 1000 Genomes Project populations for the lead variants rs1003835 and rs77599934. Retrieved from *The Geography of Genetic Variants Web* or GGV.

19 published to date. Interestingly, rs13003835 did not reach significance ($p=0.972$) in the COVID-19 HGI trans-ancestry meta-analysis including the five population groups¹⁷.

The other novel risk locus is led by the sentinel variant rs77599934, a rare intronic variant located in chromosome 11 within *DDIAS* and associated with the risk of COVID-19 hospitalization (OR=2.27, 95% CI=1.70-3.04, $p=2.26\times 10^{-8}$).

We also observed a suggestive association with rs2601183 in chromosome 15, which is located between *ZNF774* and *IQGAP1* (allele-G OR=1.20, 95% CI=1.12-1.29, $p=6.11\times 10^{-8}$, see Supplementary Table 2), which has not yet been reported in any other GWAS of COVID-19 to date.

The GWAS meta-analysis also pinpointed two significant variants at known loci, *LZTFL1* and *FOXP4*. The SNP rs35731912 was previously associated with COVID-19 severity in EUR populations¹⁷, and it was mapped to *LZTFL1*. While rs2477820 is a novel risk variant within the *FOXP4* gene, it has a moderate LD ($r^2=0.295$) with rs2496644, which has been linked to COVID-19 hospitalization¹⁸. This is consistent with the effects of LD in tag-SNPs when conducting GWAS in diverse populations.

None of the lead variants was associated with the comorbidities included in **Table 1**.

Functional mapping of novel risk variants

Variants belonging to the lead loci were prioritized by positional and expression quantitative trait loci (eQTL) mapping with FUMA, resulting in 31 mapped genes (Supplementary Table 5). Within the region surrounding the lead variant in chromosome 2, FUMA prioritized four genes in addition to *BAZ2B* (*PLA2R1*, *LY75*, *WDSUB1*, and *CD302*). rs13003835 (allele C) is an eQTL of *LY75* in the esophagus mucosa (NES=0.27) and of *BAZ2B-AS* in whole blood (NES=0.27), while rs2884110 ($R^2=0.85$) is an eQTL of *LY75* in lung (NES=0.22). As for the chromosome 11, rs77599934 (allele G) is in moderate-to-strong LD ($r^2=0.776$) with rs60606421 (G deletion, allele -) which is an eQTL associated with a reduced expression of *DDIAS* in the lungs (NES=-0.49, allele -). Associations with expression are shown in the supplementary Figure 4. The sentinel variant for the region in chromosome 16 is in perfect LD ($r^2=1$) with rs601183, an eQTL of *ZNF774* in the lung.

Bayesian fine mapping

We performed different approaches to narrow down the prioritized loci to a set of most likely genes driving the associations. First, we computed credible sets at the 95% confidence level for causal variants and annotated them with VEP (and V2G aggregate scoring. The 95% confidence credible set from the region of chromosome 2 around rs13003835 included 76 variants, which can be found in the Supplementary Table 6 (VEP and V2G annotations are included in the Supplementary Tables 7 and 8). The V2G score prioritized *BAZ2B* as the most likely gene driving the association. However, the approach was unable to converge allocating variants in a 95% confidence credible set for the region in chromosome 11.

Transcriptome-wide association study (TWAS)

Five novel genes, namely, *SLC25A37*, *SMARCC1*, *CAMP*, *TYW3*, and *S100A12* (supplementary Table 9), were found to be significantly associated in the cross-tissue TWAS. To our knowledge, these genes have not been reported previously in any COVID-19 TWAS or GWAS analyses published to date. In the single tissue analyses, *ATP5O* and *CXCR6* were significantly associated in the lungs, *CCR9* was significantly associated in whole blood, and *IFNAR2* and *SLC25A37* were associated in lymphocytes.

Likewise, we carried out TWAS analyses using the models trained in the admixed populations. However, no significant gene pairs were detected in this case. The top 10 genes with the lowest p-values for each of the datasets (Puerto-Ricans, Mexicans, African-Americans and pooled cohorts) are shown in the Supplementary Table 10. Although not significant, *KCNC3* was repeated in the four analyses, whereas *MAPKAPK3*, *NAPSA* and *THAP5* were repeated in 3 out of 4. Both *NAPSA* and *KCNC3* are located in the chromosome 19 and were reported in the latest HGI meta-analysis¹⁹.

All mapped genes from analyses conducted in AMR populations are shown in **Figure 5**.

Genetic architecture of COVID-19 hospitalization in AMR populations

Allele frequencies of rs13003835 and rs77599934 across ancestries

Neither rs13003835 (*BAZ2B*) nor rs77599934 (*DDIAS*) were significantly associated in the COVID-19 HGI B2 cross-population or population-specific meta-analyses. Thus, we investigated their allele frequencies (AF) across populations and compared their effect sizes.

According to gnomAD v3.1.2, the T allele at rs13003835 (*BAZ2B*) has an AF of 43% in admixed AMR groups, while AF is lower in the EUR populations (16%) and in the global sample (29%). Local ancestry inference (LAI) reported by gnomAD shows that within the Native-American component, the risk allele T is the major allele, whereas it is the minor allele within the African and European LAI components. These large differences in AF might be the reason underlying the association found in AMR populations. However, when comparing effect sizes between populations, we found that they were in opposite directions between SAS-AMR and EUR-AFR-EAS and that there was large heterogeneity among them (**Figure 3**). We queried SNPs within 50kb windows of the lead variant in each of the other populations that had p-value<0.01. The variant with the lowest p-value in the EUR population was rs559179177 ($p=1.72\times 10^{-4}$), which is in perfect LD ($r^2=1$) in the 1KGP EUR population with our sentinel variant (rs13003835), and in moderate LD $r^2=0.4$ in AMR populations. Since this variant was absent from the AMR analysis, probably due to its low frequency, it could not be meta-analyzed. Power calculations revealed that the EUR analysis was underpowered for this variant to achieve genome-wide significance (77.6%, assuming an effect size of 0.46, EAF= 0.0027, and number of cases/controls as shown in the HGI website for B2-EUR). In the cross-population meta-analysis (B2-ALL), rs559179177 obtained a p-value of 5.9×10^{-4} .

rs77599934 (*DDIAS*) had an AF of 1.1% for the G allele in the nonhospitalized controls (**Table 2**), in line with the recorded gnomAD AF of 1% in admixed AMR groups. This variant has the potential to be a population-specific variant, given the allele frequencies in other population groups, such as EUR (0% in Finnish, 0.025% in non-Finnish), EAS (0%) and SAS (0.042%), and its greater effect size over AFR populations (**Figure 3**). Examining the LAI, the G allele occurs at a 10.8% frequency in the African component, while it is almost absent in the Native-American and European. Due to its low MAF, rs77599934 was not analyzed in the COVID-19 HGI B2 cross-population meta-analysis and was only present in the HGI B2 AFR population-specific meta-analysis, precluding the comparison (**Figure 3**). For this reason, we retrieved the variant with the lowest p-value within a 50 kb region around rs77599934 in the COVID-19 HGI cross-population analysis to investigate whether it was in moderate-to-strong LD with our sentinel variant. The variant with the smallest p-value was rs75684040 (OR=1.07, 95% CI=1.03-1.12, $p=1.84\times 10^{-3}$). However, LD calculations using the 1KGP phase 3 dataset indicated that rs77599934 and rs75684040 were poorly correlated ($r^2=0.11$). As for AFR populations, the variant with the lowest p-value was rs138860115 ($p=8.3\times 10^{-3}$), but it was not correlated with the lead SNP of this locus.

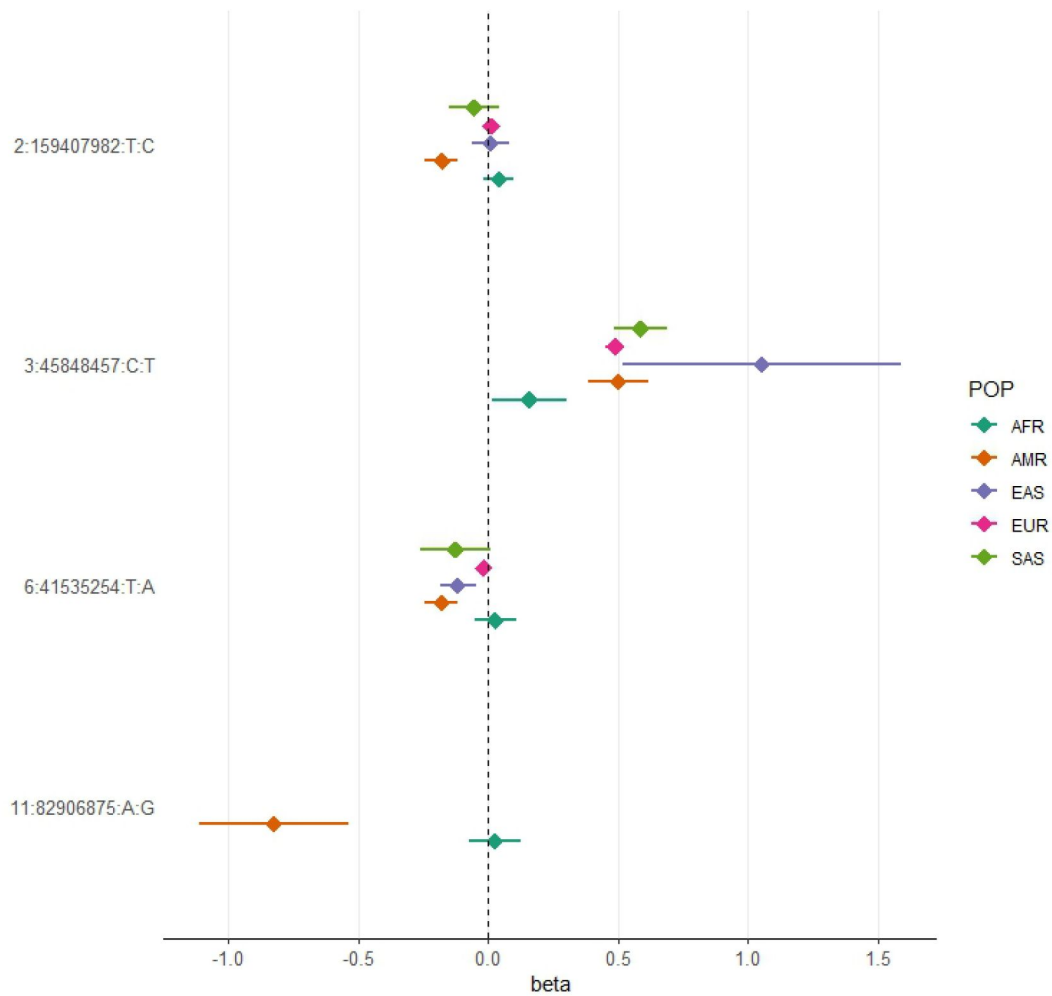


Figure 3.

Forest plot showing effect sizes and the corresponding confidence intervals for the sentinel variants identified in the AMR meta-analysis across populations.

All beta values with their corresponding CIs were retrieved from the B2 population-specific meta-analysis from the HGI v7 release, except for AMR, for which the beta value and IC from the HGI_{AMR}-SCOURGE meta-analysis are represented.

Cross-population meta-analyses

We carried out two cross-ancestry inverse variance-weighted fixed-effects meta-analyses with the admixed AMR GWAS meta-analysis results to evaluate whether the discovered risk loci replicated when considering other population groups. In doing so, we also identified novel cross-population COVID-19 hospitalization risk loci.

First, we combined the SCOURGE Latin American GWAS results with the HGI B2 ALL analysis (supplementary Table 11). We refer to this analysis as the SC-HGI_{ALL} meta-analysis. Out of the 40 genome-wide significant loci associated with COVID-19 hospitalization in the last HGI release¹, this study replicated 39, and the association was stronger than in the original study in 29 of those (supplementary Table 12). However, the variant rs13003835 located in *BAZ2B* did not replicate (OR=1.00, 95% CI=0.98-1.03, p=0.644).

In this cross-ancestry meta-analysis, we replicated two associations that were not found in HGIv7, albeit they were sentinel variants in the latest GenOMICC meta-analysis². We found an association at the *CASC20* locus led by the variant rs2876034 (OR=0.95, 95% CI=0.93-0.97, p=2.83×10⁻⁸). This variant is in strong LD with the sentinel variant of that study (rs2326788, r²=0.92), which was associated with critical COVID-19². In addition, this meta-analysis identified the variant rs66833742 near *ZBTB7A* associated with COVID-19 hospitalization (OR=0.94, 95% CI=0.92-0.96, p=2.50×10⁻⁸). Notably, rs66833742 or its perfect proxy rs67602344 (r²=1) are also associated with upregulation of *ZBTB7A* in whole blood and in esophageal mucosa. This variant was previously associated with COVID-19 hospitalization².

In a second analysis, we also explored the associations across the defined admixed AMR, EUR, and AFR ancestral sources by combining through meta-analysis the SCOURGE Latin American GWAS results with the HGI studies in EUR, AFR, and admixed AMR and excluding those from EAS and SAS (supplementary Table 13). We refer to this as the SC-HGI_{3POP} meta-analysis. The association at rs13003835 (*BAZ2B*, OR=1.01, 95% CI=0.98-1.03, p=0.605) was not replicated, and rs77599934 near *DDIAS* could not be assessed, although the association at the *ZBTB7A* locus was confirmed (rs66833742, OR=0.94, 95% CI=0.92-0.96, p=1.89×10⁻⁸). The variant rs76564172 located near *CREBBP* also reached statistical significance (OR=1.31, 95% CI=1.25-1.38, p=9.64×10⁻⁹). The sentinel variant of the region linked to *CREBBP* (in the trans-ancestry meta-analysis) was also subjected a Bayesian fine mapping (supplementary Table 6). Eight variants were included in the credible set for the region in chromosome 16 (meta-analysis SC-HGI_{3POP}).

Polygenic risk score models

Using the 49 variants associated with disease severity that are shared across populations according to the HGIv7, we constructed a polygenic risk score (PGS) model to assess its generalizability in the admixed AMR (supplementary Table 14). First, we calculated the PGS for the SCOURGE Latin Americans and explored the association with COVID-19 hospitalization under a logistic regression model. The PGS model was associated with a 1.48-fold increase in COVID-19 hospitalization risk per every PGS standard deviation. It also contributed to explaining a slightly larger variance ($\Delta R^2=1.07\%$) than the baseline model. Subsequently, we divided the individuals into PGS deciles and percentiles to assess their risk stratification. The median percentile among controls was 40, while in cases, it was 63. Those in the top PGS decile exhibited a 2.89-fold (95% CI=2.37-3.54, p=1.29×10⁻⁷) greater risk compared to individuals in the deciles between 4 and 6 (corresponding to a score of the median distribution).

We also examined the distribution of PGS scores across a 5-level severity scale to further determine if there was any correspondence between clinical severity and genetic risk. Median PGS scores were lower in the asymptomatic and mild groups, whereas higher median scores were observed in the moderate, severe, and critical patients (Figure 4). We fitted a multinomial

model using the asymptomatic class as a reference and calculated the OR for each category (supplementary Table 15), observing that the disease genetic risk was similar among asymptomatic, mild, and moderate patients. Given that the PGS was built with variants associated with critical disease and/or hospitalization and that the categories severe and critical correspond to hospitalized patients, these results underscore the ability of cross-ancestry PGS for risk stratification even in an admixed population.

Discussion

We have conducted the largest GWAS meta-analysis of COVID-19 hospitalization in admixed AMR to date. While the genetic risk basis discovered for COVID-19 is largely shared among populations, trans-ancestry meta-analyses on this disease have primarily included EUR samples. This dominance of GWAS in Europeans and the subsequent bias in sample sizes can mask population-specific genetic risks (i.e., variants that are monomorphic in some populations) or be less powered to detect risk variants having higher allele frequencies in population groups other than Europeans. In this sense, after combining data from admixed AMR patients, we found two risk loci that were first discovered in a GWAS of Latin American populations. Interestingly, the sentinel variant rs77599934 in the *DDIAS* gene is a rare coding variant (~1% for allele G) with a large effect on COVID-19 hospitalization that is nearly monomorphic in most of the other populations. This has likely led to its exclusion from the cross-population meta-analyses conducted to date, remaining undetectable.

Fine mapping of the region harboring *DDIAS* did not reveal further information about which gene could be the more prone to be causal or about the functional consequences of the risk variant, but our sentinel variant was in strong LD with an eQTL that associated with reduced gene expression of *DDIAS* in the lung. *DDIAS*, known as damage-induced apoptosis suppressor gene, is itself a plausible candidate gene. It has been linked to DNA damage repair mechanisms: research has shown that depletion of *DDIAS* leads to an increase in ATM phosphorylation and the formation of p53-binding protein (53BP1) foci, a known biomarker of DNA double-strand breaks, suggesting a potential role in double-strand break repair²⁰. Interestingly, a study found that infection by SARS-CoV-2 also triggered the phosphorylation of ATM kinase and inhibited repair mechanisms, causing the accumulation of DNA damage²¹. This gene has also been proposed as a potential biomarker for lung cancer after finding that it interacts with STAT3 in lung cancer cells, regulating IL-6^{22,23} and thus mediating inflammatory processes, while another study determined that its blockade inhibited lung cancer cell growth²⁴. Another prioritized gene from this region was *PRCP*, an angiotensinase which shares substrate specificity with ACE2 receptor. It has been positively linked to hypertension and some studies have raised hypotheses on its role in COVID-19 progression, particularly in relation to the development of pro-thrombotic events^{25,26}.

The risk region found in chromosome 2 harbors more than one gene. The lead variant rs13003835 is located within *BAZ2B*, and it increases the expression of the antisense *BAZ2B* gene in whole blood. *BAZ2B* encodes one of the regulatory subunits of the Imitation switch (ISWI) chromatin remodelers²⁷ constituting the BRF-1/BRF-5 complexes with SMARCA1 and SMARCA5, respectively. Interestingly, it was discovered that *lnc-BAZ2B* promotes macrophage activation through regulation of *BAZ2B* expression. Its over-expression resulted in pulmonary inflammation and elevated levels of *MUC5AC* in mice with asthma²⁸. This variant was also an eQTL for *LY75* (encoding lymphocyte antigen 75) in the esophageal mucosa tissue. Lymphocyte antigen 75 is involved in immune processes through antigen presentation in dendritic cells and endocytosis²⁹ and has been associated with inflammatory diseases, representing a compelling candidate for the region. Increased expression of *LY75* has been detected within hours after infection by SARS-CoV-2^{30,31}. It is worth noting that differences in AF for this variant suggest that analyses in AMR populations might be more powered to detect the association, supporting the necessity of population-specific studies.

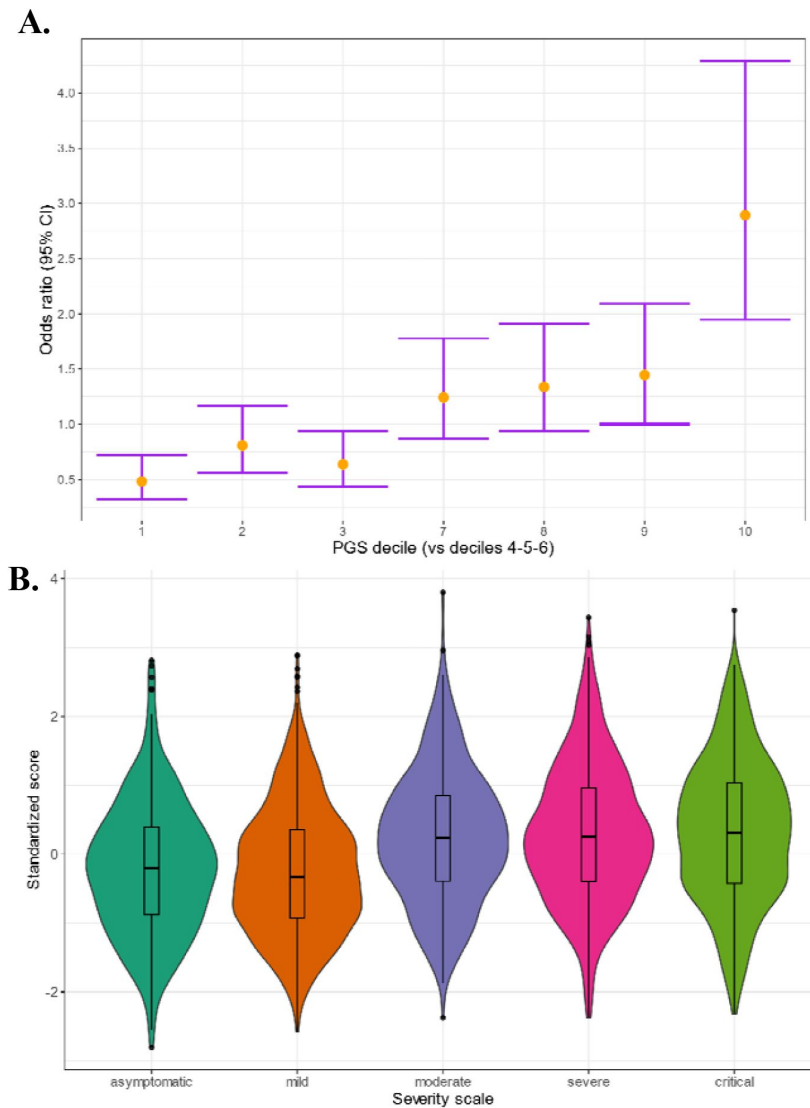


Figure 4.

(A) Polygenic risk stratified by PGS deciles comparing each risk group against the lowest risk group (OR-95% CI); (B) Distribution of the PGS scores in each of the severity scale classes.

0-Asymptomatic, 1-Mild disease, 2-Moderate disease, 3-Severe disease, 4-Critical disease.

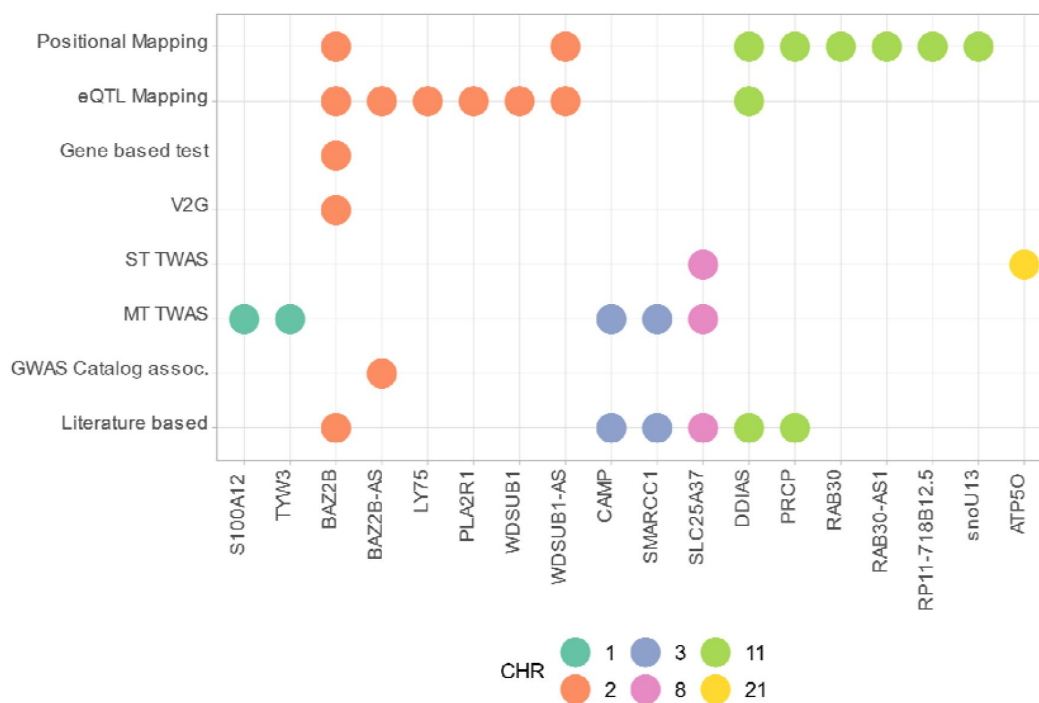


Figure 5.

Summary of the results from gene prioritization strategies used for genetic associations in AMR populations.

GWAS catalog association for *BAZ2B-AS* was with FEV/FCV ratio. Literature based evidence is further explored in discussion.

A third novel risk region was observed on chromosome 15 between the *IQGAP1* and *ZNF774* genes, although it did not reach genome-wide significance.

Secondary analyses revealed five TWAS-associated genes, some of which have already been linked to severe COVID-19. In a comprehensive multitissue gene expression profiling study³², decreased expression of *CAMP* and *S100A8/S100A9* genes in patients with severe COVID-19 was observed, while another study detected the upregulation of *SCL25A37* among patients with severe COVID-19³³. *SMARCC1* is a subunit of the SWI/SNF chromatin remodeling complex that has been identified as proviral for SARS-CoV-2 and other coronavirus strains through a genome-wide screen³⁴. This complex is crucial for *ACE2* expression and viral entry into the cell³⁵. However, it should be noted that using eQTL mostly from European populations such as those in GTEx could result in reduced power to detect associations.

To explore the genetic architecture of the trait among admixed AMR populations, we performed two cross-ancestry meta-analyses including the SCOURGE Latin American cohort GWAS findings. We found that the two novel risk variants were not associated with COVID-19 hospitalization outside the population-specific meta-analysis, highlighting the importance of complementing trans-ancestry meta-analyses with group-specific analyses. Notably, this analysis did not replicate the association at the *DSTYK* locus, which was associated with severe COVID-19 in Brazilian individuals with higher European admixture³⁶. This lack of replication aligns with the initial hypothesis of that study suggesting that the risk haplotype was derived from European populations, as we reduced the weight of this ancestral contribution in our study by excluding those individuals.

Moreover, these cross-ancestry meta-analyses pointed to three loci that were not genome-wide significant in the HGiv7 ALL meta-analysis: a novel locus at *CREBBP* and two loci at *ZBTB7A* and *CASC20* that were reported in another meta-analysis. *CREBBP* and *ZBTB7A* achieved a stronger significance when considering only the EUR, AFR, and admixed AMR GIA groups. According to a recent study, elevated levels of the *ZBTB7A* gene promote a quasihomeostatic state between coronaviruses and host cells, preventing cell death by regulating oxidative stress pathways³⁷. This gene is involved in several signaling pathways, such as B and T-cell differentiation³⁸. On a separate note, *CREBBP* encodes the CREB binding protein (CBP), which is involved in transcriptional activation and is known to positively regulate the type I interferon response through virus-induced phosphorylation of IRF-3³⁹. In addition, the CREBP/CBP interaction has been implicated in SARS-CoV-2 infection⁴⁰ via the cAMP/PKA pathway. In fact, cells with suppressed *CREBBP* gene expression exhibit reduced replication of the so-called Delta and Omicron SARS-CoV-2 variants⁴⁰.

We developed a cross-population PGS model, which effectively stratified individuals based on their genetic risk and demonstrated consistency with the clinical severity classification of the patients. Only a few polygenic scores were derived from COVID-19 GWAS data. Horowitz et al. (2022)⁴¹ developed a score using 6 and 12 associated variants (PGS ID: PGP000302) and reported an associated OR (top 10% vs rest) of 1.38 for risk of hospitalization in European populations, whereas the OR for Latin-American populations was 1.56. Since their sample size and the number of variants included in the PGS were lower, direct comparisons are not straightforward. Nevertheless, our analysis provides the first results for a PRS applied to a relatively large AMR cohort, being of value for future analyses regarding PRS transferability.

This study is subject to limitations, mostly concerning sample recruitment and composition. The SCOURGE Latino-American sample size is small, and the GWAS is likely underpowered. Another limitation is the difference in case-control recruitment across sampling regions that, yet controlled for, may reduce the ability to observe significant associations driven by different compositions of the populations. In this sense, the identified risk loci might not replicate in a cohort lacking any of the parental population sources from the three-way admixture. Likewise, we

could not explicitly control for socioenvironmental factors that could have affected COVID-19 spread and hospitalization rates, although genetic principal components are known to capture nongenetic factors. Finally, we must acknowledge the lack of a replication cohort. We used all the available GWAS data for COVID-19 hospitalization in admixed AMR in this meta-analysis due to the low number of studies conducted. Therefore, we had no studies to replicate or validate the results. These concerns may be addressed in the future by including more AMR GWAS in the meta-analysis, both by involving diverse populations in study designs and by supporting research from countries in Latin America.

This study provides novel insights into the genetic basis of COVID-19 severity, emphasizing the importance of considering host genetic factors by using non-European populations, especially of admixed sources. Such complementary efforts can pin down new variants and increase our knowledge on the host genetic factors of severe COVID-19.

Materials and methods

GWAS in Latin Americans from SCOURGE

The SCOURGE Latin American cohort

A total of 3,729 COVID-19-positive cases were recruited across five countries from Latin America (Mexico, Brazil, Colombia, Paraguay, and Ecuador) by 13 participating centers (supplementary Table 1) from March 2020 to July 2021. In addition, we included 1,082 COVID-19-positive individuals recruited between March and December 2020 in Spain who either had evidence of origin from a Latin American country or showed inferred genetic admixture between AMR, EUR, and AFR (with $< 0.05\%$ SAS/EAS). These individuals were excluded from a previous SCOURGE study that focused on participants with European genetic ancestries¹⁶. We used hospitalization as a proxy for disease severity and defined COVID-19-positive patients who underwent hospitalization as a consequence of the infection as cases and those who did not need hospitalization due to COVID-19 as controls.

Samples and data were collected with informed consent after the approval of the Ethics and Scientific Committees from the participating centers and by the Galician Ethics Committee Ref 2020/197. Recruitment of patients from IMSS (in Mexico, City) was approved by the National Committee of Clinical Research from Instituto Mexicano del Seguro Social, Mexico (protocol R-2020-785-082).

Samples and data were processed following normalized procedures. The REDCap electronic data capture tool^{42,43}, hosted at Centro de Investigación Biomédica en Red (CIBER) from the Instituto de Salud Carlos III (ISCIII), was used to collect and manage demographic, epidemiological, and clinical variables. Subjects were diagnosed with COVID-19 based on quantitative PCR tests (79.3%) or according to clinical (2.2%) or laboratory procedures (antibody tests: 16.3%; other microbiological tests: 2.2%).

SNP array genotyping

Genomic DNA was obtained from peripheral blood and isolated using the Chemagic DNA Blood 100 kit (PerkinElmer Chemagen Technologies GmbH), following the manufacturer's recommendations.

Samples were genotyped with the Axiom Spain Biobank Array (Thermo Fisher Scientific) following the manufacturer's instructions in the Santiago de Compostela Node of the National Genotyping Center (CeGen-ISCIII; <http://www.usc.es/cegen>). This array contains probes for genotyping a total

of 757,836 SNPs. Clustering and genotype calling were performed using Axiom Analysis Suite v4.0.3.3 software.

Quality control steps and variant imputation

A quality control (QC) procedure using PLINK 1.9⁴⁴ was applied to both samples and the genotyped SNPs. We excluded variants with a minor allele frequency (MAF) <1%, a call rate <98%, and markers strongly deviating from Hardy-Weinberg equilibrium expectations ($p < 1 \times 10^{-6}$) with mid-p adjustment. We also explored the excess of heterozygosity to discard potential cross-sample contamination. Samples missing >2% of the variants were filtered out. Subsequently, we kept the autosomal SNPs, removed high-LD regions and conducted LD pruning (windows of 1,000 SNPs, with a step size of 80 and a r^2 threshold of 0.1) to assess kinship and estimate the global ancestral proportions. Kinship was evaluated based on IBD values, removing one individual from each pair with $PI_HAT > 0.25$ that showed a Z0, Z1, and Z2 coherent pattern (according to the theoretical expected values for each relatedness level). Genetic principal components (PCs) were calculated with PLINK with the subset of LD pruned variants.

Genotypes were imputed with the TOPMed version r2 reference panel (GRCh38) using the TOPMed Imputation Server, and variants with $Rsq < 0.3$ or with $MAF < 1\%$ were filtered out. A total of 4,348 individuals and 10,671,028 genetic variants were included in the analyses.

Genetic admixture estimation

Global genetic inferred ancestry (GIA), referred to the genetic similarity to the used reference individuals, was estimated with ADMIXTURE⁴⁵ v1.3 software following a two-step procedure. First, we randomly sampled 79 European (EUR) and 79 African (AFR) samples from The 1000 Genomes Project (1KGP)⁴⁶ and merged them with the 79 Native American (AMR) samples from Mao *et al.*⁴⁷ keeping the biallelic SNPs. LD-pruned variants were selected from this merge using the same parameters as in the QC. We then run an unsupervised analysis with $K=3$ to redefine and homogenize the clusters and to compose a refined reference for the analyses by applying a threshold of $\geq 95\%$ of belonging to a particular cluster. As a result, 20 AFR, 18 EUR, and 38 AMR individuals were removed. The same LD-pruned variants data from the remaining individuals were merged with the SCOURGE Latin American cohort to perform supervised clustering and estimate admixture proportions. A total of 471 samples from the SCOURGE cohort with >80% estimated European GIA were removed to reduce the weight of the European ancestral component, leaving a total of 3,512 admixed Latin American (AMR) subjects for downstream analyses.

Association analysis

The results for the SCOURGE Latin American GWAS were obtained by testing for COVID-19 hospitalization as a surrogate of severity. To accommodate the continuum of GIA in the cohort, we opted for a joint testing of all the individuals as a single study using a mixed regression model, as this approach has demonstrated a greater power and sufficient control of population structure⁴⁸. The SCOURGE cohort consisted of 3,512 COVID-19-positive patients: cases ($n=1,625$) were defined as hospitalized COVID-19 patients, and controls ($n=1,887$) were defined as nonhospitalized COVID-19-positive patients.

Logistic mixed regression models were fitted using the SAIGEgds⁴⁹ package in R, which implements the two-step mixed SAIGE⁵⁰ model methodology and the SPA test. Baseline covariables included sex, age (continuous), and the first 10 PCs. To account for potential heterogeneity in the recruitment and hospitalization criteria across the participating countries, we adjusted the models by groups of the recruitment areas classified into six categories: Brazil, Colombia, Ecuador, Mexico, Paraguay, and Spain. This dataset has not been used in any previously published GWAS of COVID-19.

Meta-analysis of Latin American populations

The results of the SCOURGE Latin American cohort were meta-analyzed with the AMR HGI-B2 data, conforming our primary analysis. Summary results from the HGI freeze 7 B2 analysis corresponding to the admixed AMR population were obtained from the public repository (April 8, 2022: <https://www.covid19hg.org/results/r7/>), summing up 3,077 cases and 66,686 controls from seven contributing studies. We selected the B2 phenotype definition because it offered more power, and the presence of population controls not ascertained for COVID-19 does not have a drastic impact on the association results.

The meta-analysis was performed using an inverse-variance weighting method in METAL⁵¹. The average allele frequency was calculated, and variants with low imputation quality ($R_{sq} < 0.3$) were filtered out, leaving 10,121,172 variants for meta-analysis.

Heterogeneity between studies was evaluated with Cochran's Q test. The inflation of results was assessed based on a genomic control (λ).

Replicability of associations

The model-based method MAMBA⁵² was used to calculate the posterior probabilities of replication for each of the lead variants. Variants with $p < 1 \times 10^{-5}$ were clumped and combined with random pruned variants from the 1KGP AMR reference panel. Then, MAMBA was applied to the set of significant and non-significant variants.

Each of the lead variants was also tested for association with the main comorbidities in the SCOURGE cohort with logistic regression models (adjusted by the same base covariables as the GWAS).

Definition of the genetic risk loci and putative functional impact

Definition of lead variant and novel loci

To define the lead variants in the loci that were genome-wide significant, LD-clumping was performed on the meta-analysis data using a threshold $p\text{-value} < 5 \times 10^{-8}$, clump distance=1500 kb, independence set at a threshold $r^2 = 0.1$ and the SCOURGE cohort genotype data as the LD reference panel. Independent loci were deemed as a novel finding if they met the following criteria: 1) $p\text{-value} < 5 \times 10^{-8}$ in the meta-analysis and $p\text{-value} > 5 \times 10^{-8}$ in the HGI B2 ALL meta-analysis or in the HGI B2 AMR and AFR and EUR analyses when considered separately; 2) Cochran's Q-test for heterogeneity of effects is $< 0.05/N_{\text{loci}}$, where N_{loci} is the number of independent variants with $p < 5 \times 10^{-8}$; and 3) the nearest gene has not been previously described in the latest HGIv7 update.

Annotation and initial mapping

Functional annotation was performed with FUMA⁵³ for those variants with a $p\text{-value} < 5 \times 10^{-8}$ or in moderate-to-strong LD ($r^2 > 0.6$) with the lead variants, where the LD was calculated from the 1KGP AMR panel. Genetic risk loci were defined by collapsing LD blocks within 250 kb. Then, genes, scaled CADD v1.4 scores, and RegulomeDB v1.1 scores were annotated for the resulting variants with ANNOVAR in FUMA⁵³. Gene-based analysis was also performed using MAGMA⁵⁴ as implemented in FUMA under the SNP-wide mean model using the 1KGP AMR reference panel. Significance was set at a threshold $p < 2.66 \times 10^{-6}$ (which assumes that variants can be mapped to a total of 18,817 genes).

FUMA allowed us to perform initial gene mapping by two approaches: (1) positional mapping, which assigns variants to genes by physical distance using 10-kb windows; and (2) eQTL mapping based on GTEx v.8 data from whole blood, lungs, lymphocytes, and esophageal mucosa tissues, establishing a false discovery rate (FDR) of 0.05 to declare significance for variant-gene pairs.

Subsequently, to assign the variants to the most likely gene driving the association, we refined the candidate genes by fine mapping the discovered regions

Bayesian fine-mapping

To conduct a Bayesian fine mapping, credible sets for the genetic loci considered novel findings were calculated on the results from each of the three meta-analyses to identify a subset of variants most likely containing the causal variant at the 95% confidence level, assuming that there is a single causal variant and that it has been tested. We used *corrcoverage* (<https://cran.rstudio.com/web/packages/corrcoverage/index.html>) for R to calculate the posterior probabilities of the variant being causal for all variants with a $r^2 > 0.1$ with the leading SNP and within 1 Mb except for the novel variant in chromosome 19, for which we used a window of 0.5 Mb. Variants were added to the credible set until the sum of the posterior probabilities was ≥ 0.95 .

VEP and V2G annotation

We used the Variant-to-Gene (V2G) score to prioritize the genes that were most likely affected by the functional evidence based on expression quantitative trait loci (eQTL), chromatin interactions, in silico functional predictions, and distance between the prioritized variants and transcription start site (TSS), based on data from the Open Targets Genetics portal⁵⁵. Details of the data integration and the weighting of each of the datasets are described with detail here: <https://genetics-docs.opentargets.org/our-approach/data-pipeline>. V2G is a score for ranking the functional genomics evidence that supports the connection between variants and genes (the higher the score the more likely the variant to be functionally implicated on the assigned gene). We used VEP release 111 (URL: <https://www.ensembl.org/info/docs/tools/vep/index.html>; accessed April 10, 2024)⁵⁶ to annotate the following: gene symbol, function (exonic, intronic, intergenic, non-coding RNA, etc.), impact, feature type, feature, and biotype.

We queried the GWAS Catalog (date of accession: 01/07/2024) for evidence of association of each of the prioritized genes with traits related to lung diseases or phenotypes. Lastly, those which were linked to COVID-19, infection, or lung diseases in the revised literature were classified as “literature evidence”.

Transcription-wide association studies

Transcriptome-wide association studies (TWAS) were conducted using the pretrained prediction models with MASHR-computed effect sizes on GTEx v8 datasets^{57,58}. The results from the Latin American meta-analysis were harmonized and integrated with the prediction models through S-PrediXcan⁵⁹ for lung, whole blood, lymphocyte and esophageal mucosal tissues. Statistical significance was set at $p\text{-value} < 0.05$ divided by the number of genes that were tested for each tissue. Subsequently, we leveraged results for all 49 tissues and ran a multitissue TWAS to improve the power for association, as demonstrated recently⁶⁰. TWAS was also performed using recently published gene expression datasets derived from a cohort of African Americans, Puerto Ricans, and Mexican Americans (GALA II-SAGE)⁶¹.

Cross-population meta-analyses

We conducted two additional meta-analyses to investigate the ability of combining populations to replicate our discovered risk loci. This methodology enabled the comparison of effects and the significance of associations in the novel risk loci between the results from analyses that included

or excluded other population groups.

The first meta-analysis comprised the five populations analyzed within HGI (B2-ALL). Additionally, to evaluate the three GIA components within the SCOURGE Latin American cohort⁶², we conducted a meta-analysis of the admixed AMR, EUR, and AFR cohorts (B2). All summary statistics were retrieved from the HGI repository. We applied the same meta-analysis methodology and filters as in the admixed AMR meta-analysis.

Cross-population Polygenic Risk Score

A polygenic risk score (PGS) for critical COVID-19 was derived by combining the variants associated with hospitalization or disease severity that have been discovered to date. We curated a list of lead variants that were 1) associated with either severe disease or hospitalization in the latest HGIv7 release¹ (using the hospitalization weights) or 2) associated with severe disease in the latest GenOMICC meta-analysis² that were not reported in the latest HGI release. A total of 48 markers were used in the PGS model (see supplementary Table 13) since two variants were absent from our study.

Scores were calculated and normalized for the SCOURGE Latin American cohort with PLINK 1.9. This cross-ancestry PGS was used as a predictor for hospitalization (COVID-19-positive patients who were hospitalized vs. COVID-19-positive patients who did not necessitate hospital admission) by fitting a logistic regression model. Prediction accuracy for the PGS was assessed by performing 500 bootstrap resamples of the increase in the pseudo-R-squared. We also divided the sample into deciles and percentiles to assess risk stratification. The models were fit for the dependent variable adjusting for sex, age, the first 10 PCs, and the sampling region (in the Admixed AMR cohort) with and without the PGS, and the partial pseudo-R² was computed and averaged among the resamples.

A clinical severity scale was used in a multinomial regression model to further evaluate the power of this cross-ancestry PGS for risk stratification. These severity strata were defined as follows: 0) asymptomatic; 1) mild, that is, with symptoms, but without pulmonary infiltrates or need of oxygen therapy; 2) moderate, that is, with pulmonary infiltrates affecting <50% of the lungs or need of supplemental oxygen therapy; 3) severe disease, that is, with hospital admission and PaO₂<65 mmHg or SaO₂<90%, PaO₂/FiO₂<300, SaO₂/FiO₂<440, dyspnea, respiratory frequency≥22 bpm, and infiltrates affecting >50% of the lungs; and 4) critical disease, that is, with an admission to the ICU or need of mechanical ventilation (invasive or noninvasive).

Data availability

Summary statistics from the SCOURGE Latin American GWAS and the analysis scripts are available from the public repository <https://github.com/CIBERER/Scourge-COVID19>.

Funding

Instituto de Salud Carlos III (COV20_00622 to A.C., COV20/00792 to M.B., COV20_00181 to C.A., COV20_1144 to M.A.J.S. and A.F.R., PI20/00876 to C.F.); European Union (ERDF) ‘A way of making Europe’. Fundación Amancio Ortega, Banco de Santander (to A.C.), Estrella de Levante S.A. and Colabora Mujer Association (to E.G.-N.) and Obra Social La Caixa (to R.B.); Agencia Estatal de Investigación (RTC-2017-6471-1 to C.F.), Cabildo Insular de Tenerife (CGIEU0000219140 ‘Apuestas

científicas del ITER para colaborar en la lucha contra la COVID-19' to C.F.) and Fundación Canaria Instituto de Investigación Sanitaria de Canarias (PIFIISC20/57 to C.F.). SD-DA was supported by a Xunta de Galicia predoctoral fellowship.

Author contributions

Study design: RC, AC, CF. Data collection: SCOURGE cohort group. Data analysis: SD-DA, RC, ADL, CF, JML-S. Interpretation: SD-DA, RC, ADL. Drafting of the manuscript: SD-DA, RC, ADL, CF, AR-M, AC. Critical revision of the manuscript: SD-DA, RC, ADL, AC, CF, JAR, AR-M, and PL. Approval of the final version of the publication: all coauthors.

Acknowledgements

The contribution of the Centro Nacional de Genotipado (CEGEN) and Centro de Supercomputación de Galicia (CESGA) for funding this project by providing supercomputing infrastructures is also acknowledged. The authors are also particularly grateful for the supply of material and the collaboration of patients, health professionals from participating centers and biobanks. Namely, Biobanc-Mur, and biobanks of the Complejo Hospitalario Universitario de A Coruña, Complejo Hospitalario Universitario de Santiago, Hospital Clínico San Carlos, Hospital La Fe, Hospital Universitario Puerta de Hierro Majadahonda—Instituto de Investigación Sanitaria Puerta de Hierro—Segovia de Arana, Hospital Ramón y Cajal, IDIBGI, IdISBa, IIS Biocruces Bizkaia, IIS Galicia Sur. Also biobanks of the Sistema de Salud de Aragón, Sistema Sanitario Público de Andalucía, and Banco Nacional de ADN.

References

1. The COVID-19 Host Genetics Initiative, Ganna A. (2023) **A second update on mapping the human genetic architecture of COVID-19** *medRxiv* <https://doi.org/10.1101/2022.12.24.22283874>
2. Pairo-Castineira E., et al. (2023) **GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19** *Nature*
3. Niemi M. E. K., et al. (2021) **Mapping the human genetic architecture of COVID-19** *Nature* **600**:472–477
4. Popejoy A. B., Fullerton S. M. (2016) **Genomics is failing on diversity** *Nature* **538**:161–164
5. Sirugo G., Williams S. M., Tishkoff S. A. (2019) **The Missing Diversity in Human Genetic Studies** *Cell* **177**:26–31
6. Li Y. R., Keating B. J. (2014) **Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations** *Genome Med* **6**
7. Rosenberg N. A., et al. (2010) **Genome-wide association studies in diverse populations** *Nat. Rev. Genet* **11**:356–366
8. Kwok A. J., Mentzer A., Knight J. C. (2021) **Host genetics and infectious disease: new tools, insights and translational opportunities** *Nat. Rev. Genet* **22**:137–153
9. Karlsson E. K., Kwiatkowski D. P., Sabeti P. C. (2014) **Natural selection and infectious disease in human populations** *Nat. Rev. Genet* **15**:379–393
10. Namkoong H., et al. (2022) **DOCK2 is involved in the host genetics and biology of severe COVID-19** *Nature* **609**:754–760
11. Bastard P., et al. (2022) **A loss-of-function IFNAR1 allele in Polynesia underlies severe viral diseases in homozygotes** *J. Exp. Med* **219**
12. Duncan C. J. A., et al. (2022) **Life-threatening viral disease in a novel form of autosomal recessive IFNAR2 deficiency in the Arctic** *J. Exp. Med* **219**
13. Peterson R. E., et al. (2019) **Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations** *Cell* **179**:589–603
14. Mester R., et al. (2023) **Impact of cross-ancestry genetic architecture on GWAS in admixed populations** *bioRxiv* <https://doi.org/10.1101/2023.01.20.524946>
15. Atkinson E.G., et al. (2021) **Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power**
16. Cruz R., et al. (2022) **Novel genes and sex differences in COVID-19 severity** *Hum. Mol. Genet* **31**:3789–3806

17. Degenhardt F., et al. (2022) **Detailed stratified GWAS analysis for severe COVID-19 in four European populations** *Hum. Mol. Genet* **31**:3945–3966
18. Kousathanas A., et al. (2022) **Whole-genome sequencing reveals host factors underlying critical COVID-19** *Nature*
19. Kanai M., et al. (2023) **A second update on mapping the human genetic architecture of COVID-19** *Nature* **621**:E7–E26
20. Brunette G.J., et al. (2019) **Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes** | *PNAS*
21. Gioia U., et al. (2023) **SARS-CoV-2 infection induces DNA damage, through CHK1 degradation and impaired 53BP1 recruitment, and cellular senescence** *Nat. Cell Biol* **25**:550–564
22. Im J.-Y., et al. (2020) **DDIAS promotes STAT3 activation by preventing STAT3 recruitment to PTPRM in lung cancer cells** *Oncogenesis* **9**:1–11
23. Im J.-Y., Kang M.-J., Kim B.-K., Won M. (2023) **DDIAS, DNA damage-induced apoptosis suppressor, is a potential therapeutic target in cancer** *Exp. Mol. Med* :1–7 <https://doi.org/10.1038/s12276-023-00974-6>
24. Won K.-J., et al. (2014) **Human Noxin is an anti-apoptotic protein in response to DNA damage of A549 non-small cell lung carcinoma** *International Journal of Cancer - Wiley Online Library*
25. Angeli F., et al. (2023) **The spike effect of acute respiratory syndrome coronavirus 2 and coronavirus disease 2019 vaccines on blood pressure** *Eur. J. Intern. Med* **109**:12–21
26. Silva-Aguiar R. P., et al. (2020) **Role of the renin-angiotensin system in the development of severe COVID-19 in hypertensive patients** *Am. J. Physiol.-Lung Cell. Mol. Physiol* **319**:L596–L602
27. Li Y., et al. (2021) **The emerging role of ISWI chromatin remodeling complexes in cancer** *J. Exp. Clin. Cancer Res* **40**
28. Xia L., et al. (2021) **Inc-BAZ2B promotes M2 macrophage activation and inflammation in children with asthma through stabilizing BAZ2B pre-mRNA** *J. Allergy Clin. Immunol* **147**:921–932
29. Mahnke K, et al. (2000) **The Dendritic Cell Receptor for Endocytosis, Dec-205, Can Recycle and Enhance Antigen Presentation via Major Histocompatibility Complex Class II-Positive Lysosomal Compartments** *Journal of Cell Biology*
30. Sims A. C., et al. (2013) **Release of Severe Acute Respiratory Syndrome Coronavirus Nuclear Import Block Enhances Host Transcription in Human Lung Cells** *J. Virol* **87**:3885–3902
31. Mitchell HD, et al. (2013) **A Network Integration Approach to Predict Conserved Regulators Related to Pathogenicity of Influenza and SARS-CoV Respiratory Viruses** *PLOS ONE*
32. Gómez-Carballa A., et al. (2022) **A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity** *Environ. Res* **210**

33. Policard M., Jain S., Rego S., Dakshanamurthy S. (2021) **Immune characterization and profiles of SARS-CoV-2 infected patients reveals potential host therapeutic targets and SARS-CoV-2 oncogenesis mechanism** *Virus Res* **301**
34. Wei J., et al. (2021) **Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection** *Cell* **184**:76–91
35. Wei J., et al. (2023) **Pharmacological disruption of mSWI/SNF complex activity restricts SARS-CoV-2 infection** *Nat. Genet* **55**:471–483
36. Pereira A. C., et al. (2022) **Genetic risk factors and COVID-19 severity in Brazil: results from BRACOVID study** *Hum. Mol. Genet* **31**:3021–3031
37. Zhu X., et al. (2022) **ZBTB7A promotes virus-host homeostasis during human coronavirus 229E infection** *Cell Rep* **41**
38. Gupta S., et al. (2020) **Emerging role of ZBTB7A as an oncogenic driver and transcriptional repressor** *Cancer Lett* **483**:22–34
39. Yoneyama M., et al. (1998) **Direct triggering of the type I interferon system by virus infection: activation of a transcription factor complex containing IRF-3 and CBP/p300** *EMBO J* **17**:1087–1095
40. Yang Q., et al. (2023) **SARS-CoV-2 infection activates CREB/CBP in cellular cyclic AMP-dependent pathways** *J. Med. Virol* **95**
41. Horowitz J. E., et al. (2022) **Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease** *Nat. Genet* **54**:382–392
42. Harris P. A., et al. (2009) **Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support** *J. Biomed. Inform* **42**:377–381
43. Harris P. A., et al. (2019) **The REDCap consortium: Building an international community of software platform partners** *J. Biomed. Inform* **95**
44. Purcell S., et al. (2007) **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses** *Am. J. Hum. Genet* **81**:559–575
45. Alexander D. H., Novembre J., Lange K. (2009) **Fast model-based estimation of ancestry in unrelated individuals** *Genome Res* **19**:1655–1664
46. Auton A., et al. (2015) **A global reference for human genetic variation** *Nature* **526**:68–74
47. Mao X., et al. (2007) **A Genomewide Admixture Mapping Panel for Hispanic/Latino Populations** *Am. J. Hum. Genet* **80**:1171–1178
48. Wojcik G. L., et al. (2019) **Genetic analyses of diverse populations improves discovery for complex traits** *Nature* **570**:514–518
49. Zheng X., Davis J. W. (2021) **SAIGEgds—an efficient statistical tool for large-scale PheWAS with mixed models** *Bioinformatics* **37**:728–730

50. Zhou W., et al. (2018) **Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies** *Nat. Genet* **50**:1335–1341
51. Willer CJ, et al. (2010) **METAL: fast and efficient meta-analysis of genomewide association scans** *Bioinformatics*
52. McGuire D., et al. (2021) **Model-based assessment of replicability for genome-wide association meta-analysis** *Nat. Commun* **12**
53. Watanabe K., Taskesen E., Bochoven A., Posthuma D. (2017) **Functional mapping and annotation of genetic associations with FUMA** *Nat. Commun* **8**
54. de Leeuw CA, et al. (2015) **MAGMA: Generalized Gene-Set Analysis of GWAS Data** *PLOS Computational Biology*
55. Ghousaini M., et al. (2021) **Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics** *Nucleic Acids Res* **49**:D1311–D1320
56. McLaren W., et al. (2016) **The Ensembl Variant Effect Predictor** *Genome Biol* **17**
57. Barbeira A. N., et al. (2021) **Exploiting the GTEx resources to decipher the mechanisms at GWAS loci** *Genome Biol* **22**
58. Barbeira A. N., et al. (2019) **GWAS and GTEx QTL integration** <https://doi.org/10.5281/zenodo.3518299>
59. Barbeira A. N., et al. (2018) **Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics** *Nat. Commun* **9**
60. Barbeira A. N., et al. (2019) **Integrating predicted transcriptome from multiple tissues improves association detection** *PLOS Genet* **15**
61. Kachuri L., et al. (2023) **Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture** *Nat. Genet* **55**:952–963
62. Bryc K, et al. (2010) **Genome-wide patterns of population structure and admixture among Hispanic/Latino populations** *PNAS*

Editors

Reviewing Editor

Siming Zhao

Dartmouth College, Lebanon, United States of America

Senior Editor

Murim Choi

Seoul National University, Seoul, Korea, the Republic of

Reviewer #1 (Public Review):

Summary:

This paper conducted a GWAS meta-analysis for COVID-19 hospitalization among admixed American populations. The authors identified four genome-wide significant associations, including two novel loci (BAZ2B and DDIAS), and an additional risk locus near CREBBP using cross-ancestry meta-analysis. They utilized multiple strategies to prioritize risk variants and target genes. Finally, they constructed and assessed a polygenic risk score model with 49 variants associated with critical COVID-19 conditions.

Strengths:

Given that most of the previous studies were done in European ancestries, this study provides unique findings about the genetics of COVID-19 in admixed American populations. The GWAS data would be a valuable resource for the community. The authors conducted comprehensive analyses using multiple different strategies, including Bayesian fine mapping, colocalization, TWAS, etc., to prioritize risk variants and target genes. The polygenic risk score (PGS) result demonstrated the ability of cross-population PGS model for COVID-19 risk stratification.

Weaknesses:

- (1) One of the major limitations of this study is that the GWAS sample size is relatively small, which limits its power.
- (2) Lack of replication cohort.
- (3) Colocalization and TWAS used eQTL data from GTEx data, which are mainly from European ancestries.

Comments on latest version:

The authors addressed most of my concerns.

<https://doi.org/10.7554/eLife.93666.2.sa3>

Reviewer #2 (Public Review):

This is a genome-wide association study of COVID-19 in individuals of admixed American ancestry (AMR) recruited from Brazil, Colombia, Ecuador, Mexico, Paraguay and Spain. After quality control and admixture analysis, a total of 3,512 individuals were interrogated for 10,671,028 genetic variants (genotyped + imputed). The genetic association results for these cohorts were meta-analyzed with the results from The Host Genetics Initiative (HGI), involving 3,077 cases and 66,686 controls. The authors found two novel genetic loci associated with COVID-19 at 2q24.2 (rs13003835) and 11q14.1 (rs77599934), and other two independent signals at 3p21.31 (rs35731912) and 6p21.1 (rs2477820) already reported as associated with COVID-19 in previous GWASs. Additional meta-analysis with other HGI studies also suggested risk variants near CREBBP, ZBTB7A and CASC20 genes.

Strengths:

These findings rely on state-of-the-art methods in the field of Statistical Genomics and help to address the issue of low number of GWASs in non-European populations, ultimately contributing to reduce health inequalities across the globe.

Weaknesses:

There is no replication cohort, as acknowledged by the authors (page 29, line 587) and no experimental validation to assess the biological effect of putative causal variants/genes. Thus, the study provides good evidence of association, rather than causation, between the genetic variants and COVID-19.

Comments on latest version:

The issues identified in the first round of review were well addressed by the authors in the revised version of the manuscript.

<https://doi.org/10.7554/eLife.93666.2.sa2>

Reviewer #3 (Public Review):

Summary:

In the context of the SCOURGE consortium's research, the authors conduct a GWAS meta-analysis on 4,702 hospitalized individuals of admixed American descent suffering from COVID-19. This study identified four significant genetic associations, including two loci initially discovered in Latin American cohorts. Furthermore, a trans-ethnic meta-analysis highlighted an additional novel risk locus in the CREBBP gene, underscoring the critical role of genetic diversity in understanding the pathogenesis of COVID-19.

Strengths:

- (1) The study identified two novel severe COVID-19 loci (BAZ2B and DDIAS) by the largest GWAS meta-analysis for COVID-19 hospitalization in admixed Americans.
- (2) With a trans-ethnic meta-analysis, an additional risk locus near CREBBP was identified.

Weaknesses:

- (1) The GWAS power is limited due to the relatively small number of cases.
- (2) There is no replication study for the novel severe COVID-19 loci, which may lead to false positive findings.
- (3) The variants selected for the PGS appear arbitrary and may not leverage the GWAS findings.
- (4) The TWAS models were predominantly trained on European samples, and there is no replication study for the findings as well.

<https://doi.org/10.7554/eLife.93666.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public Review):

Summary:

This paper conducted a GWAS meta-analysis for COVID-19 hospitalization among admixed American populations. The authors identified four genome-wide significant associations, including two novel loci (BAZ2B and DDIAS), and an additional risk locus near CREBBP using cross-ancestry meta-analysis. They utilized multiple strategies to prioritize risk variants and target genes. Finally, they constructed and assessed a polygenic risk score model with 49 variants associated with critical COVID-19 conditions.

Strengths:

Given that most of the previous studies were done in European ancestries, this study provides unique findings about the genetics of COVID-19 in admixed American populations. The GWAS data would be a valuable resource for the community. The authors conducted comprehensive analyses using multiple different strategies, including Bayesian fine mapping, colocalization, TWAS, etc., to prioritize risk variants and target genes. The polygenic risk score (PGS) result demonstrated the ability of the cross-population

PGS model for COVID-19 risk stratification.

Thank you very much for the positive comments and the willingness to revise this manuscript.

Weaknesses:

(1) One of the major limitations of this study is that the GWAS sample size is relatively small, which limits its power.

(2) The fine mapping section is unclear and there is a lack of information. The authors assumed one causal signal per locus, and only provided credible sets, but did not provide posterior inclusion probabilities (PIP) for the variants to be causal.

(3) Colocalization and TWAS used eQTL data from GTEx data, which are mainly from European ancestries. It is unclear how much impact the ancestry mismatch would have on the result. The readers should be cautious when interpreting the results and designing follow-up studies.

We agree with that the sample size is relatively small. Despite that, it was sufficient to reveal novel risk loci supporting the robustness of the main findings. We have indicated this limitation at the end of the discussion section.

Thank you for rising this point. As suggested, we have also used SuSIE, which allows to assume more than one causal signal per locus. However, in this case the results were not different from those obtained with the original Bayesian colocalization performed with corrcoverage. Regarding the PIP, at the fine mapping stage we are inclined to put more weight on the functional annotations of the variants in the credible set than on the statistical contributions to the signal. This is the reason why we prefer not to put weight on the PIP of the variants but prioritize variants that were enriched functional annotations.

This is a good point regarding the lack of diversity in GTEx data. We have also used data from AMR populations (GALA II-SAGE models), although it was only available for blood tissue. Regarding the ancestry mismatch between datasets, several studies have attempted to explore the impact. Gay et al. (PMID: 32912333) studied local ancestry effects on eQTLs from the GTEx consortium and concluded that adjustment of eQTLs by local ancestry only yields modest improvement over using global ancestry (as done in GTEx). Moreover, the colocalization results between adjusting by Local Ancestry and Global Ancestry were not significantly different. Besides, Mogil et al. (PMID: 30096133) observed that genes with higher heritability share genetic architecture between populations. Nevertheless, both studies have evidenced decreased power and poorer predictive performances regarding gene expression because of reduced diversity in eQTL analyses. As consequence of the ancestry mismatch, we now warn the readers that this may compromise signal detection (Discussion, lines 531-533).

Reviewer #2 (Public Review):

This is a genome-wide association study of COVID-19 in individuals of admixed American ancestry (AMR) recruited from Brazil, Colombia, Ecuador, Mexico, Paraguay, and Spain. After quality control and admixture analysis, a total of 3,512 individuals were interrogated for 10,671,028 genetic variants (genotyped + imputed). The genetic association results for these cohorts were meta-analyzed with the results from The Host Genetics Initiative (HGI), involving 3,077 cases and 66,686 controls. The authors found two novel genetic loci associated with COVID-19 at 2q24.2 (rs13003835) and 11q14.1 (rs77599934), and other two independent signals at 3p21.31 (rs35731912) and 6p21.1 (rs2477820) already reported as associated with COVID-19 in previous GWASs. Additional meta-analysis with other HGI studies also suggested risk variants near CREBBP, ZBTB7A, and CASC20 genes.

Strengths:

These findings rely on state-of-the-art methods in the field of Statistical Genomics and help to address the issue of a low number of GWASs in non-European populations, ultimately contributing to reducing health inequalities across the globe.

Thank you very much for the positive comments and the willingness to revise this manuscript.

Weaknesses:

There is no replication cohort, as acknowledged by the authors (page 29, line 587), and no experimental validation to assess the biological effect of putative causal variants/genes. Thus, the study provides good evidence of association, rather than causation, between the genetic variants and COVID-19. Lastly, I consider it crucial to report the results for the SCOURGE Latin American GWAS, in addition to its meta-analysis with HGI results, since HGI data has a different phenotype scheme (Hospitalized COVID vs Population) compared to SCOURGE (Hospitalized COVID vs Non-hospitalized COVID).

We essentially agree with the reviewer in that one of the main limitations of the study is the lack of a replication stage because of the use of all available datasets on a one-stage analysis. To contribute to the interpretation of the findings in the absence of a replication stage, we now assessed the replicability of the novel loci using the Meta-Analysis Model-based Assessment of replicability (MAMBA) approach (PMID: 33785739) and included the posterior probabilities of replication in Table 2. We also explored further the potential replicability of signals in other populations. We agree that the results should be interpreted in terms of associations given the lack of functional validation of main findings, so we have slightly modified the discussion.

As suggested, the SCOURGE Latin American GWAS summary is now accessible by direct request to the Consortium GitHub repository (<https://github.com/CIBERER/Scourge-COVID19>) (lines 797-799). We have also included the results from the SCOURGE GWAS analysis for the replication of the 40 lead variants in the Supplementary Table 12. Results from the SCOURGE GWAS for the lead variants in the AMR meta-analysis with HGI were already included in the Supplementary Table 2. As note, we have not been able to conduct the meta-analysis with the same hospitalization scheme as in the HGI study since the population-specific results for those analyses were not publicly released. However, sensitivity analyses included within the supplementary material from the COVID-19 Host Genetics Initiative (2021) stated that there were no significant differences in effects (Odds Ratios) between analyses using population controls or just non-hospitalized COVID-19 patients.

Reviewer #3 (Public Review):

Summary:

In the context of the SCOURGE consortium's research, the authors conduct a GWAS meta-analysis on 4,702 hospitalized individuals of admixed American descent suffering from COVID-19. This study identified four significant genetic associations, including two loci initially discovered in Latin American cohorts. Furthermore, a trans-ethnic meta-analysis highlighted an additional novel risk locus in the CREBBP gene, underscoring the critical role of genetic diversity in understanding the pathogenesis of COVID-19.

Strengths:

(1) The study identified two novel severe COVID-19 loci (BAZ2B and DDIAS) by the largest GWAS meta-analysis for COVID-19 hospitalization in admixed Americans.

(2) With a trans-ethnic meta-analysis, an additional risk locus near CREBBP was identified.

Thank you very much for the positive comments and the willingness to revise this manuscript.

Weaknesses:

(1) The GWAS power is limited due to the relatively small number of cases.

(2) There is no replication study for the novel severe COVID-19 loci, which may lead to false positive findings.

We agree with that the sample size is relatively small. Despite that, it was sufficient to reveal novel risk loci supporting the robustness of the main findings. We have indicated this limitation at the end of the discussion section.

Regarding the lack of a replication study, we now assessed the replicability of the novel loci using the Meta-Analysis Model-based Assessment of replicability (MAMBA) approach (PMID: 33785739). We have included the posterior probabilities of replication in Table 2.

(3) Significant differences exist in the ages between cases and controls, which could potentially introduce biased confounders. I'm curious about how the authors treated age as a covariate. For instance, did they use ten-year intervals? This needs clarification for reproducibility.

Thank you for rising this point. Age was included as a continuous variable. This has been now indicated in line 667 (within Material and Methods).

(4)"Those in the top PGS decile exhibited a 5.90-fold (95% CI=3.29-10.60, $p=2.79 \times 10^{-9}$) greater risk compared to individuals in the lowest decile". I would recommend comparing with the 40-60% PGS decile rather than the lowest decile, as the lowest PGS decile does not represent 'normal controls'.

Thank you. In the revised version, the PGS categories was compared following the recommendation (lines 461-463).

(5) In the field of PGS, it's common to require an independent dataset for training and testing the PGS model. Here, there seems to be an overfitting issue due to using the same subjects for both training and testing the variants.

We are sorry for the misunderstanding. In fact, we have followed the standard to avoid overfitting of the PGS model and have used different training and testing datasets. The training data (GWAS) was the HGI-B2 ALL meta-analysis, in which our AMR GWAS was not included. The PRS model was then tested in the SCOURGE AMR cohort. However, it is true that we did test the combination of the PRS adding the new discovered variants in the SCOURGE cohort. To avoid potential overfitting by adding the new loci, we have excluded from the manuscript the results on which we included the newly discovered variants.

(6) The variants selected for the PGS appear arbitrary and may not leverage the GWAS findings without an independent training dataset.

Again, we are sorry for the misunderstanding. The PGS model was built with 43 variants associated with hospitalization or severity within the HGI v7 results and 7 which were discovered by the GenOMICC consortium in their latest study and were not in the latest HGI release. The variants are included within the Supplementary Table 14, but we have now annotated the discovery GWAS.

(7) The TWAS models were predominantly trained on European samples, and there is no replication study for the findings as well.

This is a good point regarding the lack of diversity in GTEx data. We have also used data from AMR populations (GALA II-SAGE models), although it was only available for blood tissue. Regarding the ancestry mismatch between datasets, several studies have attempted to explore the impact. Gay et al. (PMID: 32912333) studied local ancestry effects on eQTLs from the GTEx consortium and concluded that adjustment of eQTLs by local ancestry only yields modest improvement over using global ancestry (as done in GTEx). Moreover, the colocalization results between adjusting by Local Ancestry and Global Ancestry were not significantly different. Besides, Mogil et al. (PMID: 30096133) observed that genes with higher heritability share genetic architecture between populations. Nevertheless, both studies have evidenced decreased power and poorer predictive performances regarding gene expression because of reduced diversity in eQTL analyses. As consequence of the ancestry mismatch, we now warn the readers that this may compromise signal detection (Discussion, lines 531-533).

Recommendations for the authors:

Reviewer #1 (Recommendations For The Authors):

(1) The authors mentioned the fine mapping method did not converge for the locus in chr 11. I would consider trying a different fine-mapping method (such as SuSiE or FINEMAP). It would be helpful to provide posterior inclusion probabilities (PIP) for the variants in fine mapping results and plot the PIP values in the regional association plots.

As suggested, we have also used SuSiE, which allows to assume more than one causal signal per locus. However, in this case the results were not different from those obtained with the original Bayesian colocalization performed with corrcoverage. SuSiE's fine-mapping for chromosome 11 prioritized a single variant, which is likely due to the rare frequency. Thus, we have maintained the fine-mapping as it was originally indicated in the previous version of the manuscript but have now included the credible set in Supplementary Table 6.

Regarding the PIP, at the fine mapping stage we are inclined to put more weight on the functional annotations of the variants in the credible set than on the statistical contributions to the signal. This is the reason why we prefer not to put weight on the PIP of the variants but prioritize variants that were enriched functional annotations.

(2) Please provide more detailed information about the VEP and V2G analysis and how to interpret those results. My understanding of V2G is that it includes different sources of information (such as molecular QTLs and chromatin interactions from different tissues/cell types, etc.). It is unclear what sources of information and weight settings were used in the V2G model.

Thank you for rising this point. As suggested, we have clarified the basis for VEP and V2G and the interpretation (lines 732-743).

(3) The authors identified multiple genes with different strategies, e.g. FUMA, V2G, COLOC, TWAS, etc. How many genes were found/supported by evidence provided by multiple methods? It could be helpful to have a table summarizing the risk genes found by different strategies, and the evidence supporting the genes. e.g. which genes are found by which methods, and the biological functions of the genes, etc.

Thank you for rising this point. As suggested, we now added a new figure (Figure 5) to summarize the findings with the multiple methods used.

(4) It would be helpful to make the code/scripts available for reproducibility.

As suggested, the SCOURGE Latin American GWAS summary and the analysis scripts (<https://github.com/CIBERER/Scourage-COVID19/tree/main/scripts/novel-risk-hosp-AMR-2024>) are now accessible in the Consortium GitHub repository (<https://github.com/CIBERER/Scourage-COVID19>) (lines 806-807).

(5) The fonts in some of the figures (e.g. Figure 2) are hard to read.

Thank you. We have now included the figures as SVG files.

Reviewer #2 (Recommendations For The Authors):

- The abstract lacks a conclusion sentence.

Thank you. As suggested, we have included two additional sentences with broad conclusions from the study. We preferred to avoid relying on conclusions related to known or new biological links of the prioritized genes given the lack of functional validation of main findings.

- Regarding the association analysis (page 27, line 677), I wonder if some of the 10 principal components (PCs) are capturing information about the recruitment areas (countries). It may be relevant to test for multicollinearity among these variables.

Since we acknowledge that some of the categories might be correlated with a certain PC but not all of them do, we have calculated GVIF values for the main variables to assess the categorical variable as a single entity. The scaled $GVIF^{1/(1/2 \cdot Df)}$ value for the categorical variable is 1.52. Thus, if we square this value, we obtain 2.31, which can be then used for applying usual rule-of-thumb for VIF values.

- Still on the topic of association analysis, did the authors adjust the logistic model for comorbidities variables from Table 1? Given these comorbidities also have a genetic component and their distribution differs between non-hospitalized vs hospitalized, I am concerned that comorbidities might be confounding the association between genetic variants and COVID.

We did not adjust by comorbidities since HGI studies were not adjusted either and we aimed to be as aligned as possible with HGI. However, as suggested, we have now tested the association between each of the comorbidities in Table 1 and each of the variants in Table 2, using the comorbidities as dependent variables and adjusting for the main covariables (age, sex, PCs and country of recruitment). None of the variants were significantly associated to the comorbidities (line 333).

- If I understood correctly, the 49 genetic variants used to develop the polygenic risk score model (PRS) were based on the HGI total sample size (data release 7), which is predominantly of European ancestry. I am concerned about the prediction accuracy in the AMR population (PRS transferability issue).

We have explored literature in search of other PRS to compare the associated OR in our cohort with ORs calculated in European populations. Horowitz et al. (2022) reported an OR of 1.38 for the top 10% with respect to hospitalization risk in European individuals using a GRS with 12 variants.

We acknowledge that this might be an issue and is now explained in discussion of the revised version (lines 561-568). However, as this is the first time a PRS for COVID-19 is applied to a relatively large AMR cohort, we believe that this analysis will be of value for further analyses regarding PRS transferability, providing a source for comparison in further studies.

- On page 23, line 579, the authors acknowledge their "GWAS is underpowered". This sentence requires a sample/power calculation, otherwise, I suggest using "is likely underpowered".

Thanks for the input. We have modified the sentence as suggested.

Reviewer #3 (Recommendations For The Authors):

I wonder if the authors have an approximate date when the GWAS summary statistic will be available. I reviewed some manuscripts in the past, and the authors claimed they would deposit the data soon, but in fact it would not happen until 2 years later.

The summary statistics are already available from the SCOURGE Consortium repository <https://github.com/CIBERER/Scourge-COVID19> (lines 806-807).

<https://doi.org/10.7554/eLife.93666.2.sa0>