

1 **Novel COVID-19 phenotype definitions reveal phenotypically distinct patterns**
2 **of genetic association and protective effects**

3
4 Genevieve H.L. Roberts^{†1}, Raghavendran Partha^{†2}, Brooke Rhead^{†2}, Spencer C. Knight², Danny
5 S. Park², Marie V. Coignet², Miao Zhang², Nathan Berkowitz², David A. Turrisini², Michael
6 Gaddis², Shannon R. McCurdy², Milos Pavlovic¹, Luong Ruiz², AncestryDNA Science Team^{1,2},
7 Asher K. Haug Baltzell¹, Harendra Guturu², Ahna R. Girshick², Kristin A. Rand², Eurie L.
8 Hong², Catherine A. Ball²

9 From:

10 1. AncestryDNA, 1300 West Traverse Parkway, Lehi, UT, 84043
11 2. AncestryDNA, 153 Townsend St, Suite 800, San Francisco, CA, 94107

12
13 †These authors contributed equally.

14 Correspondence to: Kristin A. Rand, kristinmuench@gmail.com

15 INTRODUCTION PARAGRAPH

16 Multiple large COVID-19 genome-wide association studies (GWAS) have identified
17 reproducible genetic associations indicating that some infection susceptibility and severity risk is
18 heritable.¹⁻⁵ Most of these studies ascertained COVID-19 cases in medical clinics and hospitals,
19 which can lead to an overrepresentation of cases with severe outcomes, such as hospitalization,
20 intensive care unit admission, or ventilation. Here, we demonstrate the utility and validity of
21 deep phenotyping with self-reported outcomes in a population with a large proportion of mild
22 and subclinical cases. Using these data, we defined eight different phenotypes related to
23 COVID-19 outcomes: four that align with previously studied COVID-19 definitions and four
24 novel definitions that focus on susceptibility given exposure, mild clinical manifestations, and an
25 aggregate score of symptom severity. We assessed replication of 13 previously identified
26 COVID-19 genetic associations with all eight phenotypes and found distinct patterns of
27 association, most notably related to the chr3/*SLC6A20/LZTFL1* and chr9/*ABO* regions. We then
28 performed a discovery GWAS, which suggested some novel phenotypes may better capture
29 protective associations and also identified a novel association in chr11/*GALNT18* that
30 reproduced in two fully independent populations.

31 **MAIN TEXT**

32 To perform genetic studies of COVID-19, we conducted a comprehensive, 50+ question survey
33 of AncestryDNA customers that assessed exposure, risk factors, symptomatology, and
34 demographic information (**Supplementary Figure 1; Supplementary Table 1**). We collected
35 over 700,000 COVID-19 survey responses between April and August 2020 and used them to
36 develop an expanded repertoire of phenotypes to investigate. In total, we defined eight
37 COVID-19 phenotypes, summarized in **Table 1**. Four phenotypes were intended to mirror
38 susceptibility or severity phenotype definitions from other large COVID-19 GWAS^{2,3} and four
39 are novel. We hypothesized that novel phenotype definitions focusing on mild outcomes or
40 absence of infection despite a strong exposure may be better suited to detecting protective
41 genetic associations than traditional phenotypes.

42
43 Susceptibility to infection is difficult to measure because contracting the virus depends on
44 exposure. We therefore designed two novel susceptibility phenotypes that focus on respondents
45 with a known, strong exposure to the virus—those who had “household exposure.” The
46 positivity rate among respondents that reported a housemate with confirmed COVID-19 was
47 approximately 65%, the highest positivity rate for any exposure we assessed. The
48 *Exposed_Positive/Exposed_Negative* phenotype compared those with a household exposure that
49 tested positive to those with a household exposure that tested negative, and
50 *Unscreened/Exposed_Negative* focused on protection from infection by comparing those with a
51 household exposure that tested negative to a large sample of unscreened controls. We also
52 defined two novel severity phenotypes: *Symptomatic/Paucisymptomatic*, which compares cases
53 with symptomatic infections to those with very mild or asymptomatic infections, and

54 *Continuous_Severity_Score* which unifies asymptomatic and severely ill COVID-19 patients.
55 The *Continuous_Severity_Score* aggregates responses from nine survey fields. Lower scores
56 correspond to lower symptom severity, while higher correspond to increased symptom severity
57 and elevated hospitalization rates (**Figure 1**). Sample sizes for each phenotype are presented in
58 **Supplementary Table 2**. For all eight phenotypes, cases corresponded to higher risk of
59 susceptibility or severity so that all positive SNP effect estimates ($\hat{\beta}_{SNP}$) can be interpreted as
60 “risk” and all negative $\hat{\beta}_{SNP}$ can be interpreted as “protective.”
61
62 Our first goal was to explore how known COVID-19 risk loci associate with the different
63 phenotype definitions. To accomplish this, we identified 13 independent SNPs ($r^2 < 0.05$) that
64 achieved genome-wide significance in at least one of two recent, large, COVID-19 meta-
65 analyses: the October 2020 data release from the COVID-19 Host Genetics Initiative (HGI) or
66 Horowitz *et al.* (**Supplementary Table 3**). We assessed association of these 13 SNPs with all
67 eight phenotypes in a trans-ancestry meta-analysis of European (EUR), Admixed Amerindian
68 (LAT), and Admixed African-European (AA) cohorts (**Supplementary Figure 2**). We
69 considered a trans-ancestry P -value of < 0.05 and consistent direction of $\hat{\beta}_{SNP}$ with the prior
70 study evidence of replication (**Supplementary Table 4**). We note that a small percentage of
71 research participants in our study overlaps prior studies, quantified in **Supplementary Figure 3**.
72
73 Replication results are visualized in **Figure 2**. Ten of 13 SNPs replicated in at least one of our
74 phenotypes. This result demonstrates that our phenotypes, which are based on self-reported
75 outcomes, strongly recapitulate the same associations previously found by clinical phenotyping.
76 Hierarchical clustering of the replication P -values revealed two unique clusters of phenotype-

77 locus pairs: three *severity* phenotypes produced a similar pattern of replication
78 (*Hospitalized/Not_Hospitalized*, *Hospitalized/Unscreened*, *Continuous_Severity_Score*) and
79 three *susceptibility* phenotypes produced a similar pattern of replication (*Positive/Negative*,
80 *Positive/Unscreened*, *Exposed_Positive/Exposed_Negative*). Phenotypes in these clusters are
81 likely capturing similar genetic associations; however, the strength of associations differ,
82 suggesting that some phenotype definitions are more powerful than others. The two remaining
83 novel phenotypes (*Symptomatic/Paucisymptomatic*, *Unscreened/Exposed_Negative*) replicated
84 the 13 SNPs poorly and may capture different genetic associations, warranting further
85 investigation.

86

87 The patterns of locus replication are of special interest, particularly in chr3 and chr9 regions.
88 There are three independent signals in a 52Kb region on chr3 near a cluster of immune genes
89 including *LZTFL1* and *SLC6A20*. The main HGI *severity* study (“ANA_B2”) identified
90 rs35081325, which is strongly associated with the *severity* cluster of phenotypes. Thus,
91 rs35081325 appears to consistently associate with increased risk of infection severity. By
92 contrast, rs73062389 was identified in the main HGI *susceptibility* study (“ANA_C2”) and is
93 strongly associated with the *susceptibility* cluster of phenotypes. Furthermore, rs73062389 is not
94 associated with *any* of our severity cluster phenotypes and thus seems to specifically confer
95 increased *susceptibility* risk. Finally, rs2531743, a novel signal recently discovered by Horowitz
96 *et al.* in an analysis of severity, associated with only two phenotypes in our study:
97 *Symptomatic/Paucisymptomatic* and *Exposed_Positive/Exposed_Negative*. Unlike the other chr3
98 signals, the minor allele of rs2531743 is associated in the protective direction of effect. Thus, all
99 three signals in this region associate with a totally distinct set of phenotypes.

100

101 Associations near *ABO*, the gene that determines blood type, have also been observed in multiple
102 COVID-19 GWAS—somewhat inconsistently with severity phenotypes and more consistently
103 with susceptibility phenotypes (**Supplementary Table 5**). The lead *ABO* SNP, rs505922,
104 replicated in all four susceptibility phenotypes plus one severity phenotype. The only severity
105 phenotype associated with the *ABO* SNP was *Hospitalized/Unscreened*, which utilized a large
106 number of unscreened controls. We speculate that unscreened controls induce susceptibility
107 associations because hospitalized cases *must* be susceptible to infection, but the unscreened
108 control group *may or may not* be susceptible, and thus this phenotype simultaneously captures
109 aspects of both susceptibility and severity.

110

111 Our second goal was to discover novel phenotype-locus associations; we therefore conducted a
112 discovery GWAS for all eight phenotypes. Due to the novelty of the phenotypes and the
113 difficulty in obtaining a truly independent COVID-19 replication cohort, we opted to conduct the
114 discovery GWAS in the same EUR cohort used in the above trans-ancestry meta-analysis, and
115 dedicate a smaller, fully independent EUR cohort, the LAT cohort, and the AA cohort to
116 determining whether any newly identified phenotype-locus associations reproduce.

117

118 No phenotype-locus association pairs surpassed a conservative Bonferroni-corrected significance
119 threshold of discovery $P < 6.25 \times 10^{-9}$, but we examined associations that reached a suggestive
120 significance threshold of $P < 1 \times 10^{-5}$ to look for trends. In total, we identified 297 suggestive
121 phenotype-locus association pairs (**Supplementary Table 6**). Strikingly, minor alleles
122 suggestively associated with three novel phenotypes (*Exposed_Positive/Exposed_Negative*,

123 *Unscreened/Exposed_Negative, Symptomatic/Paucisymptomatic*) were nearly always associated
124 with a *protective* direction of effect, whereas for all previously studied phenotypes, the minor
125 allele was *nearly always* associated in the *risk* direction (**Figure 3**). This finding supports our
126 hypothesis that the novel phenotype definitions that focus on mild outcomes or absence of
127 infection despite a strong exposure may be better suited to detecting protective genetic
128 associations than traditional phenotypes.

129
130 Overall, we observed low rates of replication among the 297 phenotype-locus association pairs
131 (mean replication rate=3.7%; **Supplementary Table 7**) and low correlation of $\hat{\beta}_{SNP}$ across the
132 three independent populations (mean $\hat{\beta}_{SNP}$ Pearson $R=-0.23$; **Supplementary Figure 4**;
133 **Supplementary Table 7**). This result suggests that the independent replication cohorts had
134 insufficient power or that many of the suggestive phenotype-locus pairs simply represent false-
135 positive associations. Interestingly, however, two novel phenotypes generally had positive $\hat{\beta}_{SNP}$
136 correlations across independent populations: *Continuous_Severity_Score* and
137 *Exposed_Positive/Exposed_Negative* (**Supplementary Figure 4b-c**), suggesting that these
138 phenotypes might yield reproducible associations as replication cohort sample sizes grow larger.
139 There were also 15 phenotype-locus association pairs that reproduced in one independent
140 population, and one that replicated in two independent populations (**Supplementary Table 8**).
141 The phenotype-locus association that replicated in two fully independent populations was
142 *Hospitalized/Not_Hospitalized* with rs55673936 (**Figure 4**). This SNP is an intronic variant on
143 chr11 in the gene *GALNT18*. Interestingly, another SNP within *GALNT18* was previously
144 reported as associated with an increased response to Tocilizumab⁶, an IL-6 blocking monoclonal
145 antibody that has been tested in multiple clinical trials for treatment of COVID-19, albeit with

146 mixed preliminary success.⁷⁻⁹ Nonetheless, this novel association with *GALNT18* complements
147 findings by other genetic studies that point to modulation of the IL-6 pathway as a potential
148 strategy to ameliorate severe COVID-19 in some people.¹⁰

149
150 In summary, we explored genetic association with eight different COVID-19 phenotype
151 definitions, four of which have not yet been explored. We find that 10 of 13 previously identified
152 COVID-19 genetic signals associate with at least one of the eight phenotype definitions. This
153 strong replication of loci identified by clinically ascertained studies confirms that phenotyping
154 based on well-designed self-report studies is valid. Some of these replicated genetic signals
155 clearly associate more with severity phenotypes and others associate more with susceptibility
156 phenotypes, suggesting that heterogeneity in ascertainment and different case/control definitions
157 likely underlies inconsistent associations, for instance *ABO*. Our findings also show that all three
158 previously identified signals in the chr3 *LZTFL1/SLC6A20* region associate with a different set
159 of phenotypes, suggesting that variation in this region modulates multiple aspects of COVID-19
160 susceptibility and severity and thus is extremely important. In our discovery analysis, we
161 identified a novel association with rs55673936, a *GALNT18* intron variant that reproduced in
162 multiple independent populations. Whereas other groups with primary ascertainment at medical
163 clinics are better equipped to study severe outcomes, our self-reported dataset allows a
164 complementary analysis of more granular phenotypes in a population enriched for mild
165 outcomes. We find promising evidence that exploring new phenotypes in this unique population
166 will yield novel genetic associations, particularly those that confer protection against the novel
167 coronavirus.

168 **ONLINE METHODS**

169 **Ethics statement**

170 All data for this research project were from subjects who provided prior informed consent to
171 participate in AncestryDNA's Human Diversity Project, as reviewed and approved by our
172 external institutional review board, Advarra (formerly Quorum). All data were de-identified prior
173 to use.

174

175 **Study population**

176 Self-reported COVID-19 outcomes were collected through the Personal Discoveries Project®, a
177 survey platform available to AncestryDNA customers via the web and mobile applications. The
178 COVID-19 survey ranged from 39-71 questions, depending on the initial COVID-19 test result
179 reported. **Supplementary Figure 1** describes the flow of the topics assessed in each section of
180 the survey. Analyses presented here were performed with data collected between April 22-
181 August 3, 2020.

182

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

189

190

191 **Binary Phenotype Definitions**

192 In total, we assessed eight phenotypes, which are summarized in **Table 1**. Key definitions
193 include testing positive or negative, hospitalization, asymptomatic cases, and housemate
194 exposure. COVID-19 positivity or negativity was assessed by the question “Have you been swab
195 tested for COVID-19, commonly referred to as coronavirus?”. Hospitalization due to COVID-19
196 illness was used as one binary measure of severity, and was assessed with the question, “Were
197 you hospitalized due to these symptoms?”. Asymptomatic cases were defined as those that were
198 positive for COVID-19 and either answered “No” do the question “Did you experience
199 symptoms as a result of your condition?” or answered either "None", "Very mild", or "Mild" to
200 all 15 questions related to symptom severity. High exposure to COVID-19 was assessed through
201 having a positive housemate, assessed by the question, “Has someone in your household tested
202 positive for COVID-19?”.

203

204 **Continuous Severity Phenotype Creation**

205 A continuous severity score was derived by computing the first principal component across nine
206 survey fields related to COVID-19 clinical outcomes. Six of the nine questions were binary:
207 hospitalization, intensive care unit (ICU) admittance with oxygen, ICU admittance with
208 ventilation, septic shock, respiratory failure, and organ failure due to COVID-19. Binary
209 responses were encoded as 0 for “No” and 1 for “Yes”. Three symptom questions related to
210 shortness of breath, fever, and nausea/vomiting symptoms were encoded as a unit-scaled variable
211 based on the following mapping: 0=“None”, 0.2=“Very mild”, 0.4=“Mild”, 0.6=“Moderate”,
212 0.8=“Severe”, and 1.0=“Very Severe”. The three symptoms were chosen based on prior
213 literature indicating their positive association with COVID-19 hospitalization.¹¹ The following

214 assumptions were made to so that a score could be calculated for most participants who reported
215 a positive COVID-19 test:

- 216 • Participants who responded “No” to the question “Did you experience symptoms as a
217 result of your condition?” were not presented with additional questions regarding
218 symptomatology or hospitalization and thus were encoded as 0 for all individual
219 symptoms (shortness-of-breath, fever, nausea/vomiting), hospitalization, ICU admittance,
220 and severe complications due to COVID-19 illness.
- 221 • Participants who responded “No” to the question “Were you hospitalized due to these
222 symptoms?” were not presented any further questions regarding hospitalization and thus
223 were encoded as 0 for ICU admittance and supplemental oxygen.
- 224 • Participants who declined to answer a question about complications due to COVID-19
225 illness such as septic shock, respiratory failure, and organ failure were encoded as 0 for
226 those complications (<2% of all participants for whom continuous severity was scored).

227

228 **Genotyping**

229 Genotyping and quality control procedures have been previously described elsewhere.¹² Briefly,
230 customer genotype data for this study were generated using an Illumina genotyping array and
231 processed either with Illumina or with Quest/Athena Diagnostics. To ensure quality of each
232 dataset, a sample passes a number of quality control (QC) checks, which includes identifying
233 duplicate samples, removing individuals with a per-sample call rate <98%, and identifying
234 discrepancies between reported sex and genetically inferred sex. Samples that pass all quality-
235 control tests proceed to the analysis pipeline; samples that fail one or more tests must be
236 recollected or manually cleared for analysis by lab technicians. Array markers with per-variant

237 call rate <0.98 and array markers that had overall allele frequency differences of >0.10 between
238 any two array versions were additionally removed prior to downstream analyses.

239

240 **Defining ancestry cohorts**

241 We defined three separate ancestry cohorts: a European ancestry group (EUR), an Admixed
242 Amerindian ancestry (LAT), and an Admixed African ancestry group (AA) (**Supplementary**
243 **Figure 2**). We assigned COVID-19 survey respondents to one of these ancestry groups with a
244 proprietary algorithm that estimates continental admixture proportions. Briefly, this algorithm
245 uses a hidden Markov model to estimate unphased diploid ancestry across the genome by
246 comparing haplotype structure to a reference panel. The reference panel consists of a
247 combination of AncestryDNA customers and publicly available datasets and is designed to
248 reflect global diversity. From our total cohort of 736,723 individuals who participated in the
249 COVID-19 survey as of August 3, 2020, 537,512 (73%) individuals were designated to the EUR
250 group, 22,464 (3%) to the AA group, and 47,301 (6%) to the (LAT) group, and the remainder
251 were not assigned to any ancestry group (**Supplementary Table 1**).

252

253 **Removal of related individuals**

254 AncestryDNA's identity-by-descent inference algorithm¹³ was used to estimate the relationship
255 between pairs of individuals. Pairs with estimated separation of fewer than four meioses were
256 considered close relatives. For all close relative pairs, one individual was randomly selected for
257 exclusion from our study. In total, we excluded ~8% (60,379) individuals from analysis due to
258 relatedness.

259

260 **Calculation of principal components (PCs)**

261 For each population described above, genetic PCs were calculated to include in the association
262 studies to control residual population structure and were computed using FlashPCA 2.0.¹⁴ Input
263 genotypes were linkage disequilibrium (LD)-pruned using PLINK 1.9 command `--indep-pairwise 100 5 0.2 --`
264 `maf 0.05 --geno 0.001.`

265

266 **Imputation**

267 Samples were imputed to the Haplotype Reference Consortium (HRC) reference panel version
268 1.1, which consists of 27,165 total individuals and 36 million variants. The HRC reference panel
269 does not include indels; consequently, indels are not present in the results of our analyses. We
270 determined best-guess haplotypes with Eagle¹⁵ version 2.4.1 and performed imputation with
271 Minimac4 version 1.0.1. We used 1,077,214 unique variants as input and 8,187,660 imputed
272 variants were retained in the final data set. For these variants, we conservatively restricted our
273 analyses to variants with minor allele frequency (MAF) >0.01 and Minimac4 $R^2 > 0.30$ using
274 imputed dosages for all variants regardless of whether they were originally genotyped.

275

276 **Discovery GWAS**

277 Discovery GWAS were conducted in EUR ancestry only. For discovery, we conducted sex-
278 stratified GWAS and meta analyze the results via inverse-variance weighting implemented with
279 METAL¹⁶(version released 25 March 2011). For each phenotype, a GWAS assuming an additive
280 genetic model was implemented with PLINK2.0. Imputed genotype dosage value was the
281 primary predictor. The following were included as fixed-effect covariates: PCs 1-25 (described
282 above), array platform, orthogonal age, and orthogonal age². Orthogonal polynomials were used

283 to eliminate collinearity between age and age² and were calculated in R version 3.6.0 with base
284 function poly(age, degree=2). We additionally used PLINK2.0 to remove variants with
285 Minimac4 imputation quality $R^2 < 0.3$ or with MAF < 0.01. The following PLINK2.0 flags were
286 used for each analysis:

```
287 --vcf [input imputed VCF] dosage=DS  
288 --psam [file that provides sex information]  
289 --covar [covariates file]  
290 --covar-name PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11, PC12, PC13, PC15, PC15, PC16, PC17, PC18, PC19, PC20, PC21, PC22, PC23, PC24, PC25,  
291 orthogonal_age, orthogonal_age2, platform  
292 --covar-variance-standardize  
293 --extract-if-info R2 >= 0.3  
294 --freq  
295 --glm  
296 --keep [list of unrelated Europeans]  
297 --keep-females OR keep-males  
298 --maf 0.01  
299 --pheno [1 of 8 phenotype files]  
300 --pheno-name [phenotype column name]
```

301
302 Unless otherwise noted, all EUR discovery variant effect estimates are adjusted for the 28
303 covariates described above. To establish significance, we implemented a stringent, Bonferroni-
304 corrected significance threshold by dividing the typical genome-wide significance threshold in
305 Europeans of $P < 5 \times 10^{-8}$ by the eight phenotypes, which results in $P < 6.25 \times 10^{-9}$. Suggestive
306 significance followed the definition used by the HGI consortium of $P < 1 \times 10^{-5}$.

307

308 **Independent Replication GWAS**

309 We used the AA, LAT, and a smaller, fully independent EUR cohort to replicate our findings.
310 We began reserving respondents for the replication EUR cohort at the conclusion of our previous
311 study¹² on May 28, 2020, and thus the replication EUR cohort is not representative of the full
312 period of survey collection. The AA and LAT cohorts were steadily collected throughout the
313 entire collection period that spanned April 22, 2020 to August 3, 2020. We conducted separate
314 GWAS for each of these three replication cohorts and for each of the eight phenotypes. The same
315 association procedure that was used for the discovery study was applied for replication cohorts,
316 except sample sizes for these cohorts were smaller (**Supplementary Table 2**), and thus a single
317 GWAS was conducted for males and females together with genetic sex included as a covariate.
318

319 **Trans-Ancestry Meta-Analysis**

320 For each phenotype, we additionally performed a trans-ancestry meta-analysis of the discovery
321 EUR cohort, AA, and LAT summary statistics, again using fixed-effect inverse-variance
322 weighting implemented in METAL. The replication EUR cohort was not included in the trans-
323 ancestry meta-analysis. These summary statistics were used to assess replication of the 13 loci
324 defined in the next section.

325

326 **Replication of 13 Independent SNPs from Previous Studies**

327 We manually curated a list of 13 independent SNPs that represent lead loci identified by either
328 HGI or Horowitz *et al.* Eight of the 13 SNPs were lead SNPs in HGI's most recent data release
329 (October 2020; without 23andMe data included). These eight SNPs were the most-associated
330 marker at any locus achieving $P < 5 \times 10^{-8}$ in the Hospitalization vs. Population ("ANA_B2") or
331 COVID-19(+) vs. Population ("ANA_C2"). The remaining five SNPs were selected from Figure

332 1 of a recent trans-ancestry meta-analysis.² We note that a subset of AncestryDNA survey
333 respondents overlap those included in the large meta analyses conducted by HGI and Horowitz *et*
334 *al.* and thus replication in our study is not completely independent (**Supplementary Figure 3**).
335 All 13 SNPs in the final list are independent of one another ($r^2 < 0.05$) and represent 11
336 positionally distinct loci (>500Kb apart). One of the 11 loci encompasses three independent
337 SNPs that span a 52Kb region near *SLC6A20/LZTFL1* on chr3. For these 13 index SNPs, we
338 extracted corresponding summary statistics from the trans-ancestry meta-analysis for each
339 phenotype. We computed the $-\log_{10}(P\text{-value})$ from the trans-ancestry meta-analysis, setting any
340 trans-ancestry $P > 0.05$ or with inconsistent directions of effect compared to the previous study
341 equal to zero. From the resulting matrix of $-\log_{10}(P\text{-values})$, we generated a heatmap with R
342 package pheatmap, and used hierarchical clustering to order the phenotype rows and the SNP
343 columns in an unsupervised fashion.

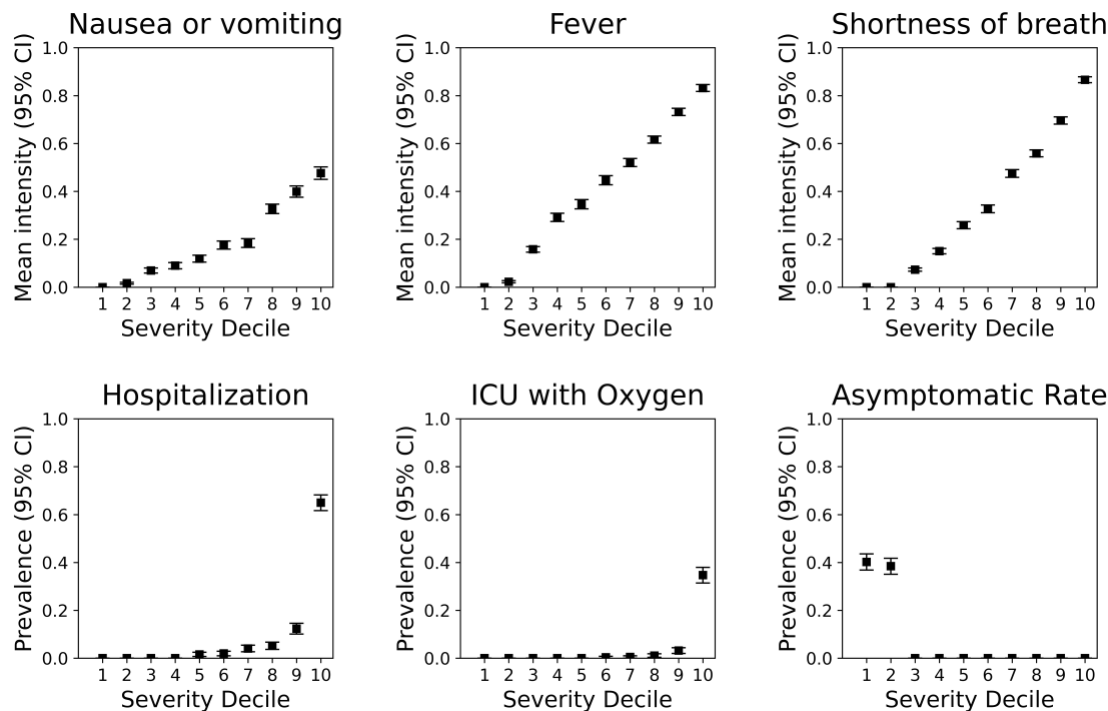
344

345 **Discovery of Novel Phenotype-Locus Associations**

346 Within the discovery EUR GWAS, we identified all loci that were suggestively associated
347 (discovery EUR $P < 1 \times 10^{-5}$) with any phenotype. For each of these suggestive associations, we
348 designated the SNP with the lowest EUR P -value within a 500kb window the index SNP. From
349 the resulting set of suggestive phenotype-locus association pairs, we determined whether the
350 association replicated in one or more independent GWAS (replication EUR, LAT, or AA). We
351 considered consistent direction of $\hat{\beta}_{SNP}$ and replication population $P < 0.05$ evidence of
352 replication. Some of the index SNPs selected in the discovery EUR GWAS were not analyzed in
353 the LAT and AA cohorts because the index SNP did not meet variant QC requirements
354 (MAF > 0.01 and Imputation $R^2 > 0.3$) in one or both of those populations. For five of such

355 phenotype-locus association pairs, there was another SNP that surpassed discovery EUR
356 $P < 1 \times 10^{-5}$ in the same region (<500Kb from the index SNP) *and* the alternative SNP was included
357 in both non-EUR GWAS, so we used this alternative SNP to assess replication in the non-EUR
358 cohorts. We also measured the Pearson correlation coefficient of $\hat{\beta}_{SNP}$ between the discovery
359 EUR study and each of the three independent replication cohorts.

360 **FIGURES**



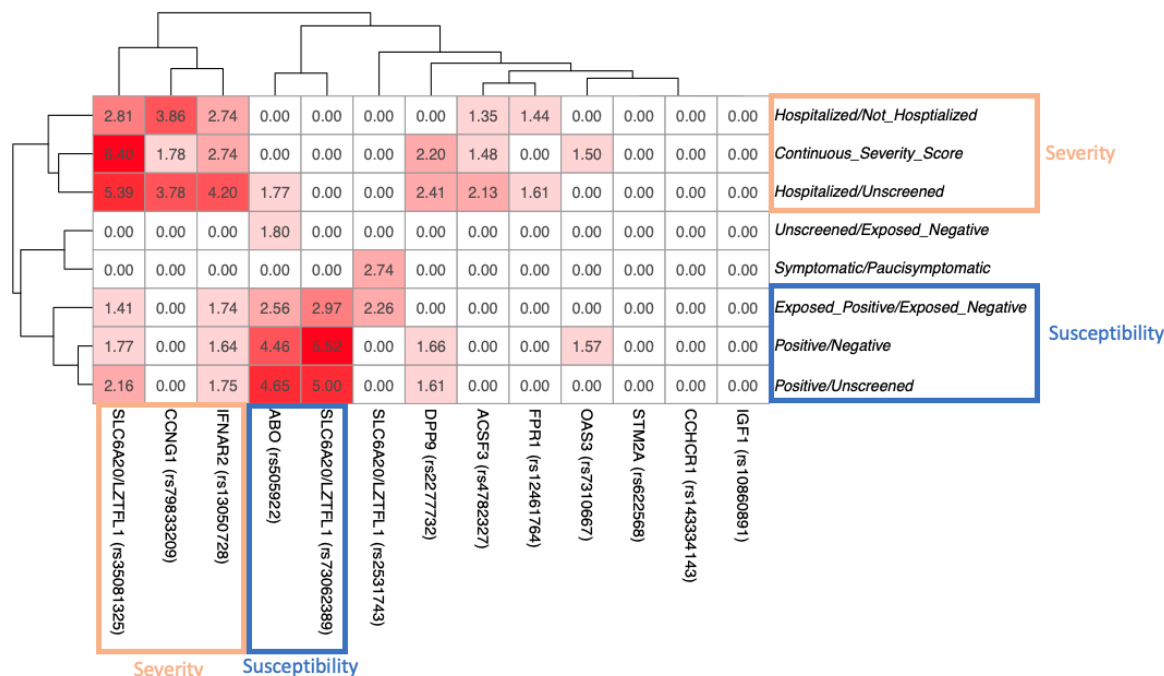
361
362

363 **Figure 1. COVID-19 Continuous Severity Score Captures Multiple Aspects of Symptom**

364 **Severity Among COVID-19(+) Individuals.** The continuous severity score was derived from
365 the first principal component across nine survey fields related to COVID-19 clinical outcomes,
366 including three symptoms, hospitalization, ICU admittance, and other severe complications due
367 to COVID-19 illness (see Methods). Plots reflect mean symptom severity (top three panels) or
368 prevalence (bottom three panels) for several fields as a function of ascending severity decile.

369 Symptom information was encoded as follows: 0=None, 0.2=Very Mild, 0.4=Mild,
370 0.6=Moderate, 0.8=Severe, and 1.0=Very Severe. A paucisymptomatic case corresponds to
371 reporting symptoms of mild intensity or less. Squares represent the estimate and vertical lines
372 represent the 95% confidence intervals for each estimate.

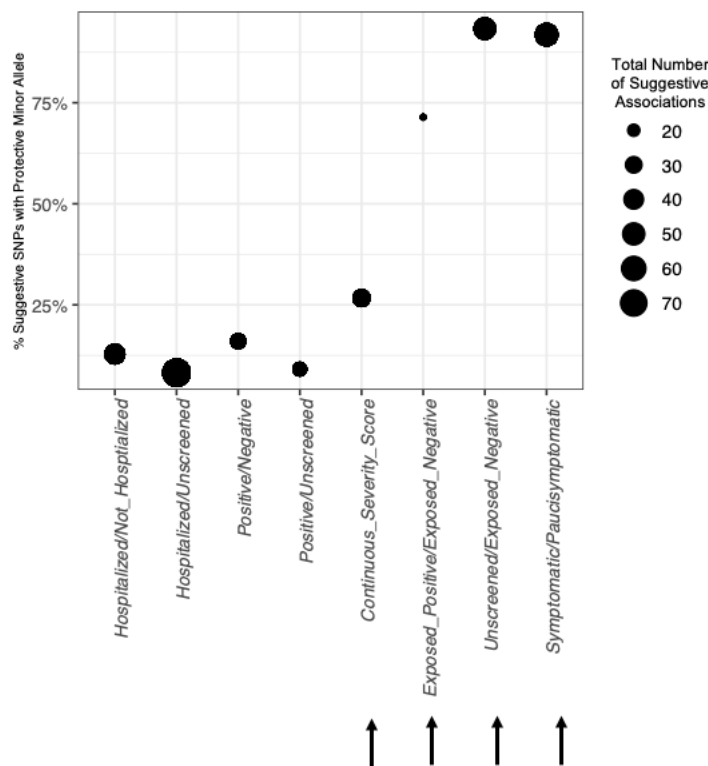
373



374

375 **Figure 2: Heatmap of replication at 13 lead SNPs identified by previous studies.** Each
 376 pairwise block represents the trans-ancestry meta-analysis $-\log_{10}(P\text{-value})$ for the association
 377 between one of the eight phenotypes we examined, and one of 13 loci previously identified by
 378 Horowitz *et al.* and/or HGI. Red blocks denote replication, with darker shades of red
 379 corresponding to lower trans-ancestry P -values in our analysis, and white blocks representing no
 380 association. All associations with trans-ancestry $P > 0.05$ or with inconsistent directions of effect
 381 relative to the previous study were forced to have $-\log_{10}(P\text{-value}) = 0$. SNP and phenotype labels
 382 were ordered by hierarchical clustering, with corresponding dendrograms shown on the top and
 383 left of the figure. Orange rectangles annotate phenotypes or loci that appear to associate more
 384 strongly with severity whereas blue rectangles annotate phenotypes or loci that appear to
 385 associate more strongly with susceptibility. Extended summary statistics for all associations in
 386 all studies are available in **Supplementary Table 4**.

387



388

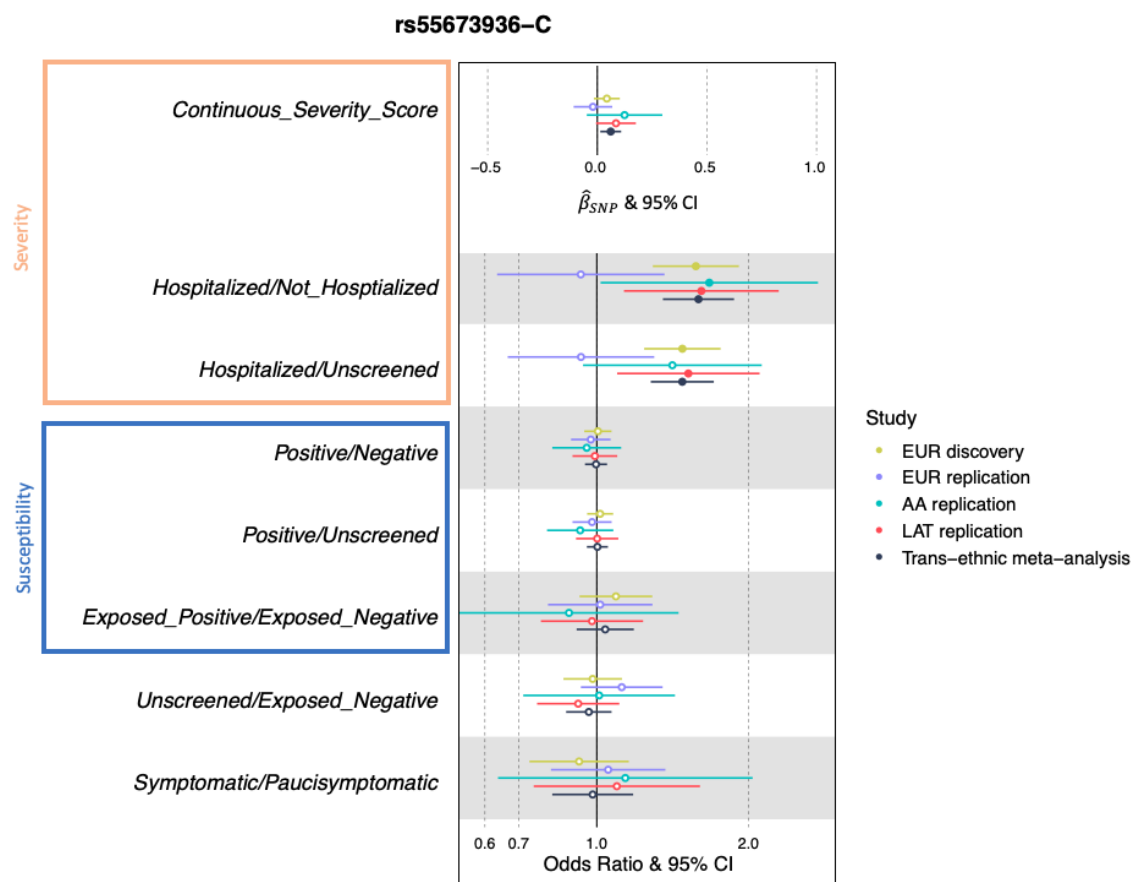
389 **Figure 3. Novel Phenotypes Detect More Associations with a Protective Minor Allele.** The

390 size of each point represents the total number of novel, suggestive SNPs (discovery EUR

391 $P < 1 \times 10^{-5}$) for each of the eight phenotypes. The y-axis position of each point shows the

392 percentage of suggestively associated SNPs for which the discovery EUR minor allele was in the

393 protective direction of effect. Arrows show the four novel phenotype definitions.



394
395

396 **Figure 4. Forest Plot of Novel Association with *GALNT18* intronic SNP, rs55673936-C, with**
 397 **the eight phenotypes.** Circles indicate effect estimates and horizontal lines represent 95%
 398 confidence intervals. *Continuous_Severity_Score* was the only continuous phenotype and
 399 therefore the reported effect estimate is the $\hat{\beta}_{SNP}$, which can be interpreted as severity score
 400 standard deviations from the mean per each copy of the “C” minor allele. For all other
 401 phenotypes, per-allele odds ratios are reported. Filled circles indicate $P < 0.05$. The orange
 402 rectangle annotates phenotypes in the severity cluster and the blue rectangle annotates the
 403 susceptibility cluster, with clusters defined in **Figure 2**.

404 **TABLES**

405 **Table 1. Summary of Eight Phenotype Definitions**

Phenotype Code ¹	Case Description ²	Control Description ²	Novelty	Type	Goal
<i>Positive/Negative</i>	COVID-19(+)	COVID-19(-)	Traditional	Susceptibility	Reproduce other studies
<i>Positive/Unscreened</i>	COVID-19(+)	Unscreened, but not known to be COVID-19(+)	Traditional	Susceptibility	Reproduce other studies
<i>Hospitalized/Not_Hospitalized</i>	COVID-19(+) and hospitalized	COVID-19+ and not hospitalized	Traditional	Severity	Reproduce other studies
<i>Hospitalized/Unscreened</i>	COVID-19(+) and hospitalized	Unscreened, but not known to be COVID-19(+)	Traditional	Severity	Reproduce other studies
<i>Exposed_Positive/Exposed_Negative</i>	COVID-19(+) and had a housemate with a confirmed COVID-19 diagnosis	COVID-19(-) and had a housemate with a confirmed COVID-19 diagnosis	Novel	Susceptibility	Study genetic susceptibility in individuals thought to have a strong exposure event
<i>Unscreened/Exposed_Negative</i>	Unscreened, but not known to be COVID-19(+)	COVID-19(-) and had a housemate with a confirmed COVID-19 diagnosis	Novel	Susceptibility	Study genetic protection from infection in individuals thought to have a strong exposure event
<i>Symptomatic/Paucisymptomatic</i>	COVID-19(+) and symptomatic	COVID-19(+) and asymptomatic or paucisymptomatic	Novel	Severity	Study genetic protection from severe outcomes if infected
<i>Continuous_Severity_Score</i> ³	COVID-19(+) score that combines nine different measures of COVID-19 symptom severity. Higher scores correspond to more severe outcomes.		Novel	Severity	Study genetic variants associated with both severe and mild outcomes simultaneously

1. Nomenclature for phenotype codes: Case_definition/Control_definition

2. Case or Control Descriptions in bold represent the minority group

3. The Continuous_Severity_Score phenotype is continuous, and thus there are no cases and controls. Instead, a score is computed for each person.

406

407 **AUTHOR CONTRIBUTIONS**

408 GHLR, RP, and BR contributed equally to the manuscript. GHLR wrote the manuscript with
409 substantial input from BR, RP, SCK, and DSP. DSP defined ancestry cohorts. RP and GHLR
410 conducted all GWAS and meta-analyses with support from DSP. BR conducted literature review.
411 MZ, DSP, DAT, SCK, MP, MG, LR, AHB, HG performed genotype imputation and data
412 preparation. MVC and KAR designed the COVID-19 survey questionnaire and GHLR, SCK,
413 MVC, KAR and SRM designed novel phenotypes. NB and MVC created the demographic table.
414 ARG, AHB, and HG facilitated forward progression of the manuscript and provided input and
415 guidance. The AncestryDNA Science Team contributed to additional work, allowing for the
416 completion of the COVID-19 research and manuscript. KAR led the COVID-19 research and
417 data teams. KAR, ELH, and CAB provided project guidance. All authors have contributed to and
418 reviewed the final manuscript.

419

420 **AncestryDNA Science Team:** Yambazi Banda, Ke Bi, Robert Burton, Marjan Champine, Ross
421 Curtis, Karen Delgado, Abby Drokhlyansky, Ashley Elrick, Cat Foo, Jialiang Gu, Heather
422 Harris, Shea King, Christine Maldonado, Evan McCartney-Melstad, Patty Miller, Keith Noto,
423 Jingwen Pei, Jenna Petersen, Chodon Sass, Alisa Sedghifar, Andrey Smelter, Sarah South, Barry
424 Starr, Cecily Vaughn, Yong Wang

425

426 **COMPETING INTERESTS**

427 The authors declare competing financial interests: authors affiliated with AncestryDNA are
428 employed by Ancestry and may have equity in Ancestry.

429

430 **ACKNOWLEDGEMENTS**

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

436

437 **DATA AVAILABILITY**

438 This study replicates findings by large consortia, for which full summary statistics can be found
439 at <https://rgc-covid19.regeneron.com> and <https://www.covid19hg.org/results/>.

440 REFERENCES

- 441 1. Ellinghaus, D. *et al.* Genomewide Association Study of Severe Covid-19 with
442 Respiratory Failure. *N Engl J Med* **383**, 1522-1534 (2020).
- 443 2. Horowitz, J.E. *et al.* Common genetic variants identify therapeutic targets for COVID-19
444 and individuals at high risk of severe disease. *medRxiv*, 2020.12.14.20248176 (2020).
- 445 3. The COVID-19 Human Genetics Initiative, COVID-19 HGI Results for Data Freeze 4
446 (October 2020).
- 447 4. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in Covid-19. *Nature*
448 (2020).
- 449 5. Shelton, J.F. *et al.* Trans-ethnic analysis reveals genetic and non-genetic associations
450 with COVID-19 susceptibility and severity. *medRxiv*, 2020.09.04.20188318 (2020)
- 451 6. Maldonado-Montoro, M., Canadas-Garre, M., Gonzalez-Utrilla, A., Plaza-Plaza, J.C. &
452 Calleja-Hernandez, M.Y. Genetic and clinical biomarkers of tocilizumab response in
453 patients with rheumatoid arthritis. *Pharmacol Res* **111**, 264-271 (2016).
- 454 7. The REMAP-CAP Investigators *et al.* Interleukin-6 Receptor Antagonists in Critically Ill
455 Patients with Covid-19 – Preliminary report. *medRxiv*, 2021.01.07.21249390 (2021)
- 456 8. Sanders, J.M., Monogue, M.L., Jodlowski, T.Z. & Cutrell, J.B. Pharmacologic
457 Treatments for Coronavirus Disease 2019 (COVID-19): A Review. *JAMA* **323**, 1824-
458 1836 (2020).
- 459 9. Takahashi, T., Luzum, J.A., Nicol, M.R. & Jacobson, P.A. Pharmacogenomics of
460 COVID-19 therapies. *NPJ Genom Med* **5**, 35 (2020).
- 461 10. Larsson, S.C., Burgess, S. & Gill, D. Genetically proxied interleukin-6 receptor
462 inhibition: opposing associations with COVID-19 and pneumonia. *Eur Respir J* (2020).
- 463 11. Knight, S.C. *et al.* COVID-19 Susceptibility and Severity Risks in a Survey of Over
464 500,000 People. *medRxiv*, 2020.10.08.20209593 (2020).
- 465 12. Roberts, G.H.L. *et al.* AncestryDNA COVID-19 Host Genetic Study Identifies Three
466 Novel Loci. *medRxiv*, 2020.10.06.20205864 (2020).
- 467 13. Ball, C.A. *et al.* AncestryDNA Matching White Paper: Discovering genetic matches
468 across a massive, expanding genetic database. (2020).
- 469 14. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of
470 Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776-2778 (2017).
- 471 15. Loh, P.R. Eagle v2.4.1 user manual. (2018).
- 472 16. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of
473 genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
- 474
- 475