ARTICLES
https://doi.org/10.1038/s41588-022-01042-x

nature genetics

Expanded COVID-19 phenotype definitions reveal distinct patterns of genetic association and protective effects

Genevieve H. L. Roberts[1,3], Raghavendran Partha[2,3], Brooke Rhead[2,3], Spencer C. Knight[2], Danny S. Park[2], Marie V. Coignet[2], Miao Zhang[2], Nathan Berkowitz[2], David A. Turrisini[2], Michael Gaddis[2], Shannon R. McCurdy[2], Milos Pavlovic[1], Luong Ruiz[2], Chodon Sass[1], AncestryDNA Science Team*, Asher K. Haug Baltzell[1], Harendra Guturu[2], Ahna R. Girshick[2], Catherine A. Ball[2], Eurie L. Hong[2] and Kristin A. Rand[2] ✉

Multiple COVID-19 genome-wide association studies (GWASs) have identified reproducible genetic associations indicating that there is a genetic component to susceptibility and severity risk. To complement these studies, we collected deep coronavirus disease 2019 (COVID-19) phenotype data from a survey of 736,723 AncestryDNA research participants. With these data, we defined eight phenotypes related to COVID-19 outcomes: four phenotypes that align with previously studied COVID-19 definitions and four 'expanded' phenotypes that focus on susceptibility given exposure, mild clinical manifestations and an aggregate score of symptom severity. We performed a replication analysis of 12 previously reported COVID-19 genetic associations with all eight phenotypes in a trans-ancestry meta-analysis of AncestryDNA research participants. In this analysis, we show distinct patterns of association at the 12 loci with the eight outcomes that we assessed. We also performed a genome-wide discovery analysis of all eight phenotypes, which did not yield new genome-wide significant loci but did suggest that three of the four 'expanded' COVID-19 phenotypes have enhanced power to capture protective genetic associations relative to the previously studied phenotypes. Thus, we conclude that continued large-scale ascertainment of deep COVID-19 phenotype data would likely represent a boon for COVID-19 therapeutic target identification.

SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), the virus that causes COVID-19, emerged in late 2019 and rapidly spread worldwide. The ongoing COVID-19 pandemic now represents one of the most severe pandemics in modern history and thus has been a central focus of biomedical research around the world. Early epidemiologic studies demonstrated that infection with SARS-CoV-2 results in a wide range of outcomes, ranging from asymptomatic infection to life-threatening viral pneumonia[1]. These studies were conducted during an early phase of the pandemic that predates the emergence of new variants of the SARS-CoV-2 virus[2,3], and thus, variations in viral strain do not explain the observed variability in outcomes. Furthermore, certain clinical risk factors, including increasing age, high body mass index and male sex, are associated with greater predisposition to severe COVID-19 disease[4–7], but these factors alone do not fully explain the remarkable variation in COVID-19 outcomes.

In addition to unexplained variation in disease severity once infected, there also may be host factors associated with susceptibility to infection with the SARS-CoV-2 virus; however, such associations remain poorly characterized for two reasons. The first is that the relationship between susceptibility to infection (defined as the probability of becoming infected given a SARS-CoV-2 exposure) and a host factor is confounded by the nature and degree of environmental exposure to the virus. Because it is not possible to precisely measure an individual's past environmental exposure to SARS-CoV-2, it is challenging to account for this confounding in epidemiologic studies of COVID-19. Second, there is dependence between infection severity and susceptibility to infection: if a person has severe COVID-19, then they must also be susceptible to infection. Therefore, host factors that appear to contribute to disease severity can be difficult to disentangle from factors that contribute to susceptibility to infection. For these reasons, the host factors that may contribute to susceptibility to infection with SARS-CoV-2 remain incompletely understood.

To investigate whether host genetic variation contributes to COVID-19 susceptibility and severity, multiple large GWASs of COVID-19 have been conducted[8–12]. These GWASs have identified multiple reproducible associations. For example, the earliest GWAS of a COVID-19 outcome, which investigated respiratory failure due to COVID-19 compared to unselected controls with <5% SARS-CoV-2 seropositivity, identified two loci: 3p21.31 (near *SLC6A20, LZTFL1, CCR9, FYCO1, CXCR6* and *XCR1*) and 9q34.2 (near *ABO*)[8]. Associations at these loci were later reproduced in independent GWAS cohorts[10,11]. Interestingly, however, there have been conflicting reports of whether the *ABO* locus associates with infection severity[8], susceptibility to infection[9,11] or both[10].

Although previous GWAS provide important insight into COVID-19 susceptibility and severity, most of these studies ascertained COVID-19 cases in medical clinics and hospitals, which can lead to an overrepresentation of cases with severe outcomes, such as hospitalization, intensive care unit (ICU) admission or ventilation. Thus, these previously studied ('established') phenotypes that

[1]AncestryDNA, Lehi, UT, USA. [2]AncestryDNA, San Francisco, CA, USA. [3]These authors contributed equally: Genevieve H. L. Roberts, Raghavendran Partha, Brooke Rhead. *A list of members and their affiliations appears in the Supplementary Information. ✉e-mail: kristinmuench@gmail.com

Table 1 | AncestryDNA COVID-19 survey cohort demographics

COVID-19 status			
Answer to 'Have you been swab tested for COVID-19, commonly referred to as coronavirus?'			*n* (%)
Yes, and was positive			8,868 (1%)
Yes, and was negative			64,161 (9%)
Yes, and my results are pending			5,117 (<1%)
No, but I have had flu-like symptoms with a fever at some point since the beginning of February 2020			86,035 (12%)
No, and I have not had flu-like symptoms at some point since the beginning of February 2020			572,542 (78%)
Basic demographics			
	COVID-19$^+$	**COVID-19$^-$**	**Full Ancestry COVID-19 cohort**
	n **= 8,868**	***n*** **= 64,161**	***n*** **= 736,723**
Median age (years)	49	56	57
Genetic sex, *n* (%)			
Female	5,684 (64%)	42,974 (67%)	490,912 (67%)
Male	3,164 (36%)	21,058 (33%)	244,353 (33%)
Genetic ancestry continental groupings[a], *n* (%)			
European ancestry (EUR)	5,524 (62%)	44,542 (69%)	537,512 (73%)
Admixed Amerindian ancestry (LAT)	1,118 (13%)	5,061 (8%)	47,301 (6%)
Admixed African-European ancestry (AA)	513 (6%)	2,695 (4%)	22,464 (3%)
Other ancestry	1,713 (19%)	11,863 (19%)	129,446 (18%)
Risk factors, *n* (%)			
Any preexisting health condition[b]	3,521 (44%)	33,034 (53%)	338,623 (47%)
Body mass index, median	28.5	28.35	28.29
Symptomatic (yes)	7,296 (87%)	—	29,849 (37%)[c]
Hospitalization (yes)	734 (10%)	—	2,057 (2%)[c]
Oxygen[d]	325 (45%)	—	567 (28%)[c]
Ventilation[d]	77 (11%)	—	118 (6%)[c]

[a]Genetic ancestry grouping definitions: admixed African-European ancestry includes 100% African ancestry; admixed Amerindian ancestry also includes 100% Amerindian ancestry. Extended Data Fig. 1 shows principal-component plots of ancestry groupings. [b]Preexisting health conditions include asthma, bone marrow transplant, cancer, cardiovascular disease, kidney disease, chronic obstructive pulmonary disease, diabetes, hypertension, organ failure requiring transplant, autoimmune disease, immunodeficiency and/or 'other' lung conditions. [c]Includes COVID-19$^+$, COVID-19 test result pending and individuals who were not tested but felt sick. [d]Percentages of total hospitalized.

can be readily gathered from medical clinics and hospitals consist of individuals with COVID-19 with symptoms that are severe enough to seek medical care. Given the wide symptom profile of those with COVID-19, the voluntary, self-reported nature of the AncestryDNA dataset allows for a complementary analysis of additional phenotypes that explore associations with mild COVID-19.

Here, we demonstrate that 'deep' phenotyping based on self-reported outcomes in a population with a large proportion of mild and subclinical cases can complement studies with primary case ascertainment in clinics. We used the AncestryDNA platform to conduct a comprehensive, 50+ question survey collecting detailed information of an individual's COVID-19 experience. In less than 4 months during an early phase of the pandemic, we collected survey data from 736,723 AncestryDNA research participants who consented to research. Because of the fast, detailed and large-scale nature of this ascertainment strategy, we are uniquely poised to investigate genetic associations with multiple COVID-19 phenotypes that probe a range of questions related to COVID-19 severity and susceptibility.

In this work, we used the self-reported survey responses to construct four phenotypes that align with previously studied COVID-19 phenotypes plus four additional, 'expanded' phenotypes that would be challenging to collect at large scale using typical hospital/clinical case ascertainment. For each of the eight phenotypes, we conducted a trans-ancestry GWAS meta-analysis. The results of this multiphenotype analysis reproduce the majority of associations identified by previous COVID-19 GWASs, shed light on whether previously identified loci associate with COVID-19 susceptibility or severity and suggest that certain 'expanded' phenotype definitions may yield new loci of therapeutic relevance.

Results

COVID-19 survey collection. To perform genetic studies of COVID-19, we conducted a comprehensive, 50+ question survey (Supplementary Note 1) of AncestryDNA customers that have consented to research to assess exposure, risk factors, symptomatology and demographic information, described in Supplementary Fig. 1. We collected 736,723 COVID-19 survey responses between April and August 2020 and used them to develop an expanded repertoire of phenotypes to investigate. Cohort demographics are available in Table 1 and Supplementary Table 1.

COVID-19 phenotype construction. In total, we defined eight COVID-19 phenotypes, which are summarized in Table 2. Four of these phenotypes were intended to mirror 'established' susceptibility or severity phenotype definitions from other large COVID-19 GWAS consortia[10,12] (*Hospitalized/Not_Hospitalized, Hospitalized/Unscreened, Positive/Negative, Positive/Unscreened*).

We hypothesized that phenotypes that focus on mild outcomes or absence of infection despite a strong exposure might be better

Table 2 | Summary of eight phenotype definitions

Phenotype code[a]	Case description[b]	Control description[b]	Goal	Cases	Controls
Positive/Negative	**COVID-19$^+$**	COVID-19$^-$	Reproduce clinically ascertained studies	5,373	35,901
Positive/Unscreened	**COVID-19$^+$**	Unscreened, but not known to be COVID-19$^+$	Reproduce clinically ascertained studies	5,373	95,027
Hospitalized/Not_Hospitalized	**COVID-19$^+$ and hospitalized**	COVID-19$^+$ and not hospitalized	Reproduce clinically ascertained studies	474	4,159
Hospitalized/Unscreened	**COVID-19$^+$ and hospitalized**	Unscreened, but not known have been COVID-19$^+$	Reproduce clinically ascertained studies	474	99,198
Exposed_Positive/Exposed_Negative	COVID-19$^+$ and had a cohabitant with a confirmed COVID-19 diagnosis	**COVID-19$^-$ and had a cohabitant with a confirmed COVID-19 diagnosis**	Study genetic susceptibility in individuals with a strong exposure event	2,022	1,060
Unscreened/Exposed_Negative	Unscreened, but not known to be COVID-19$^+$	**COVID-19$^-$ and had a cohabitant with a confirmed COVID-19 diagnosis**	Study genetic susceptibility in individuals with a strong exposure event	98,507	1,060
Symptomatic/Paucisymptomatic	COVID-19$^+$ and symptomatic	**COVID-19$^+$ and asymptomatic or paucisymptomatic**	Study genetic protection from severe outcomes if infected	4,353	391
Continuous_Severity_Score	COVID-19$^+$ score that combines nine different measures of COVID-19 symptom severity; higher scores correspond to more severe outcomes		Study genetic variants associated with both severe and mild outcomes simultaneously	4,952	N/A[c]

[a]Nomenclature for phenotype codes: Case_definition/Control_definition. Phenotype names are italicized to aid identification in the main text. [b]Case or control descriptions in bold represent the minority group. [c]The *Continuous_Severity_Score* phenotype is continuous, and thus, there are no cases and controls. Instead, a score is computed for each COVID-19$^+$ person.

able to detect protective genetic associations. Therefore, in addition to these 'established' phenotypes, we designed four 'expanded' phenotypes that assess either susceptibility to infection given a strong exposure or symptom severity. Biological susceptibility to infection is difficult to measure because the probability of contracting the virus depends on the nature of exposure, which is often unknown. We attempted to capture biological susceptibility to SARS-CoV-2 infection by conditioning on a known, strong exposure to the virus: household exposure. The positivity rate among respondents that reported cohabiting with a person with confirmed COVID-19 was ~65%, the highest positivity rate for any exposure we assessed. The *Exposed_Positive/Exposed_Negative* phenotype compared those with a household exposure who tested positive to those with a household exposure who tested negative, and *Unscreened/Exposed_Negative* focused on protection from infection by comparing those with a household exposure who tested negative to a large sample of unscreened controls.

To more deeply investigate symptom severity, we collected ordinal information (e.g., mild, moderate or severe) for 15 different COVID-19 symptoms (Supplementary Note 1). From responses to symptom severity questions, we created one binary phenotype, *Symptomatic/Paucisymptomatic*, which compares individuals with COVID-19 with at least one moderate or greater symptom to the remaining individuals with mild or asymptomatic infections. In addition, we aggregated responses from nine survey fields that were previously associated with severe outcomes[7] to create the *Continuous_Severity_Score* phenotype, in which lower scores correspond to lower symptom severity and higher scores correspond to increased symptom severity and elevated hospitalization rates (Fig. 1).

For all eight phenotypes, as cases corresponded to higher susceptibility or severity, all positive single-nucleotide polymorphism (SNP) effect estimates (β_{SNP}) can be interpreted as 'risk', and all negative estimates can be interpreted as 'protective.' By leveraging mild and asymptomatic cases and creating a continuous phenotype to reflect symptom severity, we attempt to capture the wide phenotypic spectrum of outcomes associated with COVID-19.

Trans-ancestry GWAS meta-analysis power. Total sample sizes for our trans-ancestry meta-analyses of AncestryDNA European (EUR), Admixed Amerindian (LAT) and Admixed African-European (AA) cohorts for all eight phenotypes are presented in Table 2 and Supplementary Table 2. Principal-component plots of the three cohorts are presented in Extended Data Fig. 1. Based on these prevalent sample sizes, we estimated power for the seven binary phenotypes assuming a causal variant relative risk (RR) of 1.25 and a type I error rate of 0.05 across a range of causal variant minor allele frequencies (MAFs) (Fig. 2a and Supplementary Table 3). All seven binary phenotypes were well powered to replicate common, moderate effect size variants at this nominal significance threshold ($P < 0.05$), all with power of at least 79% assuming causal variant MAF = 0.20. Interestingly, we found that several phenotypes with smaller numbers of screened controls had slightly higher power than the analogous phenotypes with a very large number of unscreened controls. For example, assuming a causal variant with MAF = 0.20 and RR = 1.25, the *Hospitalized/Unscreened* phenotype had power of 83% with 99,198 unscreened controls, whereas the *Hospitalized/Not_Hospitalized* phenotype had a higher power of 87% despite having only 4,159 screened controls. This observation is not surprising, given that using unscreened controls decreases power by increasing outcome misclassification[13,14] and that adding additional controls beyond a case/control ratio of ~1:4 typically yields very small increases in power[15].

Replication of 12 COVID-19-associated SNPs with eight phenotypes. To explore how known COVID-19 risk loci associate with these different phenotype definitions, we investigated 12 independent SNPs ($r^2 < 0.05$) that were identified in at least one of two recent, large COVID-19 meta-analyses: the October 2020 data release from the COVID-19 Host Genetics Initiative (HGI) or

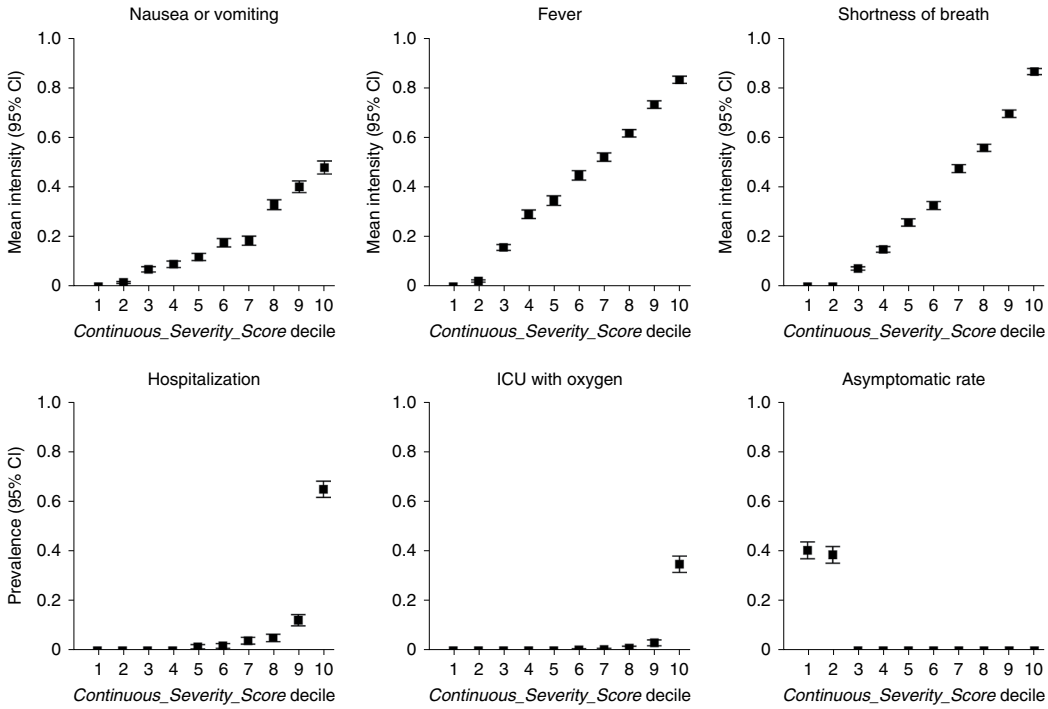


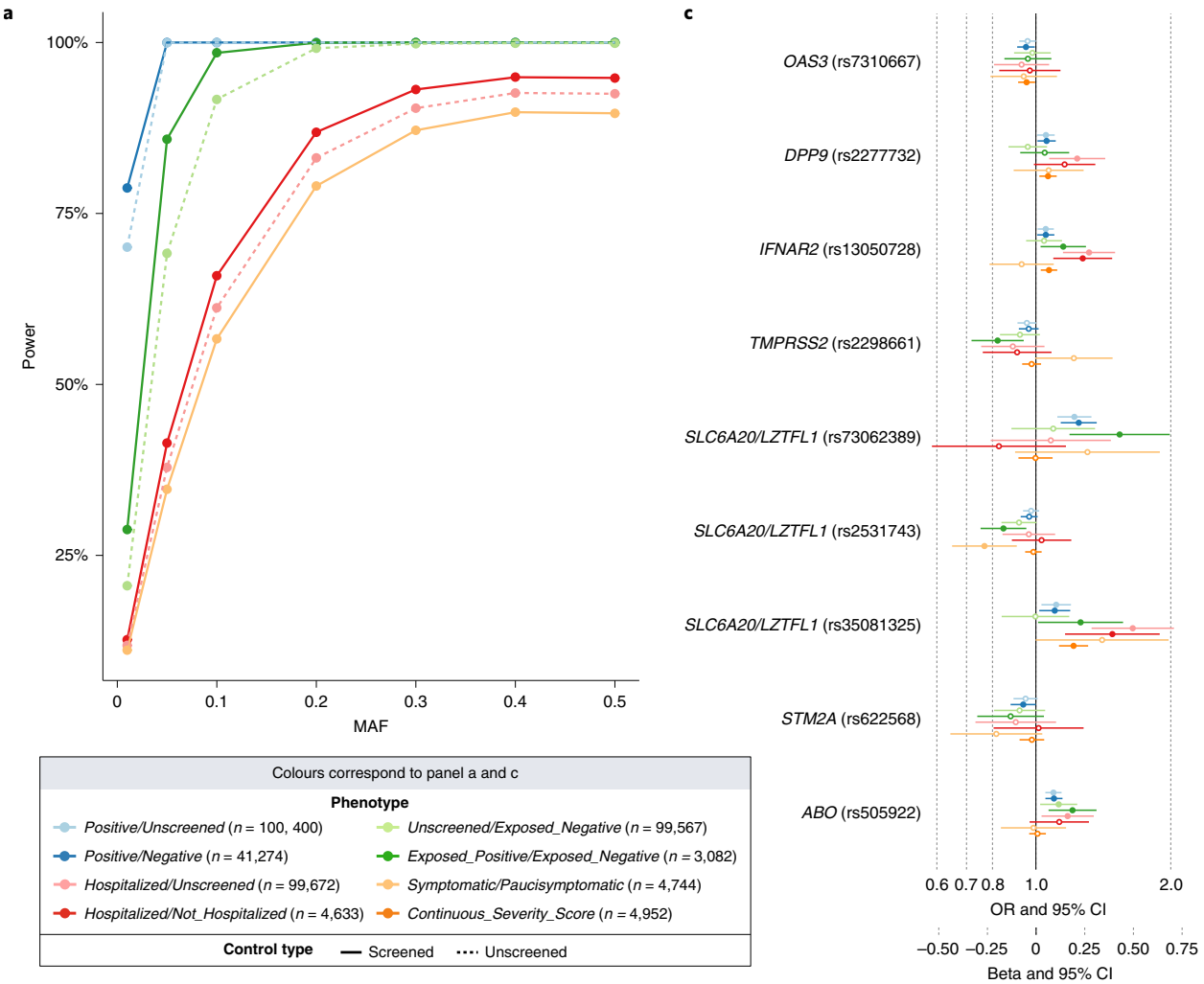
Fig. 1 | COVID-19 *Continuous_Severity_Score* captures multiple aspects of symptom severity among COVID-19⁺ individuals. The *Continuous_Severity_Score* was derived from the first principal component across nine survey fields related to COVID-19 clinical outcomes, including three symptoms, hospitalization, ICU admittance and other severe complications due to COVID-19 illness (Methods). Plots reflect mean symptom severity (top three panels) or prevalence (bottom three panels) for several fields as a function of ascending *Continuous_Severity_Score* decile. Symptom information was encoded as follows: 0 = none, 0.2 = very mild, 0.4 = mild, 0.6 = moderate, 0.8 = severe and 1.0 = very severe. A paucisymptomatic case corresponds to an individual reporting symptoms of mild intensity or less. Squares represent the estimate, and error bars represent the 95% confidence intervals (CIs) for each estimate ($n = 4{,}952$ biologically independent samples).

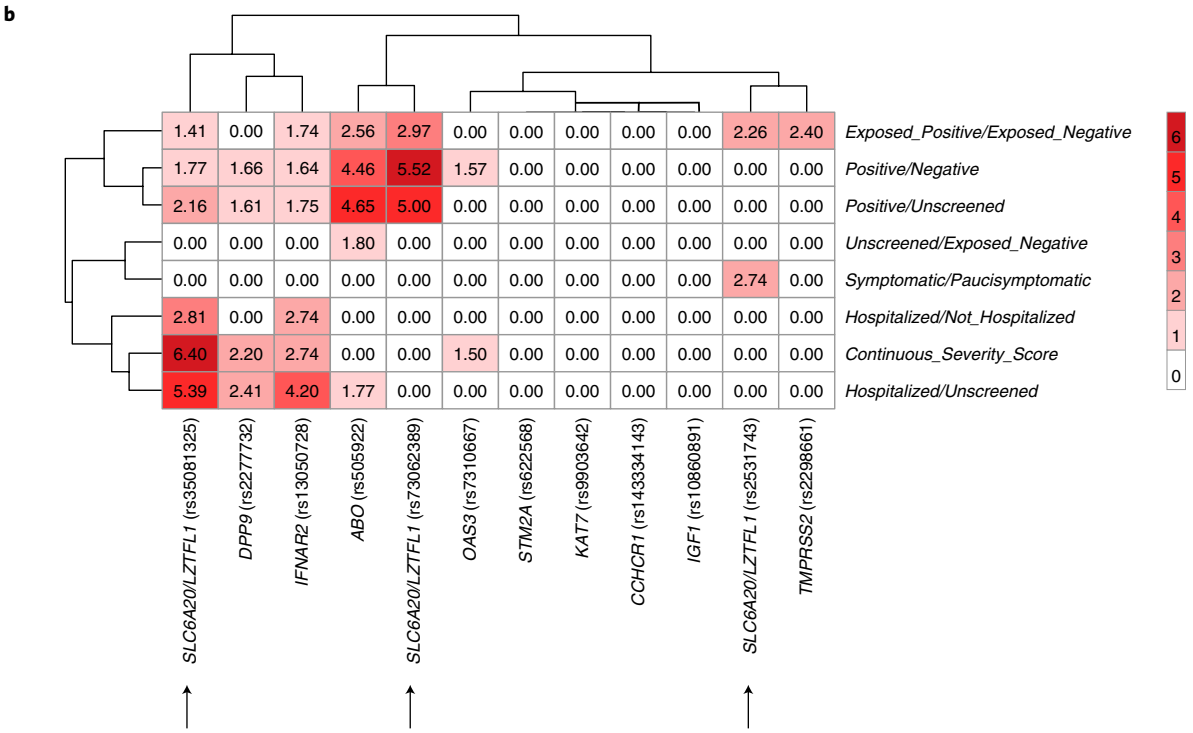
Horowitz et al. (Supplementary Table 4). We assessed association of these 12 SNPs with all eight phenotypes, defining evidence of replication as an AncestryDNA trans-ancestry $P < 0.05$ and consistent direction of effect with the prior study. We note that a small percentage of research participants in these prior studies overlap our study, quantified in Supplementary Fig. 2.

Eight of 12 SNPs replicated in at least one of our phenotypes (Fig. 2b,c). This result demonstrates that our phenotypes, which are based entirely on self-reported outcomes, recapitulate many of the same associations previously observed in clinically ascertained studies. Hierarchical clustering of the replication *P* values at these known loci revealed two clusters of phenotypes: three severity phenotypes produced a similar pattern of replication (*Hospitalized/Not_Hospitalized*, *Hospitalized/Unscreened*, *Continuous_Severity_Score*), and three susceptibility phenotypes produced a similar pattern of replication (*Positive/Negative*, *Positive/Unscreened*, *Exposed_Positive/Exposed_Negative*). Phenotypes in these clusters are likely capturing similar genetic associations. The two remaining phenotypes (*Symptomatic/Paucisymptomatic*, *Unscreened/Exposed_Negative*) replicated the 12 SNPs poorly and may capture different genetic associations, warranting further investigation.

Power for nominal replication ($P < 0.05$) was broadly similar for most phenotypes at most MAF thresholds, although the severity phenotypes had somewhat lower power overall compared to the susceptibility phenotypes (Fig. 2a). Despite greater power among the susceptibility phenotypes, there are instances of more significant association with the less powerful hospitalization phenotypes (e.g., rs13050728 and rs35081325), suggesting that the observed

Fig. 2 | Replication of 12 independent SNPs identified by previous studies. a, Estimated power to replicate a SNP for all binary phenotypes at seven different MAFs assuming a type I error rate of 0.05 and per-allele RR of 1.25. Solid lines represent phenotypes with screened controls, and dashed lines represent unscreened controls. *Continuous_Severity_Score* was excluded from this panel because SNP effect size is on a different scale. Phenotypes are presented in the same order as in Fig. 1a,c. In the legend, (*n*) represents the number of biologically independent samples for each phenotype. **b**, Heatmap of the trans-ancestry meta-analysis $-\log_{10}(P \text{ values})$ for each of the eight phenotypes and the 12 independent SNPs identified in a previous study. Red blocks denote replication, with darker shades of red corresponding to more significant trans-ancestry *P* values. All associations that failed to replicate (trans-ancestry $P > 0.05$) or had inconsistent effect direction relative to the discovery study were assigned $-\log_{10}(P \text{ value}) = 0$. SNP and phenotype labels were ordered by hierarchical clustering, with corresponding dendrograms shown on the top and left of the figure. Arrows identify the three independent SNPs in the chr3p21 locus. **c**, Forest plot that shows minor allele effect sizes for nine of the 12 SNPs that nominally replicated (trans-ancestry $P < 0.05$, any phenotype). *STM2A* is included, but had inconsistent direction of effect with the discovery study. Circles indicate effect estimates, colours correspond to the phenotypes in panel **a**, and horizontal lines represent 95% confidence intervals. Filled circles indicate $P < 0.05$. *Continuous_Severity_Score* was the only continuous phenotype, and therefore, the reported effect estimate is β_{SNP}, which can be interpreted as standard deviations from the mean per each copy minor allele. For all other phenotypes, per-allele ORs are reported. All *P* values referred to in this figure are not adjusted for multiple comparisons and are two-sided and based on an inverse-variance-weighted fixed-effects trans-ancestry meta-analysis.


a
100%
75%
50%
25%
Power
0
0.1
0.2
0.3
0.4
0.5
MAF
Colours correspond to panel a and c
Phenotype
Positive/Unscreened (n = 100, 400)
Unscreened/Exposed_Negative (n = 99,567)
Positive/Negative (n = 41,274)
Exposed_Positive/Exposed_Negative (n = 3,082)
Hospitalized/Unscreened (n = 99,672)
Symptomatic/Paucisymptomatic (n = 4,744)
Hospitalized/Not_Hospitalized (n = 4,633)
Continuous_Severity_Score (n = 4,952)
Control type
Screened
Unscreened
c
OAS3 (rs7310667)
DPP9 (rs2277732)
IFNAR2 (rs13050728)
TMPRSS2 (rs2298661)
SLC6A20/LZTFL1 (rs73062389)
SLC6A20/LZTFL1 (rs2531743)
SLC6A20/LZTFL1 (rs35081325)
STM2A (rs622568)
ABO (rs505922)
0.6 0.7 0.8 1.0 2.0
OR and 95% CI
−0.50 −0.25 0 0.25 0.50 0.75
Beta and 95% CI


b
1.41 0.00 1.74 2.56 2.97 0.00 0.00 0.00 0.00 0.00 2.26 2.40 Exposed_Positive/Exposed_Negative
1.77 1.66 1.64 4.46 5.52 1.57 0.00 0.00 0.00 0.00 0.00 0.00 Positive/Negative
2.16 1.61 1.75 4.65 5.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 Positive/Unscreened
0.00 0.00 0.00 1.80 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 Unscreened/Exposed_Negative
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 2.74 0.00 Symptomatic/Paucisymptomatic
2.81 0.00 2.74 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 Hospitalized/Not_Hospitalized
6.40 2.20 2.74 0.00 0.00 1.50 0.00 0.00 0.00 0.00 0.00 0.00 Continuous_Severity_Score
5.39 2.41 4.20 1.77 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 Hospitalized/Unscreened
SLC6A20/LZTFL1 (rs35081325)
DPP9 (rs2277732)
IFNAR2 (rs13050728)
ABO (rs505922)
SLC6A20/LZTFL1 (rs73062389)
OAS3 (rs7310667)
STM2A (rs622568)
KAT7 (rs9903642)
CCHCR1 (rs143334143)
IGF1 (rs10860891)
SLC6A20/LZTFL1 (rs2531743)
TMPRSS2 (rs2298661)
6
5
4
3
2
1
0

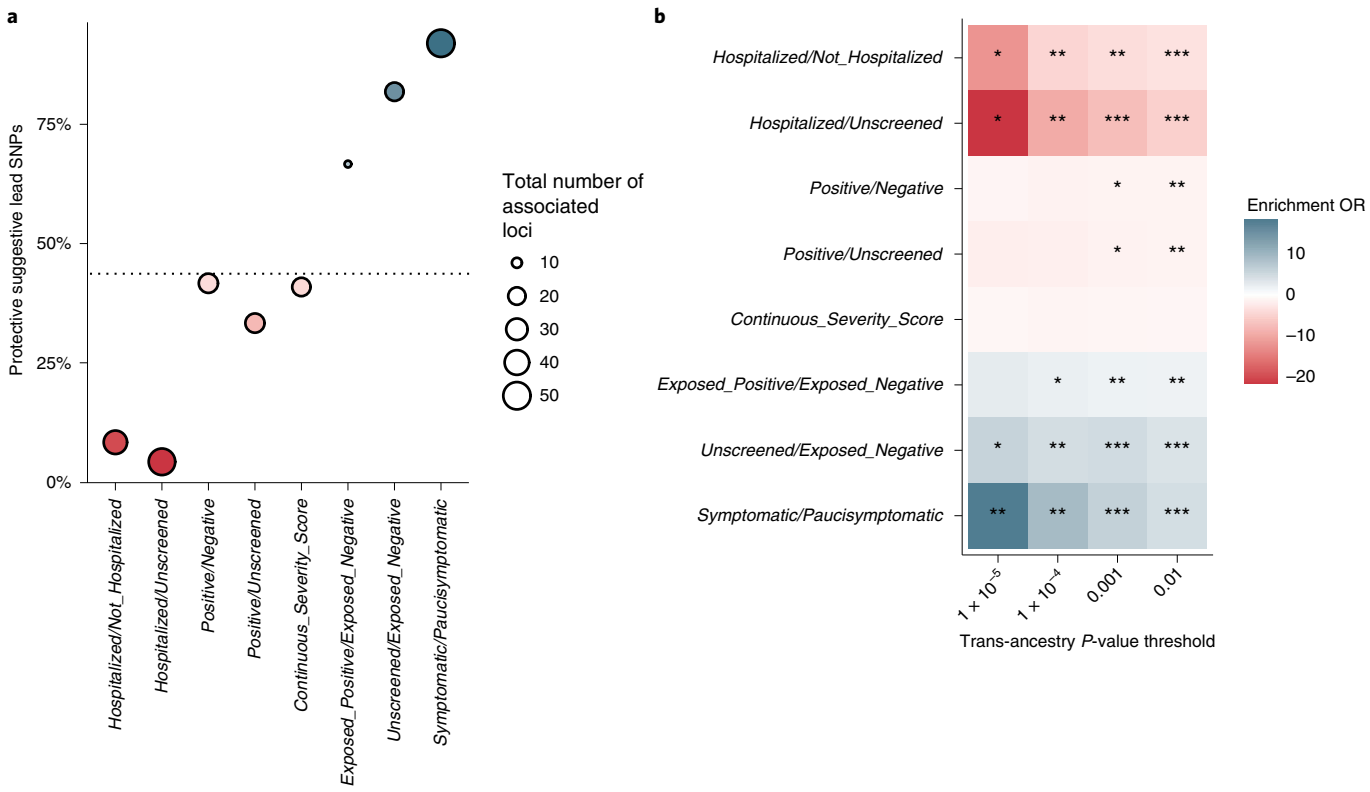


Fig. 3 | Three 'expanded' phenotypes are enriched for minor alleles associated with a protective direction of effect. a, The *y* axis represents the percentage of suggestively associated lead SNPs (two-sided inverse-variance-weighted trans-ancestry $P<1\times10^{-5}$) for which the minor allele was protective. The size of each point represents the total number of independent suggestive SNPs for each of the eight phenotypes. The horizontal dashed line represents the mean proportion of protective minor alleles across all eight phenotypes. Point colors match the protective enrichment color scale described in panel **b**. **b**, Protective (blue) or risk (red) enrichment at four two-sided CMH enrichment *P*-value thresholds for each of the eight phenotypes relative to the remaining seven phenotypes. Darker shades correspond to greater magnitude of enrichment (larger CMH OR). Asterisks represent significance of CMH enrichment test (unadjusted two-sided $^{*}P<0.0016$; $^{**}P<1\times10^{-10}$; $^{***}P<1\times10^{-100}$).

patterns of association are not simply driven by differences in power but rather that these phenotypes capture different aspects of biology related to the COVID-19 host response.

The associations in the chr3 and chr9 regions are of special interest, as associations with these loci were among the first to be observed[8]. The chr3p21 region is complex (Supplementary Note 2 and Extended Data Fig. 2), with at least three independent SNPs ($r^2<0.04$; Supplementary Table 5) identified in previous literature. These three SNPs span a 54-kb region within chr3p21 near a cluster of immune genes, including *CCR9, CXCR6* and *XCR1*, but nearest to *LZTFL1* and *SLC6A20*. The HGI severity study identified rs35081325 (HGI hospitalized COVID$^+$ versus population (B2) $P=6.89\times10^{-52}$; HGI B2 odds ratio (OR)=1.82), which is most associated with the severity cluster of phenotypes in our study: *Continuous_Severity_Score* ($P=4\times10^{-7}$, $\beta_{SNP}=0.19$ standard deviations), *Hospitalized/Unscreened* ($P=4.07\times10^{-7}$, $OR=1.65$), *Hospitalized/Not_Hospitalized* ($P=1.54\times10^{-3}$, $OR=1.48$). This SNP also associated with three susceptibility phenotypes but with weaker effect estimates (ORs ranging from 1.10-1.26). Thus, rs35081325, which lies in an intron of *LZTFL1*, appears to have a consistently stronger association with increased infection severity. By contrast, rs73062389, which lies 54 kb away in an intron of *SLCA620*, was identified in the HGI susceptibility study (HGI COVID$^+$ versus population (C2) $P=1.09\times10^{-9}$; HGI C2 $OR=1.23$) and is strongly associated with the susceptibility cluster of phenotypes in our analysis: *Positive/Negative* ($P=3.02\times10^{-6}$, $OR=1.25$), *Positive/Unscreened* ($P=9.90\times10^{-6}$, $OR=1.22$) and *Exposed_Positive/Exposed_Negative* ($P=1.07\times10^{-3}$, $OR=1.54$). Furthermore, rs73062389 is not associated with any of our severity cluster phenotypes and thus seems to specifically confer increased susceptibility risk. Yet another independent SNP in chr3p21, rs2531743, was discovered near *SLCA620* by Horowitz et al. using a COVID$^+$ versus COVID$^-$ definition. Unlike the other chr3p21 signals, the minor allele of rs2531743 was reported as protective (Horowitz et al. $P=2\times10^{-6}$, Horowitz OR=0.94). We also observed a protective effect in two phenotypes for this SNP: *Symptomatic/Paucisymptomatic* ($P=1.82\times10^{-3}$, $OR=0.77$) and *Exposed_Positive/Exposed_Negative* ($P=5.46\times10^{-3}$, $OR=0.85$). Thus, all three independent signals in this region associate with a distinct set of phenotypes.

Associations near *ABO* on chr9q34.2, the gene that determines blood type, have also been observed in multiple, early COVID-19 GWASs, somewhat inconsistently with severity phenotypes and more consistently with susceptibility phenotypes (Supplementary Table 6). In our analysis, the lead *ABO* SNP, rs505922, replicated in all four susceptibility phenotypes and only one severity phenotype, the *Hospitalized/Unscreened* analysis, which used a large number of unscreened controls. We speculate that the *ABO* locus specifically confers increased susceptibility to infection, but inclusion of unscreened controls in severity studies can induce susceptibility associations as hospitalized cases must be susceptible to infection, but an unscreened control group may or may not be susceptible. Thus, the *Hospitalized/Unscreened* phenotype simultaneously captures aspects of both susceptibility and severity.

The four SNPs that did not replicate in our study, near *STM2A, KAT7, CCHCR1* and *IGF1*, were all originally identified using severity phenotype definitions, and all had MAFs in the range of 0.08–0.13 (Supplementary Table 4). Power for severity phenotypes in our study for MAFs in that range was only about 50–70% (Fig. 2a), so the lack of replication is likely due to insufficient statistical power.

Genome-wide discovery analysis reveals an enrichment of protective effects for three phenotypes. We additionally conducted a genome-wide trans-ancestry discovery study for each of the eight phenotypes. No phenotype–SNP association pairs surpassed a conservative Bonferroni-corrected genome-wide discovery significance threshold of $P < 1.25 \times 10^{-9}$; however, we examined SNPs that reached a less stringent significance threshold of $P < 1 \times 10^{-5}$ (Supplementary Table 7) to investigate trends. In this genome-wide analysis, we found that three of the 'expanded' phenotypes identified a much larger proportion of protective minor alleles than other phenotypes (Fig. 3a and Supplementary Table 8; *Symptomatic/Paucisymptomatic* = 92% protective; *Unscreened/Exposed_Negative* = 82% protective; *Exposed_Positive/Exposed_Negative* = 67% protective). Furthermore, the 'established' phenotypes more often identified minor alleles associated in the risk direction (*Hospitalized/Unscreened* = 96% risk; *Hospitalized/Not_Hospitalized* = 92% risk; *Positive/Unscreened* = 67% risk; *Positive/Negative* = 58% risk). To explore these trends further, we tested for enrichment of protective or risk effects for each phenotype at four different suggestive significance thresholds (Fig. 3b). There was significant enrichment for protective effects for lead SNPs at all four *P*-value thresholds for both *Symptomatic/Paucisymptomatic* and *Unscreened/Exposed_Negative*, with the magnitude of protective enrichment increasing with increasing significance threshold. This result is not surprising but strongly suggests that using additional phenotype definitions that interrogate the full spectrum of symptom severity can identify additional, often protective, associations.

Discussion

We examined genetic association with eight different COVID-19 phenotype definitions, four of which had not yet been explored due to the difficulty of large-scale clinical ascertainment of detailed phenotype information pertaining to COVID-19 exposure and symptom severity. A power analysis of these phenotypes revealed that using a smaller number of screened controls versus a very large number of unscreened controls can be more powerful when the screened control/case ratio is already large. This result highlights the need for thoughtful approaches around case and control selection and shows that a well-powered genetic study of COVID-19 may not require a large number of unscreened controls.

We observe that eight of 12 previously identified COVID-19 genetic signals associate with at least one of the eight phenotypes. This strong replication of loci identified by clinically ascertained studies confirms that phenotyping based on well-designed self-report studies is valid. There are distinct patterns of association with these known COVID-19 loci, suggesting there are true biological differences in genetic susceptibility to severe infection, symptomatic infection and the probability of becoming infected if exposed. Additionally, our findings demonstrate that three previously identified signals in the chr3p21 *LZTFL1*/*SLC6A20* region each associate with a different set of the phenotypes we investigated, suggesting that variation in this region modulates multiple aspects of COVID-19 susceptibility and severity. The independent associations in this region with a range of COVID-19 phenotypes may represent a kind of 'allelic series' or natural dose–response curve between genetic variation and related disease phenotypes. For other diseases, the identification of genes with allelic series has indicated a central role for the causal gene in disease pathobiology, and such genes are often considered to represent strong candidates as therapeutic targets[16–19]. We thus we believe that further investigation into this region is extremely important.

Genome-wide, we observed that suggestive associations for 'established' hospitalization phenotypes were enriched for risk effects whereas three of the new, 'expanded' phenotypes (*Symptomatic/Paucisymptomatic*, *Unscreened/Exposed_Negative*, *Exposed_Positive/Exposed_Negative*) were enriched for protective effects. This finding is important, as it suggests that certain ways of defining phenotypes may be more likely to identify a noteworthy subcategory of protective variants, protective loss-of-function variants, which are desirable drug targets because it is generally easier to mimic their effects pharmacologically[20]. Thus, specifically investigating mild and subclinical self-reported phenotypes may yield new genetic associations that are more likely to be therapeutically actionable[21].

In summary, the AncestryDNA self-reported dataset allowed a complementary analysis of more granular phenotypes in a population enriched for mild outcomes compared to clinically ascertained studies enriched for severe outcomes. We found promising evidence that exploring new phenotypes in such populations will yield new genetic associations, particularly those that confer protection against the novel coronavirus SARS-CoV-2. Self-reported medical data have been previously demonstrated to support replication and new discovery of associations as well as expand the number of cases for GWAS[22,23]. Further studies that leverage mild and subclinical self-reported symptoms to probe the full symptom profile of disease may yield new insights into this disease and other diseases with complex symptom profiles, ultimately helping uncover disease mechanisms and providing potential therapeutic targets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01042-x.

Received: 22 January 2021; Accepted: 2 March 2022;
Published online: 11 April 2022

References

1. Sun, J. et al. COVID-19: epidemiology, evolution, and cross-disciplinary perspectives. *Trends Mol. Med.* **26**, 483–495 (2020).
2. Tregoning, J. S., Flight, K. E., Higham, S. L., Wang, Z. & Pierce, B. F. Progress of the COVID-19 vaccine effort: viruses, vaccines and variants versus efficacy, effectiveness and escape. *Nat. Rev. Immunol.* **21**, 626–636 (2021).
3. Viana, R. et al. Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
4. Zhou, F. et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
5. Cummings, M. J. et al. Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *Lancet* **395**, 1763–1770 (2020).
6. Lopez, L. III et al. Racial and ethnic health disparities related to COVID-19. *JAMA* **325**, 719–720 (2021).
7. Knight, S. C. et al. COVID-19 susceptibility and severity risks in a survey of over 500,000 people. Preprint at *medRxiv* https://www.medrxiv.org/content/10.1101/2020.10.08.20209593 (2021).
8. Ellinghaus, D. et al. Genomewide association study of severe covid-19 with respiratory failure. *N. Engl. J. Med.* **383**, 1522–1534 (2020).
9. Horowitz, J. E. et al. Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. *Nat. Genet.* https://doi.org/10.1038/s41588-021-01006-7 (2022).
10. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).
11. Shelton, J. F. et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nat. Genet.* **53**, 801–808 (2021).

12. Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
13. Hodge, S. E., Subaran, R. L., Weissman, M. M. & Fyer, A. J. Designing case-control studies: decisions about the controls. *Am. J. Psychiatry* **169**, 785–789 (2012).
14. Moskvina, V., Holmans, P., Schmidt, K. M. & Craddock, N. Design of case-controls studies with unscreened controls. *Ann. Hum. Genet.* **69**, 566–576 (2005).
15. Hong, E. P. & Park, J. W. Sample size and statistical power calculation in genetic association studies. *Genomics Inform.* **10**, 117 (2012).
16. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
17. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P. R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021).
18. Lichou, F. & Trynka, G. Functional studies of GWAS variants are gaining momentum. *Nat. Commun.* **11**, 1–4 (2020).
19. Plenge, R. M. Priority index for human genetics and drug discovery. *Nat. Genet.* **51**, 1073–1075 (2019).
20. Harper, A. R., Nayee, S. & Topol, E. J. Protective alleles and modifier variants in human health and disease. *Nat. Rev. Genet.* **16**, 689–701 (2015).
21. Farrelly, C. 'Positive biology' as a new paradigm for the medical sciences: focusing on people who live long, happy, healthy lives might hold the key to improving human well-being. *EMBO Rep.* **13**, 186–188 (2012).
22. Tung, J. Y. et al. Efficient replication of over 180 genetic associations with self-reported medical data. *PLoS One* **6**, e23473 (2011).
23. DeBoever, C. et al. Assessing digital phenotyping to enhance genetic studies of human diseases. *Am. J. Hum. Genet.* **106**, 611–622 (2020).

Methods

Ethics statement. All data for this research project were from individuals who provided prior informed consent to participate in AncestryDNA's Human Diversity Project, as reviewed and approved by our external institutional review board, Advarra (formerly Quorum). All data were deidentified before use.

Study population. Self-reported COVID-19 outcomes were collected through the Personal Discoveries Project, a survey platform available to AncestryDNA customers via the web and mobile applications. The COVID-19 survey ranged from 39 to 71 questions, depending on the initial COVID-19 test result reported. Supplementary Fig. 1 describes the flow of the topics assessed in each section of the survey, and the complete set of questions is in the Supplementary Note 1. Analyses presented here were performed with data collected between 22 April and 3 August 2020.

To participate in the COVID-19 survey, participants were required to meet the following criteria: 18 years of age or older, resident of the United States, existing AncestryDNA customer who has consented to participate in research and able to complete a short survey. The survey was designed to assess self-reported COVID-19 positivity and severity, as well as susceptibility and known risk factors, including community exposure and known contacts with individuals diagnosed with COVID-19.

Binary phenotype definitions. In total, we assessed eight phenotypes, which are summarized in Table 2. The full survey is available in Supplementary Note 1, but key definitions for this work include testing positive or negative, hospitalization, asymptomatic cases and cohabitant exposure. COVID-19 positivity or negativity was assessed by the question 'Have you been swab tested for COVID-19, commonly referred to as coronavirus?'. Hospitalization due to COVID-19 illness was used as one binary measure of severity and was assessed with the question 'Were you hospitalized due to these symptoms?'. Asymptomatic individuals were defined as those that were positive for COVID-19 and either answered 'No' to the question 'Did you experience symptoms as a result of your condition?' or answered one of 'None', 'Very mild' or 'Mild' to all 15 questions related to symptom severity. High exposure to COVID-19 was assessed through having a COVID-19 positive cohabitating person, assessed by the question 'Has someone in your household tested positive for COVID-19?'.

Continuous severity phenotype creation. A continuous severity score was derived by computing the first principal component across nine survey fields related to COVID-19 clinical outcomes. Six of the nine questions were binary: hospitalization, ICU admittance with oxygen, ICU admittance with ventilation, septic shock, respiratory failure and organ failure due to COVID-19. Binary responses were encoded as 0 for 'No' and 1 for 'Yes'. Three questions related to shortness of breath, fever and nausea/vomiting symptoms were encoded as a unit-scaled variable based on the following mapping: 0 = none, 0.2 = very mild, 0.4 = mild, 0.6 = moderate, 0.8 = severe and 1.0 = very severe. These three symptoms were chosen based on prior literature indicating their positive association with COVID-19 hospitalization[7]. The following assumptions were made so that a score could be calculated for most participants who reported a positive COVID-19 test result: (1) participants who responded 'No' to the question 'Did you experience symptoms as a result of your condition?' were not presented with additional questions regarding symptomatology or hospitalization and thus were encoded as 0 for all individual symptoms (shortness of breath, fever or nausea/vomiting), hospitalization, ICU admittance and severe complications due to COVID-19 illness; (2) participants who responded 'No' to the question 'Were you hospitalized due to these symptoms?' were not presented any further questions regarding hospitalization and thus were encoded as 0 for ICU admittance and supplemental oxygen; and (3) participants who declined to answer a question about complications due to COVID-19 illness such as septic shock, respiratory failure and organ failure were encoded as 0 for those complications (<2% of all participants for whom continuous severity was scored).

Power analysis. Power analysis was performed with the Purcell Power Calculator for case–control discrete traits (https://zzz.bwh.harvard.edu/gpc/cc2.html, created 24 October 2008). Power was not computed for the *Continuous_Severity_Score*, because this phenotype is continuous and thus the effect size is on a different scale that is not comparable to the other seven binary phenotypes. Allelic power for each phenotype was computed based on the 'Prevalence', 'Minority Class Cases', 'Control:Case' and 'Unselected controls?' fields as defined in Supplementary Table 3. For all seven phenotypes, the same effect size, type I error rates and MAFs were assumed: (1) 'Genotype relative risk Aa' = 1.25 (equivalent protective RR Aa = 0.80), (2) 'Genotype relative risk AA' = 1.25^2 = 1.56, (3) 'User-defined type I error rate' = 0.05 and (4) 'High risk allele frequency (A)' ∈ {0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50}. For all phenotypes, perfect linkage disequilibrium (LD) tagging of the causal variant was assumed, and thus 'D-prime' = 1 and 'Marker allele frequency (B)' = 'High risk allele frequency (A)'.

Genotyping. Customer genotype data for this study were generated using one of ten versions of a custom Illumina Human OmniExpress genotyping array. Array-based genotyping and SNP calling were performed by either Illumina with the GenotypeStudio Platform or Quest/Athena Diagnostics, a Clinical Laboratory Improvement Amendments-certified genotyping lab. These providers return called genotypes. To ensure quality of each dataset, a sample passes a number of quality control checks, including identifying duplicate samples, removing individuals with a per-sample call rate <98%, and identifying discrepancies between reported sex and genetically inferred sex. Samples that pass all quality control tests proceed to the analysis pipeline; samples that fail one or more tests must be recollected or manually cleared for analysis by lab technicians. Array markers with per-variant call rate <0.98 and array markers that had overall allele frequency differences of >0.10 between any two array versions were additionally removed before downstream analyses.

Defining ancestry cohorts. We defined three separate ancestry cohorts: EUR, LAT and AA (Extended Data Fig. 1). We assigned COVID-19 survey respondents to one of these ancestry groups with a proprietary algorithm that estimates continental admixture proportions. Briefly, this algorithm uses a hidden Markov model to estimate unphased diploid ancestry across the genome by comparing haplotype structure to a reference panel of ~56,000 samples that represent 77 global regions (details available in the Ancestry DNA Ethnicity Estimate White Paper at https://www.ancestrycdn.com/support/us/2021/09/ethnicity2021whitepaper.pdf). The reference panel consists of a combination of AncestryDNA customers and publicly available datasets and is designed to reflect global diversity. From our total cohort of 736,723 individuals who participated in the COVID-19 survey as of 3 August 2020, 537,512 (73%) individuals were designated to the EUR group, 22,464 (3%) to the AA group and 47,301 (6%) to the LAT group, and the remainder were not assigned to any ancestry group (Table 2 and Supplementary Table 2).

Removal of related individuals. AncestryDNA's identity-by-descent inference algorithm was used to estimate the relationship between pairs of individuals (details available in the Ancestry DNA Matching White Paper at https://www.ancestry.com/dna/resource/whitePaper/AncestryDNA-Matching-White-Paper.pdf). Pairs with estimated separation of fewer than four meioses were considered close relatives. For all close relative pairs, one individual was randomly selected for exclusion from our study. In total, we excluded 60,379 (~8%) individuals from analysis due to relatedness.

Calculation of principal components. For each population described above, genetic principal components were calculated to include in the association studies to control residual population structure and were computed using FlashPCA 2.0.2 (ref. [24]). Input genotypes were LD-pruned and filtered using the PLINK 1.9 indep-pairwise command with a window size of 100 SNPs, step size of 5 SNPs and r^2 threshold of 0.2 and were filtered to remove SNPs with MAF < 0.05 and missing call rate >0.001.

Imputation. Samples were imputed to the Haplotype Reference Consortium reference panel version 1.1, which consists of 27,165 total individuals and 36 million variants. The Haplotype Reference Consortium reference panel does not include indels; consequently, indels are not present in the results of our analyses. We determined best-guess haplotypes with Eagle[25] version 2.4.1 and performed imputation with Minimac4 version 1.0.1. We used 1,077,214 unique variants as input and 8,187,660 imputed variants were retained in the final dataset. For these variants, we conservatively restricted our analyses to variants with MAF > 0.01 and Minimac4 R^2 > 0.30 using imputed dosages for all variants regardless of whether they were originally genotyped.

GWASs in each Ancestry population. We conducted separate GWASs for the EUR, LAT and AA populations described above. For the EUR population only, we conducted sex-stratified GWASs and meta-analyzed the results via inverse-variance weighting implemented in METAL[26] (version released 25 March 2011).

For each individual population (EUR female, EUR male, LAT and AA) and each of the eight phenotypes, we conducted a GWAS assuming an additive genetic model with PLINK2.0. Imputed genotype dosage value was the primary predictor. The following were included as fixed-effect covariates: principal components 1–25 (described above), array platform, orthogonal age and orthogonal age^2. Orthogonal polynomials were used to eliminate collinearity between age and age^2, and were calculated in R version 3.6.0 with base function poly(age, degree = 2). We additionally used PLINK2.0 to remove variants with Minimac4 imputation quality R^2 < 0.3 or with MAF < 0.01. See the Supplementary Methods for a list of PLINK2.0 flags used for each analysis. All variant effect estimates are adjusted for the 28 covariates described above.

Trans-ancestry meta-analysis. For each phenotype, we additionally performed a trans-ancestry meta-analysis of the sex-combined EUR, AA and LAT summary statistics, again using fixed-effect inverse-variance weighting implemented in METAL[26] (version released 25 March 2011). These summary statistics were used to assess replication of the 12 SNPs and investigate enrichment of protective effects, as defined in the next sections. In the genome-wide analysis, we only examined SNPs that passed quality control in all individual populations. We implemented a

stringent, Bonferroni-corrected significance threshold by dividing a trans-ancestry genome-wide significance threshold of $P < 1 \times 10^{-8}$ by the eight phenotypes, which results in $P < 1.25 \times 10^{-9}$. Suggestive significance followed the definition used by the HGI consortium of $P < 1 \times 10^{-5}$. All reported P values are two sided. Manhattan plots for each of the eight trans-ancestry meta-analyses are provided in Extended Data Fig. 3, and corresponding quantile–quantile plots and genomic inflation factors are provided in Extended Data Fig. 4.

Replication of 12 independent SNPs from previous studies. We manually curated a list of 12 independent SNPs that represent lead loci identified by either HGI or Horowitz et al. Eight of the 12 SNPs were lead SNPs in HGI's October 2020 data freeze 4, without 23andMe data included the data release. These eight SNPs were the most-associated marker at any locus achieving $P < 5 \times 10^{-8}$ in the European hospitalization versus population ('ANA_B2') or European COVID-19^{+} versus population ('ANA_C2') study. The remaining four SNPs were selected from Fig. 1 of a recent trans-ancestry meta-analysis by Horowitz et al.[27] We note that a subset of AncestryDNA survey respondents overlap those included in the large meta-analyses conducted by HGI and Horowitz et al., and thus, replication in our study is not completely independent (Supplementary Fig. 2). All 12 SNPs in the final list are independent of one another ($r^2 < 0.05$ in the EUR, AA and LAT cohorts) and represent ten positionally distinct loci (>500 kb apart). One of the ten loci encompasses three independent SNPs that span a 54-kb region near *SLC6A20/LZTFL1* on chr3. For these 12 index SNPs, we extracted corresponding summary statistics from the trans-ancestry meta-analysis for each phenotype. We computed two-sided $-\log_{10}(P \text{ value})$ from the trans-ancestry meta-analysis, setting any trans-ancestry $P > 0.05$ or with inconsistent directions of effect compared to the previous study equal to zero. From the resulting matrix of $-\log_{10}(P \text{ values})$, we generated a heatmap with R package pheatmap version 1.0.12 and used hierarchical clustering to order the phenotype rows and the SNP columns in an unsupervised fashion. We generated a forest plot of corresponding minor allele effect estimates and 95% confidence intervals with the R package ggforestplot version 0.1.0.

Analysis of enrichment of protective effects. For each of the eight phenotypes, we identified all SNPs that were associated at different P-value significance thresholds (α), where $\alpha \in \{0.01, 0.001, 1 \times 10^{-4}, 1 \times 10^{-5}\}$. For each suggestive association, we designated the SNP with the lowest trans-ancestry P value within a 500-kb window as the index SNP. We recorded the total number of associated index SNPs and the proportion of index SNPs for which the minor allele was associated with a protective direction of effect. Each index SNP was also binned into one of six MAF bins: [0.01,0.05], (0.05,0.1], (0.1,0.2], (0.2,0.3] (0.3,0.4], (0.4,0.5].

To test for enrichment of protective or risk effects for each phenotype and α threshold, we used a Cochran–Mantel–Haenszel (CMH) test of two proportions with the base R (version 3.6.0) function mantelhaen.test. We compared the proportion of protective index SNPs for each phenotype to the proportion of protective index SNPs identified in the remaining seven phenotypes at the same α. MAF bins were used as strata. We considered a Bonferroni-corrected two-sided CMH P value $< 0.05/(8 \text{ phenotypes} \times 4\ \alpha \text{ thresholds}) = 0.00156$ evidence for enrichment of risk or protective effects relative to the other phenotypes. We plotted the MAF-adjusted CMH ORs in Fig. 3b, where positive numbers represent enrichment for protective effects. In cases where the protective enrichment OR was less than 1, we reported $-(1/\text{OR})$, so that lower numbers represent greater enrichment of risk effects.

Statistics and reproducibility. This is a cross-sectional genetic association study using extant genetic data from AncestryDNA customers and disease data collected via survey over an approximately 3.5-month period early in the COVID-19 pandemic. No statistical method was used to predetermine sample size. Individuals who could not be assigned to one of the three ancestry cohorts were excluded from analyses. For all close relative pairs, one randomly selected individual was excluded from analyses. The experiments were otherwise not randomized. This was an observational study from self-reported data, and as such, the investigators were not blinded to allocation during experiments and outcome assessment.

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

Data availability

This study replicates findings by large consortia, for which full summary statistics can be found at https://rgc-covid19.regeneron.com and https://www.covid19hg.org/results/. GWAS summary statistics for each cohort and each of the eight phenotypes will be available to qualified researchers upon request and approval by the data access committee through the European Genome-phenome Archive (https://ega-archive.org/studies/EGAS00001005099). The top 10,000 SNPs in the trans-ancestry meta-analysis for each of the eight phenotypes have also been deposited in the GWAS Catalog (*Positive/Negative*: GCST90094643, *Positive/Unscreened*: GCST90094644, *Hospitalized/Not_Hospitalized*: GCST90094645, *Hospitalized/Unscreened*: GCST90094646, *Exposed_Positive/Exposed_Negative*: GCST90094647, *Unscreened/Exposed_Negative*: GCST90094648, *Symptomatic/Paucisymptomatic*: GCST90094649, *Continuous_Severity_Score*: GCST90094650). AncestryDNA cannot make the individual-level data or summary data used in the reference panel to infer continental admixture proportions available to the academic community in light of our commitment to customer privacy.

Code availability

Code to reproduce the analyses presented here is available in the Ancestry public GitHub Repository (https://github.com/Ancestry/NatGenCOVIDpaper2022/; https://doi.org/10.5281/zenodo.5808281). Code to estimate continental admixture proportions and code to infer identity by descent cannot be made available, because this information is proprietary.

References

24. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobannk scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
25. Loh, P. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
26. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
27. Horowitz, J. E. et al. Common genetic variants identify therapeutic targets for COVID-19 and individuals at high risk of severe disease. Preprint at *medRxiv* https://doi.org/10.1101/2020.12.14.20248176 (2020).

Acknowledgements

We thank our AncestryDNA customers who made this study possible by voluntarily contributing information about their experience with COVID-19 through our survey. Without them, this work would not be possible. We additionally thank our collaborators at Regeneron Genetics Center and the COVID-19 Host Genetics Initiative for including us in ongoing meta-analyses aimed to improve understanding of COVID-19 infection susceptibility and severity. We acknowledge and thank J. Rhead for creating publication-ready figures.

Author contributions

G.H.L.R., R.P. and B.R. contributed equally to the manuscript. G.H.L.R. wrote the manuscript with substantial input from B.R., R.P., S.C.K. and D.S.P. D.S.P. defined ancestry cohorts. R.P. and G.H.L.R. conducted all GWASs and meta-analyses with support from D.S.P. B.R. conducted literature review. M.Z., D.S.P., D.A.T., S.C.K., M.P., M.G., L.R., A.H.B. and H.G. performed genotype imputation and data preparation. M.V.C. and K.A.R. designed the COVID-19 survey questionnaire, and G.H.L.R., S.C.K., M.V.C., K.A.R. and S.R.M. designed new phenotypes. N.B. and M.V.C. created the demographic table. C.S. compiled the appendix of survey questions. A.R.G., A.K.H.B. and H.G. facilitated forward progression of the manuscript and provided input and guidance. The AncestryDNA Science Team contributed to additional work, allowing for the completion of the COVID-19 research and manuscript. K.A.R. led the COVID-19 research and data teams. K.A.R., E.L.H. and C.A.B. provided project guidance. All authors contributed to and reviewed the final manuscript.

Competing interests

The authors declare competing financial interests. All authors are employed by Ancestry and may have equity in Ancestry.

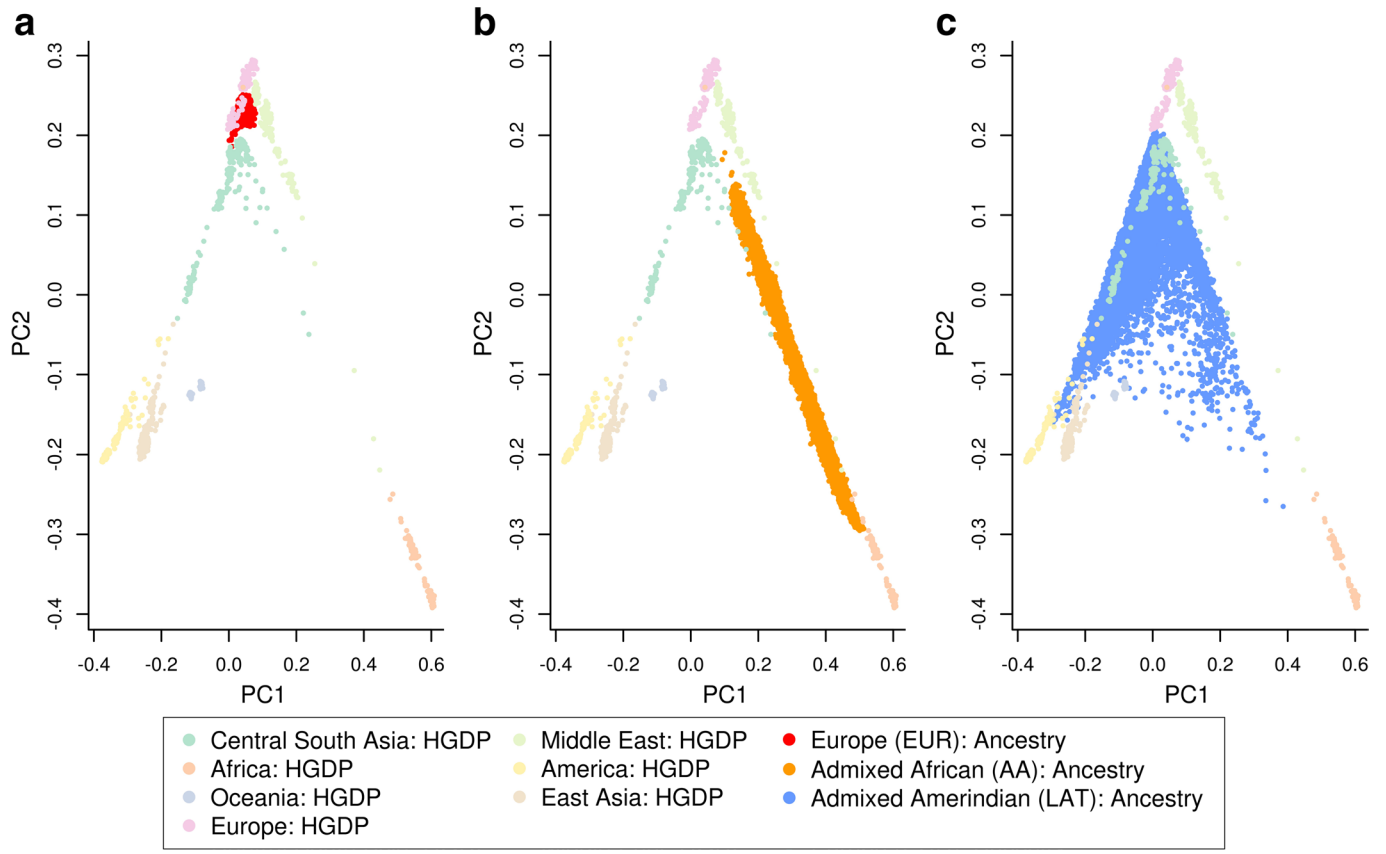
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-022-01042-x.

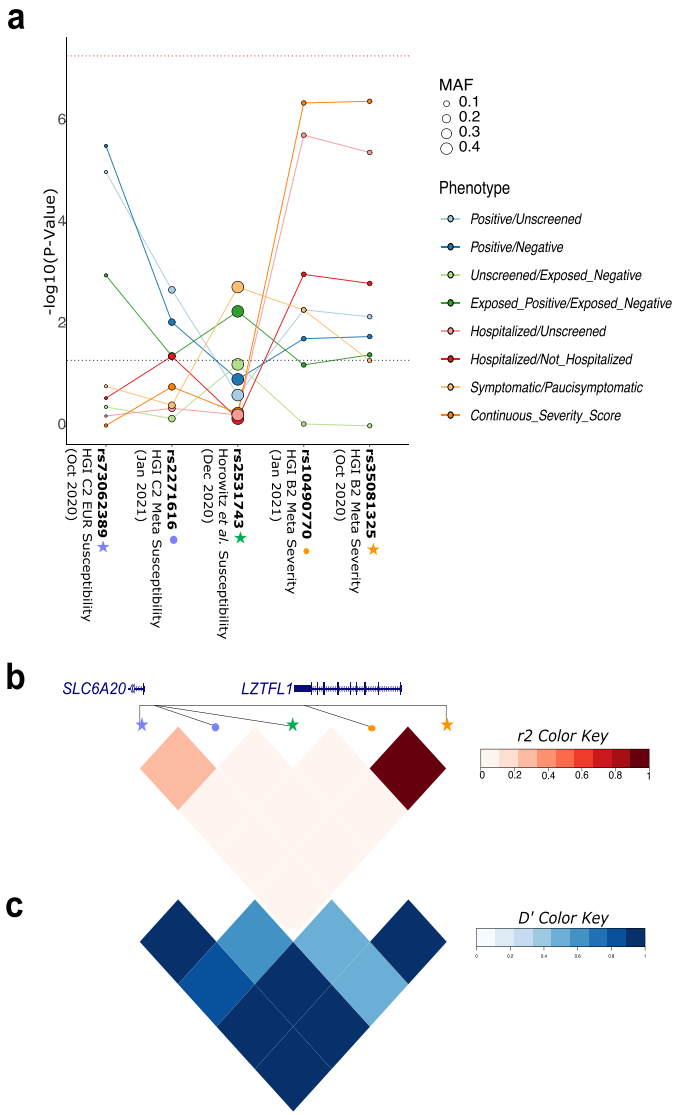
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01042-x.

Correspondence and requests for materials should be addressed to Kristin A. Rand.

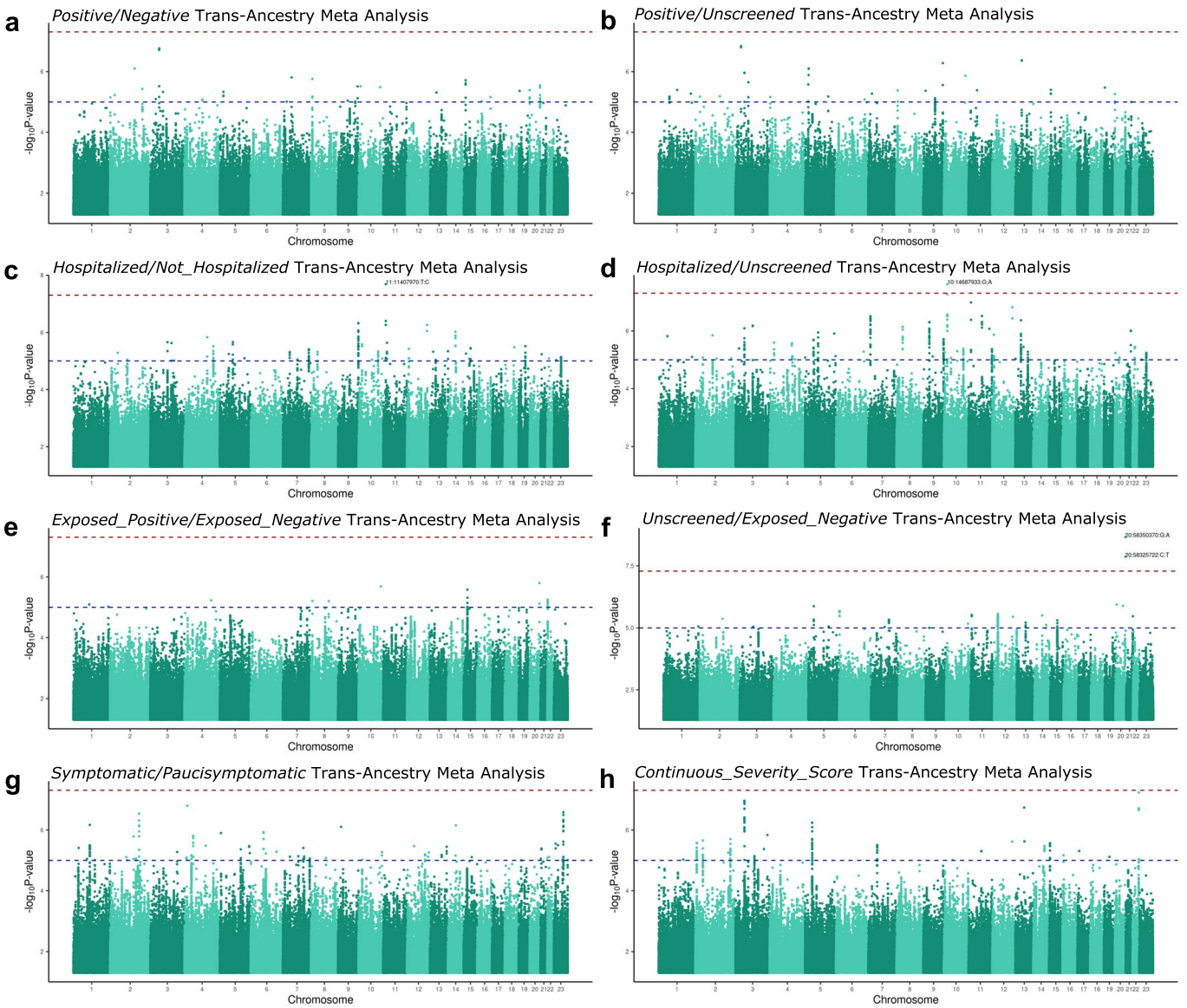
Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



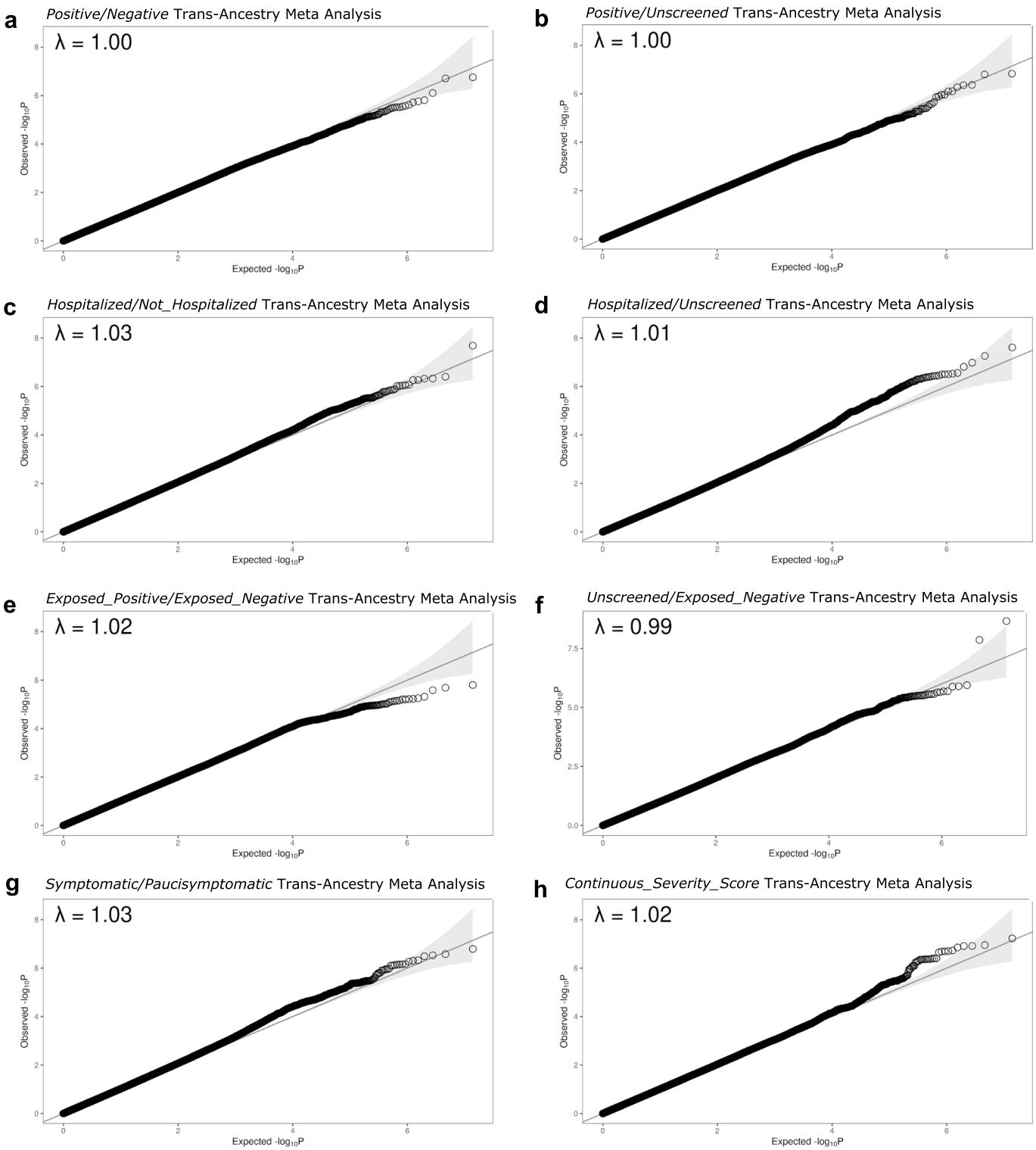
Extended Data Fig. 1 | Principal components plot of the three genetic ancestry cohorts. Principal-component (PC) plots of the survey cohort together with samples from the Human Genetic Diversity Project superpopulations, denoted 'HGDP'. The PC1 versus PC2 plots are shown for the following cohorts: (**a**) European (EUR; red), (**b**) Admixed African-European including 100% African ancestry (AA; orange), and (**c**) Admixed Amerindian including 100% Amerindian ancestry (LAT; blue).



Extended Data Fig. 2 | Association of Five Lead SNPs within the Chr3p21 Region. (**a**) The *y* axis represents -$\log_{10}$(*P*-values) for the association between the three SNPs presented in the main text (stars) plus two additional SNPs more recently reported as lead SNPs for COVID-19 severity or susceptibility by HGI (circles; see Supplementary Note). The three colors of the SNP annotation shapes represent three independent signals ($r^2 < 0.04$). The plotted points represent -$\log_{10}$(*P*-values) for the association with each phenotype in our analysis, where *P*-values are two-sided and based on an inverse-variance-weighted fixed-effects meta-analysis. The plotted points are colored by phenotype, and the point's size corresponds to the MAF from the combined EUR, LAT, AA meta-analysis cohort. The black horizontal dotted line represents $P = 0.05$ and the red horizontal dotted line represents $P = 5 \times 10^{-8}$. Each SNP's position and the genes within the chr3:45835417-45889921 (hg19) are shown below the main plot. For all five SNPs, we also show (**b**) pairwise r^2 and (**c**) pairwise *D'* calculated in the CEU population from 1000 Genomes Project.



Extended Data Fig. 3 | Manhattan plots for trans-ancestry meta-analyses of European (EUR), Admixed Amerindian (LAT), Admixed African-European (AA) cohorts in eight phenotypes. Chromosomal position is represented on the x axes and -$\log_{10}$(*P*-values) on the y axes. P-values are two-sided and based on an inverse-variance-weighted fixed-effects meta-analysis. Blue dashed lines are drawn at the suggestive significance threshold of $P=1\times10^{-5}$ and red dashed lines at the genome-wide significance threshold of $P=5\times10^{-8}$. The phenotypes are: (**a**) *Positive/Negative*, (**b**) *Positive/Unscreened*, (**c**) *Hospitalized/Not_Hospitalized*, (**d**) *Hospitalized/Unscreened*, (**e**) *Exposed_Positive/Exposed_Negative*, (**f**) *Unscreened/Exposed_Negative*, (**g**) *Symptomatic/Paucisymptomatic*, and (**h**) *Continuous_Severity_Score*.



Extended Data Fig. 4 | Quantile-quantile plots for trans-ancestry meta-analyses of European (EUR), Admixed Amerindian (LAT), Admixed African-European (AA) cohorts in eight phenotypes. Expected -$\log_{10}$(*P*-values) are represented on the x axes and observed -$\log_{10}$(*P*-values) on the y axes. Observed P-values are two-sided and based on an inverse-variance-weighted fixed-effects meta-analysis. Lambda (λ) values are genomic inflation factors. The phenotypes are: (**a**) *Positive/Negative*, (**b**) *Positive/Unscreened*, (**c**) *Hospitalized/Not_Hospitalized*, (**d**) *Hospitalized/Unscreened*, (**e**) *Exposed_Positive/Exposed_Negative*, (**f**) *Unscreened/Exposed_Negative*, (**g**) *Symptomatic/Paucisymptomatic*, and (**h**) *Continuous_Severity_Score*.

nature portfolio

Corresponding author(s): Kristin A. Rand

Last updated by author(s): 2/16/2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Self-reported COVID 19 outcomes were collected through the Personal Discoveries Project®, a survey platform available to AncestryDNA customers via the web and mobile applications. The COVID 19 survey ranged from 39-71 questions, depending on the initial COVID 19 test result reported. Individuals that had access to the survey were existing AncestryDNA customers with microarray genotype data.

Data analysis	All analysis was performed with previously published software. Genotype phasing was conducted with Eagle version 2.4.1 and imputation was conducted with Minimac4 version 1.0.1. Principal Components Analysis (PCA) was conducted with FlashPCA 2.0. Genome-wide association studies (GWAS) were conducted with PLINK version 2.0. GWAS meta-analyses were conducted with METAL (version released 25 March 2011). Hierarchical clustering of -log10(replication P-values) was conducted with the R package pheatmap(). Protective effect enrichment was performed at 4 different trans-ancestry P-value thresholds with a Cochran–Mantel–Haenszel test (CMH) of two proportions using the R package mantelhaen.test(). Power analysis was performed with the Purcell Power Calculator for case-control discrete traits (https://zzz.bwh.harvard.edu/gpc/cc2.html, created 24.Oct.2008). PCs were calculated to include in the association studies to control residual population structure and were computed using FlashPCA 2.0.2. Samples were imputed to the Haplotype Reference Consortium (HRC) reference panel version 1.1, which consists of 27,165 total individuals and 36 million variants. The HRC reference panel does not include indels; consequently, indels are not present in the results of our analyses. We determined best-guess haplotypes with Eagle version 2.4.1 and performed imputation with Minimac4 version 1.0.1. For each individual population (EUR female, EUR male, LAT, and AA) and each of the 8 phenotypes, we conducted a GWAS assuming an additive genetic model with PLINK2.0. Orthogonal polynomials were used to eliminate collinearity between age and age2 and were calculated in R version 3.6.0 with base function poly(age, degree=2). We additionally used PLINK2.0 to remove variants with Minimac4 imputation quality R2<0.3 or with MAF<0.01. For each phenotype, we additionally performed a trans-ancestry meta-analysis of the sex-combined EUR, AA, and LAT summary statistics, again using fixed-effect inverse-variance weighting implemented in METAL (version released 25 March 2011). We generated a heatmap of -log10(P-values) with R package pheatmap version 1.0.12, and used hierarchical clustering to order the phenotype rows and the SNP columns in an unsupervised fashion. We generated a forest plot of corresponding minor allele effect estimates and 95% confidence intervals with the R package ggforestplot version 0.1.0. A DOI-minted repository for analyses presented here is available at https://doi.org/10.5281/zenodo.5808281.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

This study replicates findings by large consortia, for which full summary statistics can be found at https://rgc-covid19.regeneron.com and https://www.covid19hg.org/results/. GWAS summary statistics for each cohort and each of the eight phenotypes will be available to qualified researchers upon request and approval by the data access committee through the European Genome-Phenome Archive (EGA) (https://ega-archive.org/studies/EGAS00001005099). The top 10,000 SNPs in the trans-ancestry meta-analysis for each of the eight phenotypes have also been deposited in the GWAS Catalog (Positive/Negative: GCST90094643, Positive/Unscreened: GCST90094644, Hospitalized/Not_Hospitalized: GCST90094645, Hospitalized/Unscreened: GCST90094646, Exposed_Positive/Exposed_Negative: GCST90094647, Unscreened/Exposed_Negative: GCST90094648, Symptomatic/Paucisymptomatic: GCST90094649, Continuous_Severity_Score: GCST90094650). AncestryDNA cannot make the individual-level data or summary data used in the reference panel to infer continental admixture proportions available to the academic community in light of our commitment to customer privacy.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based on the number of respondents to our online survey and we judged GWAS sample sizes appropriate based on a power analysis. Power for nominal replication (P<0.05) for all 7 binary phenotypes was >79% based assuming a causal variant relative risk (RR) of 1.25 minor allele frequency (MAF)=0.20. All sample sizes are shown in Supplementary Table 2.
Data exclusions	Individuals were excluded from GWAS for the following reasons: (1) if they did not answer a survey question required to create the phenotype, (2) if they were close relatives of another individual that participated, (3) if they did not meet the requirements to be included in one of the three ancestry cohorts (European ancestry/EUR, an Admixed Amerindian/LAT, or Admixed African/AA)), or (4) due to the large scale of the data, a subset of "unscreened" EUR individuals that did not report a COVID-19 test result were randomly excluded after May 28, 2020 to reduce the cost of whole genome imputation and data storage while maintaining high GWAS power.
Replication	To explore how known COVID-19 risk loci associate with these different phenotype definitions, we investigated 12 independent SNPs (r2<0.05) that were identified in at least one of two recent, large, COVID-19 meta-analyses: the October 2020 data release from the COVID-19 Host Genetics Initiative (HGI) or Horowitz et al. (as shown in Supplementary Table 4 of the paper). We assessed association of these 12 SNPs with all eight phenotypes, defining evidence of replication as an AncestryDNA trans-ancestry P<0.05 and consistent direction of effect with the prior study. Eight of 12 SNPs replicated in at least one of our eight phenotypes.
Randomization	This is an observational study, for which randomization is not applicable; however, Individuals in the EUR cohort were randomly assigned to

Randomization	be imputed if they were unscreened for COVID-19 and imputed EUR individuals were randomly assigned to the EUR independent replication cohort or the EUR discovery cohort after May 28, 2020, in all cases using R function sample().
Blinding	This is an observational study, for which blinding is not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	To participate in the COVID-19 survey, participants must meet the following criteria: they must be 18 years of age or older, a resident of the United States, be an existing AncestryDNA customer who has consented to participate in research and be able to complete a short survey. The survey is designed to assess self-reported COVID-19 positivity and severity, as well as susceptibility and known risk factors including community exposure and known contacts with individuals diagnosed with COVID-19. Related demographic information is presented in Table 1 and Supplementary Table 1.
Recruitment	To perform genetic studies of COVID-19, we conducted a comprehensive, 50+ question survey of AncestryDNA customers (Supplementary Appendix). All subjects provided informed consent before being allowed access to the survey. The survey assessed exposure, risk factors, symptomatology, and demographic information, as summarized in Table 1. An expanded set of cohort demographics are available in Supplementary Table 1. We collected 736,723 COVID-19 survey responses between April and August 2020. Given the nature of self-reported data through our survey engine, one would assume a participant to be healthy enough to voluntarily participate in the survey, thus our data may represent a more mild COVID-19 phenotype on average, with lower representation of severe COVID-19 outcomes. We have utilized this fact to better study more mild phenotypes and believe it can be used to better understand protective effects associated with disease. Participants were self-selected and therefore may not be representative of the larger United States population.
Ethics oversight	All data for this research project were from subjects who provided prior informed consent to participate in AncestryDNA's Human Diversity Project, as reviewed and approved by our external institutional review board, Advarra (formerly Quorum). All data were de-identified prior to use.

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