Alternative splicing of *OAS1* alters the risk for severe COVID-19

Authors

Jennifer Huffman[1], Guillaume Butler-Laporte[2], Atlas Khan[3], Theodore G. Drivas[4,5], Gina M. Peloso[1,6], Tomoko Nakanishi[7-11], Anurag Verma[4], Krzysztof Kiryluk[3,12], J. Brent Richards[2,13] and Hugo Zeberg[14,15*]

Affiliations

[1] Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Bostoon, MA, USA, 02130

[2] Departments of Medicine, Human Genetics, Epidemiology, Biostatistics and Occupational Health, McGill University, Lady Davis Institute, Jewish General Hospital, Montréal, Québec, Canada

[3] Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY

[4] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

[5] Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia

[6] Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118

[7] Institute for Molecular Medicine Finland, Univerisity of Helsinki, Helsinki, Finland.

[8] Department of Human Genetics, McGill University, Montréal, Québec, Canada.

[9] Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada.

[10] Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan.

[11] Research Fellow, Japan Society for the Promotion of Science, Tokyo, Japan.

[12] Institute for Genomic Medicine, Columbia University, New York, NY

[13] Department of Twin Research, King's College London, London, UK

[14] Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany.

[15] Department of Neuroscience, Karolinska Institutet, SE-17177 Stockholm, Sweden.

***Corresponding author:** hugo.zeberg@ki.se

ABSTRACT

A locus containing *OAS1/2/3* has been identified as a risk locus for severe COVID-19 among Europeans ancestry individuals, with a protective haplotype of ~75 kilobases derived from Neanderthals. Here, we show that among several potentially causal variants at this locus, a splice variant of *OAS1* occurs in people of African ancestry independently of the Neanderthal haplotype and confers protection against COVID-19 of a magnitude similar to that seen in individuals without African ancestry.

MAIN TEXT

The COVID-19 pandemic has haunted the world for over a year. During this period, several large international efforts[1–4] have been launched to identify the genetic determinants of COVID-19 susceptibility and severity. These efforts have identified more than a dozen genomic regions associated with severe COVID-19. However, the causal variants in these regions are yet to be identified, hampering our ability to understand COVID-19 pathophysiology.

When risk haplotypes are long it is more challenging to disentangle causal variants due to linkage disequilibrium (LD). This is especially problematic for haplotypes derived from Neanderthals and Denisovans that often span several tens of kilobases or more. Two notable COVID-19 examples are the major risk locus on chromosome 3 (3p21.31) and the *OAS1/2/3* locus on chromosome 12 (12q24.13), which both carry haplotypes of Neanderthal origin[5,6]. The *OAS* genes encode enzymes catalyzing the synthesis of short polyadenylates, which activate ribonuclease L that in turn degrades intracellular double-stranded RNA and triggers several other antiviral mechanisms[7]. The protective Neanderthal-derived haplotype confers ~23% reduced risk of becoming critically ill upon infection with SARS-CoV-2[3]. Supporting this, a recent Mendelian randomization study found that increased circulating levels of OAS1 were associated with reduced risk of very severe COVID-19, hospitalization for COVID-19 and susceptibility to this disease[8]. However, other evidence from a transcription-wide association study, suggested a stronger association with OAS3 levels[3]. Thus, efforts are required to disentangle the causal gene, or genes, at this locus.

The *OAS* region was identified as a COVID-19 risk locus in association studies[1,3] of mainly Europeans. The protective haplotype is derived from Neanderthals, is ~75 kilobases long and covers the three genes *OAS1/2/3*[6]. A candidate causal variant in the region is rs10774671, which falls in a splice acceptor site of *OAS1* and where the protective (G) allele results in a longer and more active OAS1 enzyme[9]. However, this variant is as associated with COVID-19 severity as any of the hundreds of variants in LD. In European ancestry individuals we find 130 variants co-segregating

(r^2>0.8) with the splice-acceptor variant (**Fig. 1a**). Thus, further methods are required to disentangle the causal SNP(s) at this locus. Doing so, could help to identify the causal gene.

One method to better identify causal SNPs at an associated locus is to test associations in different ancestries, particularly when these other populations have different LD structure and shorter haplotypes. Therefore, to examine if we could identify a population with which we could test this variant independently, we investigated the presence of co-segregating variants in populations in the 1000 Genomes project[10]. In South Asians, there are 129 such variants and in East Asian 128 variants. In stark contrast, no variants co-segregate with rs10774671 in Africans at a LD of r^2>0.6 (**Fig. 1b**). Thus, the African ancestry population offers a possibility to independently test if rs10774671 is associated with COVID-19 severity.

To test the association of splice acceptor variant rs10774671 in people of African ancestry with COVID-19 outcomes we combined five studies that had assessed COVID-19 severity (UK BioBank, Penn Medicine BioBank, Columbia University COVID-19 Biobank, Biobanque Québec COVID-19 and the VA Million Veteran Program), comprising 1,842 cases and 118,631 controls. We found that the rs10774671-G allele conferred a protection against COVID-19 hospitalization in this population (**Fig. 2**, p = 0.03) of similar magnitude (OR = 0.92, 95% confidence interval [CI]: 0.86-0.99) as in Europeans (OR = 0.89, 95% CI: 0.86-0.93), a population in which the rs10774671-G allele is less common (35% allele frequency versus 66% among African ancestry individuals[10]). Moreover, we find no evidence of heterogeneity accross the five studies (Cochran's Q = 2.00, p = 0.74; I^2 = 0.0% [0.0%-58.4%]; τ^2 = 0.00 [0.00-0.09], 95% CI in brackets, see Methods). Thus, rs10774671 is associated with COVID-19 severity independently of the variants with which it is associated in non-African populations.

This observation is compatible with the fact that Neanderthal haplotypes are rare or absent in African populations[11,12] and that ancestral alleles seen in Neanderthals, such as the G allele at rs10774671, can also exist today as a result of inheritance from the ancestral population common to both modern humans and Neanderthals. In the latter case, variants have existed in modern humans in the order of half a million years ago and therefore co-segregate with other variants than when they are derived from gene flow from Neanderthals into modern humans that occurred about 60,000 years ago[13]. Here, we leverage this fact to show that the ancestral splice variant, encoding a longer and more active enzyme[9], is responsible for the protective effect associated with this locus[14].

These findings provide evidence that the splice-site variant at this locus influences COVID-19 outcomes by manipulating splicing of *OAS1*. Further, this rapid insight highlights the importance of including populations of different ancestries in genetic association studies and rapidly sharing data through large, international consortia.

ACKNOWLEDGMENTS

We thank the COVID-19 Host Genetics Initiative and Regeneron for making the data from the genome-wide association study available. Genotyping and phenotyping of the Columbia University cohort was made possible by the Columbia University COVID-19 Biobank Genomics Workgroup members, including Richard Mayeux, Muredach P. Reilly, Wendy Chung, David B. Goldstein, Christine K. Garcia, Iuliana Ionita-Laza, Andrea Califano, Sheila M. O'Byrne, Danielle Pendrick, Soumitra Sengupta, Peter Sims, and Anne-Catrin Uhlemann. The Columbia University COVID-19 Biobank was supported by Columbia University and the National Center for Advancing Translational Sciences, NIH, through Grant Number UL1TR001873. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The Richards research group is supported by the Canadian Institutes of Health Research (CIHR) (365825 and 409511), the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation (CFI), the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada, the McGill Interdisciplinary Initiative in Infection and Immunity and the Fonds de Recherche Québec Santé (FRQS). GBL is supported by a CIHR scholarship and a joint FRQS and Québec Ministry of Health and Social Services scholarship. TN is supported by a research fellowship of the Japan Society for the Promotion of Science for Young Scientists. JBR is supported by an FRQS Clinical Research Scholarship. Support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded by the Welcome Trust, the Medical Research Council, the European Union, the National Institute for Health Research-funded BioResource and the Clinical Research Facility and Biomedical Research Centre based at Guy's and St. Thomas' NHS Foundation Trust in partnership with King's College London. The Biobanque Québec COVID19 is funded by FRQS, Genome Québec and the Public Health Agency of Canada, the McGill Interdisciplinary Initiative in Infection and Immunity and the Fonds de Recherche Québec Santé. HZ is supported by Jeanssons Stiftelser and Magnus Bergsvalls Stiftelse. We thank Svante Pääbo for careful reading of the manuscript and helpful comments. These funding agencies had no role in the design, implementation or interpretation of this study.

METHODS

Study participants

Our analysis pooled hospitalized COVID-19 patients of African ancestry (n = 1,842) from five cohorts. The UK BioBank cohort contained 74 cases and 9,305 controls, the Penn Medicine BioBank 114 cases and 8,901 controls, the Columbia University Biobank 304 cases and 2,246 controls, Québec Covid-19 50 cases and 50 controls, and VA Million Veteran Program 1,300 cases and 98,129 controls. All participants gave appropriate consent and ethical approvals were obtained from the relevant research ethics boards.

Summary statistics - VA *Million Veteran Program*

The VA Million Veteran Program (MVP) is a US-based longitudinal research program investigating how genes, lifestyle, and military exposures influences health and illness in Veterans, with study recruitment commencing in 2011[15]. Study participants were genotyped using a customized Affymetrix Axiom biobank array (the MVP 1.0 Genotyping Array), containing 723,305 variants[16]. Imputation was performed to a hybrid imputation panel comprised of the African Genome Resources panel (https://imputation.sanger.ac.uk/?about=1#referencepanels) and 1000G v3p5. COVID-19 cases were identified using an algorithm developed by the VA COVID National Surveillance Tool (NST)[17]. COVID-19-related hospitalizations were defined as hospital admissions between 7 days before and 30 days after an individual's positive SARS-CoV-2 test. The association of hospitalized COVID-19 cases versus all other MVP participants was tested under an additive logistic model and was corrected for age, age^2, sex, age-by-sex, and ethnicity-specific PCs. Individuals who died before March 1, 2020 were excluded as was one individual from each related pair. The analysis was restricted to only African American MVP participants (as defined by HARE[18]) resulting in 1,300 cases and 98,129 controls.

Summary statistics - Biobanque Québécoise de la COVID-19

The Biobanque Québécoise de la COVID-19 (BQC-19) is a prospective hospital-based biobank recruiting patients with proven or suspected COVID-19 (Jewish General Hospital research ethics board no. 2020-2137). Whole genome genotyping was performed for all participants, with imputation using the TOPMed server. Individuals with African ancestry were determined by projecting genetic principal components on the 1000G reference panel. Our 50 cases were defined as patients hospitalized with COVID-19 or who died from the infection. Controls were the 50 other African ancestry participants, of which 32 had a clinical presentation consistent with COVID-19, but never had a positive test. An additive logistic regression model with the first 10 genetic principal

components, age, sex, age^2, age-by-sex, and age^2-by-sex as covariates was used to determine the effect of the protective G allele on the risk of being a case.

Summary statistics - Penn Medicine Biobank

The Penn Medicine Biobank has ~60,000 participants consented for genetic studies, which are followed for those SARS-CoV-2 infection and hospitalization. The cohort is prospectively consenting and obtaining DNA from COVID-19 inpatients from Penn Medicine hospitals and controls who were tested by nasal swab and confirmed negative with consenting for DNA banking and genetic studies. Study participants were genotyped using the Illumina Global Screening Array (Version 2) and further imputed using TOPMed. SNPs with a call rate of 99% were included and with a minor allele frequency of at least 1%. The association was corrected for age, age^2, sex, age-by-sex, and ethnicity-specific PCs. The collection of samples to the Penn Medicine Biobank was approved by the Institutional Review Board of Perelman School of Medicine at University of Pennsylvania.

Summary statistics - Columbia University Biobank

The Columbia University COVID-19 Biobank was established in response to the New York City infection surge in March 2020. The biobank recruited COVID-19 cases of diverse ancestry among all patients who were treated at Columbia University Irving Medical Center between March and May 2020. All cases were diagnosed by positive SARS-CoV-2 PCR test based on nasopharyngeal samples. The mean age of cases was 62.89 years, and the percentage of females was 43%. DNA of whole blood samples was extracted using standard procedures and genotyping was performed using the Illumina Global Diversity Array (GDA) chip. The controls were genotyped using the Illumina Multiethnic Global Ancestry (MEGA) chip. The analysis of intensity clusters and genotype calls were performed in Illumina Genome Studio software; all SNPs were called on forward DNA strand and standard quality control (QC) filters were applied, including per-SNP genotyping rate > 95%, per-individual genotyping rate > 90%, minor allele frequency (MAF) > 0.01, and Hardy–Weinberg equilibrium (HWE) test p-value > 10-08 in controls. The duplicates and cryptic relatedness in the given cohort were determined and excluded based on the estimated pairwise kinship coefficients > 0.0884. After QC, the dataset consisted of 6,757 individuals (1,029 cases and 5,728 controls) genotyped for 1,096,321 SNPs with overall genotyping rate of 99.9%. The imputation analysis was performed using TopMed imputation server. A total of 13,439,413 common markers imputed at high quality (R2 > 0.8 and MAF > 0.01) were used in downstream analyses. To define the African ancestry cluster, we used PCA against 1000 Genomes reference populations followed by k-means clustering on significant PCs of ancestry. The African ancestry cluster contained 332 cases positive

for SARS-CoV-2 and 2,246 population controls. Of the 332 African ancestry cases, 304 had severe COVID-19 requiring hospitalization. Among the 304 cases included, 78 (26%) had respiratory failure requiring intubation and invasive ventilatory support, and 86 (28%) died due to COVID-19. We then tested the effect of rs10774671-G on the risk of hospitalization using SAIGE (Scalable and Accurate Implementation of GEneralized mixed model), after adjustment for sex and five principal components of ancestry. The collection of samples to the Columbia University COVID-19 Biobank was approved by the Institutional Review Board (IRB) of Columbia University (IRB protocol number AAAS7370), while the genetic analyses were approved under Columbia University IRB protocol number AAAS7948.

Summary statistics - UK BioBank

Association analyses were performed using the Firth logistic regression test implemented in `REGENIE`, including as covariates age, age^2, sex, age-by-sex, age^2-by-sex, and ten ancestry-informative principle components. Association analyses were performed using the Firth logistic regression test implemented in REGENIE, including as covariates age, age2, sex, age-by-sex, age2-by-sex, and ten ancestry-informative principle components. The data was downloaded at https://rgc-covid19.regeneron.com/results [2021-02-02].

Meta-analysis

The meta-analysis was done using inverse-variance weighting in the R-package `meta`. Heterogeneity was measured using Cochran's Q, Higgin's & Thompson's I^2, and τ^2 using the DerSimonian-Laird estimator.

Linkage disequilibrium

Linkage disequilibrium was calculated using LDlink[19] 4.1 in the genomic region 113.30-113.45 Mb (*hg19*) using data from the 1000 Genomes Project[10].

REFERENCES

1. COVID-19 Host Genetics Intitative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet* **28**, 715–718 (2020).
2. Ellinghaus, D. *et al.* Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med* (2020) doi:10.1056/NEJMoa2020283.
3. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in Covid-19. *Nature* (2020) doi:10.1038/s41586-020-03065-y.
4. Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, (2020).
5. Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).
6. Zeberg, H. & Pääbo, S. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proc Natl Acad Sci U S A* **118**, (2021).
7. Choi, U. Y., Kang, J.-S., Hwang, Y. S. & Kim, Y.-J. Oligoadenylate synthase-like (OASL) proteins: dual functions and associations with diseases. *Exp Mol Med* **47**, e144 (2015).
8. Zhou, S. *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat Med* (2021) doi:10.1038/s41591-021-01281-1.
9. Bonnevie-Nielsen, V. *et al.* Variation in antiviral 2’,5’-oligoadenylate synthetase (2’5’AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene. *Am J Hum Genet* **76**, 623–633 (2005).
10. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
11. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
12. Chen, L., Wolf, A. B., Fu, W., Li, L. & Akey, J. M. Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* **180**, 677-687.e16 (2020).
13. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The date of interbreeding between Neandertals and modern humans. *PLoS Genet* **8**, e1002947 (2012).
14. Sams, A. J. *et al.* Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol* **17**, 246 (2016).

15. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214–223 (2016).

16. Hunter-Zinck, H. *et al.* Genotyping Array Design and Data Quality Control in the Million Veteran Program. *Am J Hum Genet* **106**, 535–548 (2020).

17. Chapman, A. *et al.* A Natural Language Processing System for National COVID-19 Surveillance in the US Department of Veterans Affairs. in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020* (Association for Computational Linguistics, 2020).

18. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet* **105**, 763–772 (2019).

19. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

FIGURES

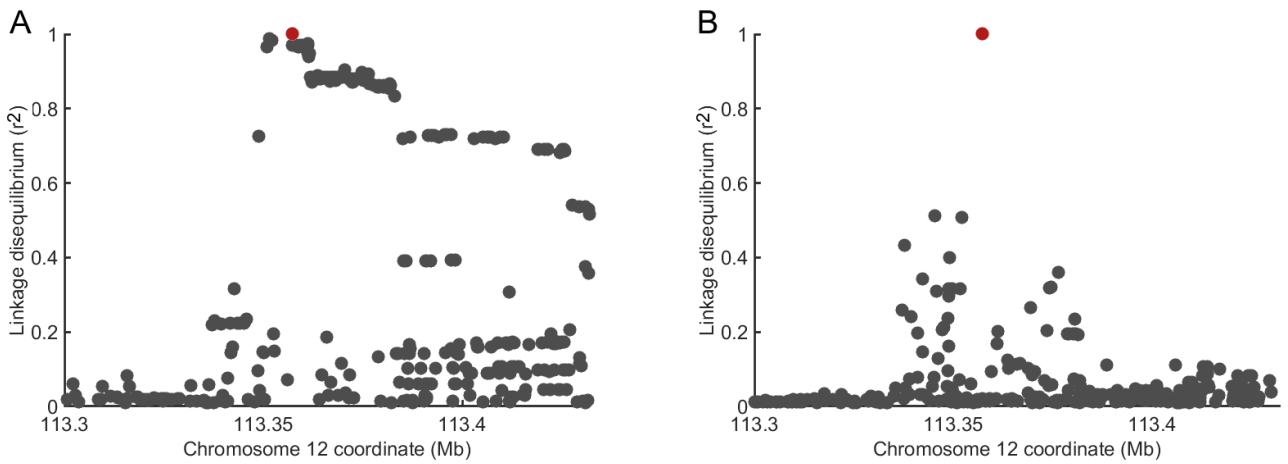


Fig 1. Linkage disequilibrium of the splice acceptor variant in individuals of European and African ancestries. A) Linkage decay in individuals European ancestry (n = 503). 130 variants are in LD ($r^2 > 0.8$) with the splice acceptor variant rs10774671 (marked in red). **B**) Same as in A) but for African ancestry (n = 661). No variants were found to be in LD with the splice acceptor variant. Data from the 1000 Genomes project[10]. *X*-axis gives *hg19* coordinates.

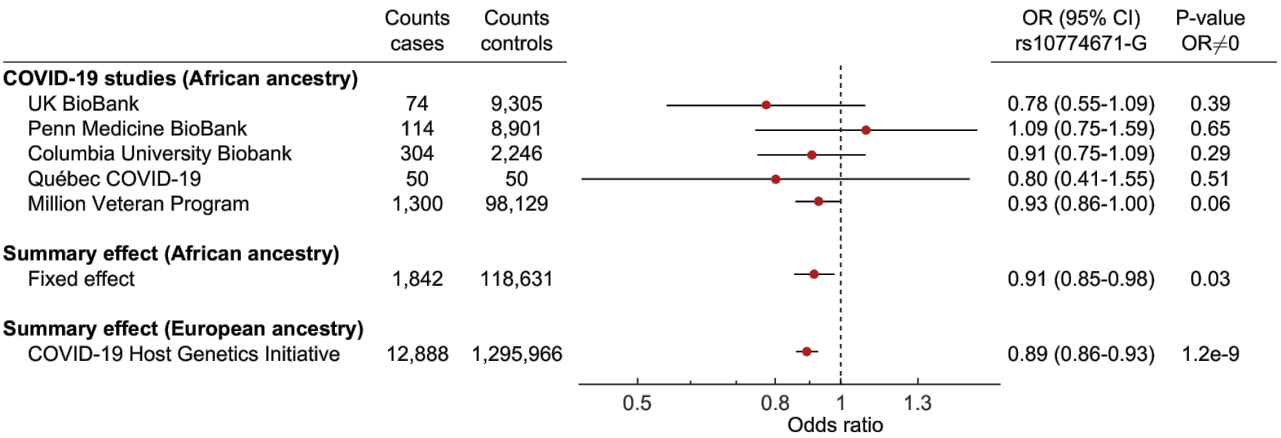
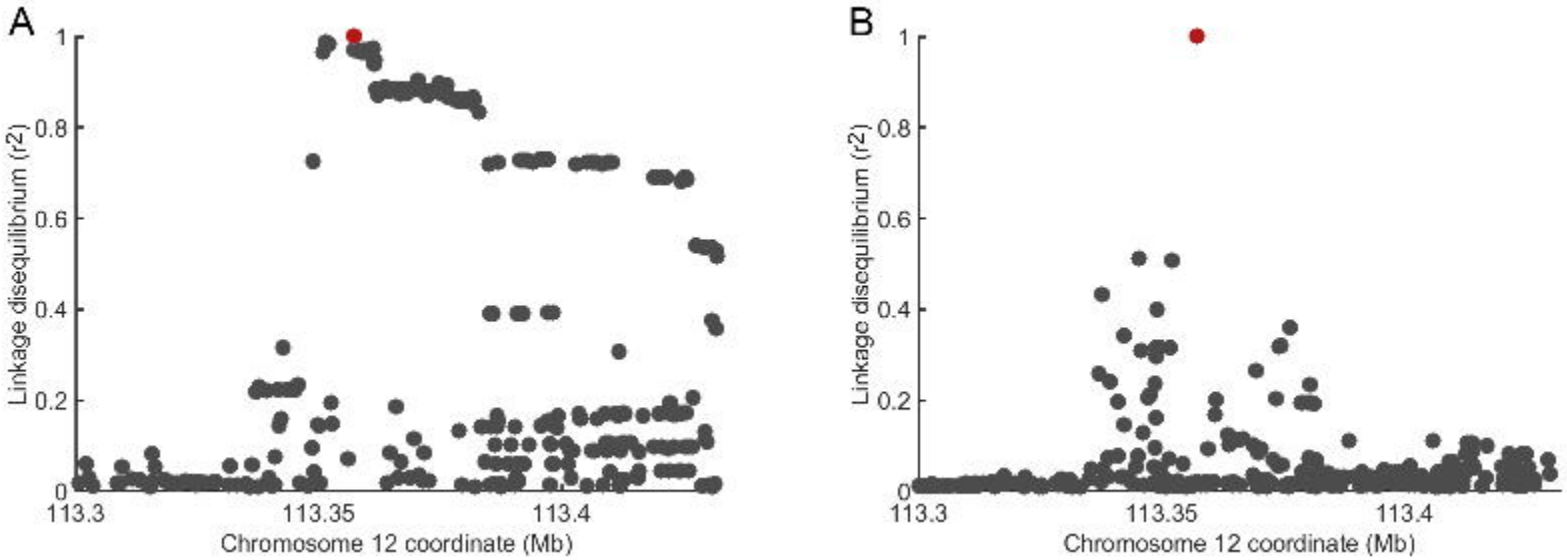
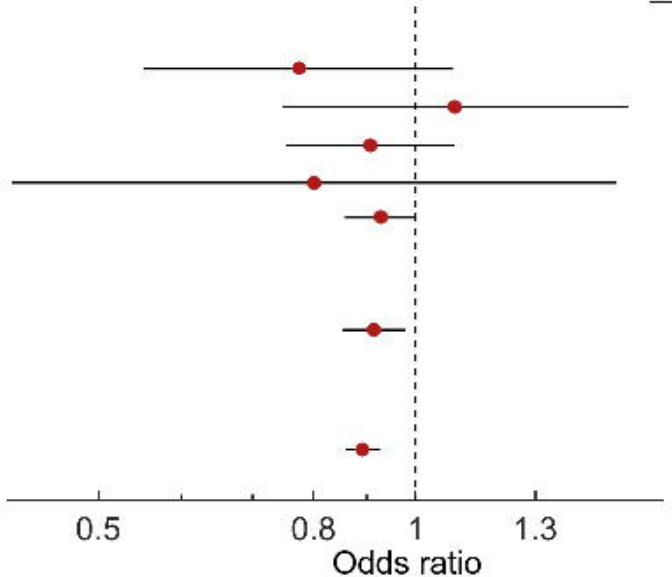


Fig 2. Odds ratios for COVID-19 hospitalization for African ancestry carriers of the ancestral splice variant. Summary effect in African ancestry individuals by a fixed effect meta-analysis of the five cohorts. Error bars give 95% confidence intervals.


A
B
Linkage disequilibrium (r2)
1
0.8
0.6
0.4
0.2
0
113.3
113.35
113.4
Chromosome 12 coordinate (Mb)

	Counts cases	Counts controls	OR (95% CI) rs10774671-G	P-value OR≠0
COVID-19 studies (African ancestry)				
UK BioBank	74	9,305	0.78 (0.55-1.09)	0.39
Penn Medicine BioBank	114	8,901	1.09 (0.75-1.59)	0.65
Columbia University Biobank	304	2,246	0.91 (0.75-1.09)	0.29
Québec COVID-19	50	50	0.80 (0.41-1.55)	0.51
Million Veteran Program	1,300	98,129	0.93 (0.86-1.00)	0.06
Summary effect (African ancestry)				
Fixed effect	1,842	118,631	0.91 (0.85-0.98)	0.03
Summary effect (European ancestry)				
COVID-19 Host Genetics Initiative	12,888	1,295,966	0.89 (0.86-0.93)	1.2e-9