

# 1 Novel risk loci for COVID-19 hospitalization among 2 admixed American populations

3 Ángel Carracedo on behalf of Spanish COalition to Unlock Research on host GENetics on  
4 COVID-19 (SCOURGE)

5 Email address for correspondence: [angel.carracedo@usc.es](mailto:angel.carracedo@usc.es)

6 Author list:

7 [https://docs.google.com/document/d/1gJVHCOM59Ycz6BduHfpyOOEmxIXcBXr/edit?usp=sharing&oid=117105](https://docs.google.com/document/d/1gJVHCOM59Ycz6BduHfpyOOEmxIXcBXr/edit?usp=sharing&oid=117105050981428441732&rtpof=true&sd=true)  
8 [050981428441732&rtpof=true&sd=true](https://docs.google.com/document/d/1gJVHCOM59Ycz6BduHfpyOOEmxIXcBXr/edit?usp=sharing&oid=117105050981428441732&rtpof=true&sd=true)

9

## 10 Abstract

11 The genetic basis of severe COVID-19 has been thoroughly studied and many genetic  
12 risk factors shared between populations have been identified. However, reduced sample  
13 sizes from non-European groups have limited the discovery of population-specific  
14 common risk loci. In this second study nested in the SCOURGE consortium, we have  
15 conducted the largest GWAS meta-analysis for COVID-19 hospitalization in admixed  
16 Americans, comprising a total of 4,702 hospitalized cases recruited by SCOURGE and  
17 other seven participating studies in the COVID-19 Host Genetic Initiative. We identified  
18 four genome-wide significant associations, two of which constitute novel loci and first  
19 discovered in Latin-American populations (*BAZ2B* and *DDIAS*). A trans-ethnic meta-  
20 analysis revealed another novel cross-population risk locus in *CREBBP*. Finally, we  
21 assessed the performance of a cross-ancestry polygenic risk score in the SCOURGE  
22 admixed American cohort.

## 23 Introduction

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

24 To date more than 50 loci associated to COVID-19 susceptibility, hospitalization, and  
25 severity have been identified using genome-wide association studies (GWAS)<sup>1,2</sup>. The  
26 COVID-19 Host Genetics Initiative (HGI) has made significant efforts<sup>4</sup> to augment the  
27 power to identify disease loci by recruiting individuals from diverse populations and  
28 conducting a trans-ancestry meta-analysis. Despite this, the lack of genetic diversity and  
29 a focus on cases of European ancestries still predominate in the studies<sup>5,6</sup>. Besides, while  
30 trans-ancestry meta-analyses are a powerful approach for discovering shared genetic risk  
31 variants with similar effects across populations<sup>7</sup>, they may fail to identify risk variants  
32 that have larger effects on particular underrepresented populations. Genetic disease risk  
33 has been shaped by the particular evolutionary history of populations and the  
34 environmental exposures<sup>8</sup>. Their action is particularly important for infectious diseases  
35 due to the selective constrains that are imposed by the host-pathogen interactions<sup>9,10</sup>.  
36 Literature examples of such population specificities in COVID-19 severity includes a  
37 *DOCK2* gene variant in East Asians<sup>11</sup>, and frequent loss of function variants in *IFNAR1*  
38 and *IFNAR2* genes in Polynesian and Inuit populations, respectively<sup>12,13</sup>.

39 Including diverse populations in case-control GWAS studies with unrelated participants  
40 usually require a prior classification of individuals in genetically homogeneous groups,  
41 which are typically analysed separately to control the population stratification effects<sup>14</sup>.  
42 Populations with recent admixture impose an additional challenge to the GWAS due to  
43 their complex genetic diversity and linkage disequilibrium (LD) patterns, requiring the  
44 development of alternative approaches and a careful inspection of results to reduce the  
45 false positives due to population structure<sup>8</sup>. In fact, there are benefits in study power from  
46 modelling the admixed ancestries either locally, at regional scale in the chromosomes, or  
47 globally, across the genome, depending on factors such as the heterogeneity of the risk  
48 variant in frequencies or the effects among the ancestry strata<sup>15</sup>. Despite the development

49 of novel methods specifically tailored for the analysis of admixed populations<sup>16</sup>, the lack  
50 of a standardized analysis framework and the difficulties to confidently cluster the  
51 admixed individuals into particular genetic groups often leads to their exclusion from  
52 GWAS.

53 The Spanish Coalition to Unlock Research on Host Genetics on COVID-19 (SCOURGE)  
54 recruited COVID-19 patients between March and December 2020 from hospitals across  
55 Spain and from March 2020 to July 2021 in Latin-America ([https://www.scourge-](https://www.scourge-covid.org)  
56 [covid.org](https://www.scourge-covid.org)). A first GWAS of COVID-19 severity among Spanish patients of European  
57 descent revealed novel disease loci and explored age and sex varying effects of the genetic  
58 factors<sup>17</sup>. Here we present the findings of a GWAS meta-analysis in admixed American  
59 (AMR) populations, comprising individuals from the SCOURGE Latin-American cohort  
60 and the HGI studies, which allowed to identify two novel severe COVID-19 loci, *BAZ2B*  
61 and *DDIAS*. Further analyses modelling the admixture from three genetic ancestral  
62 components and performing a trans-ethnic meta-analysis led to the identification of an  
63 additional risk locus near *CREBBP*. We finally assessed a cross-ancestry polygenic risk  
64 score model with variants associated with critical COVID-19.

## 65 **Results**

### 66 **Meta-analysis of COVID-19 hospitalization in admixed Americans**

#### 67 *Study cohorts*

68 Within the SCOURGE consortium, we included 1,608 hospitalized cases and 1,887  
69 controls (not hospitalized COVID-19 patients) from Latin-American countries and from  
70 recruitments of individuals of Latin-American descent conducted in Spain  
71 (Supplementary Table 1). Quality control details and estimation of global genetic inferred  
72 ancestry (GIA) (supplementary Figure 1) are described in Methods, whereas clinical and

73 demographic characteristics of patients included in the analysis are shown in Table 1.  
74 Summary statistics from the SCOURGE cohort were obtained under a logistic mixed  
75 model with the SAIGE model (Methods). Another seven studies participating in the  
76 COVID-19 HGI consortium were included in the meta-analysis of COVID-19  
77 hospitalization in admixed Americans (Figure 1).

#### 78 *GWAS meta-analysis*

79 We performed a fixed-effects GWAS meta-analysis using the inverse of the variance as  
80 weights for the overlapping markers. The combined GWAS sample size consisted of  
81 4,702 admixed AMR hospitalized cases and 68,573 controls.

82 This GWAS meta-analysis revealed genome-wide significant associations at four risk  
83 loci, two of which (*BAZ2B* and *DDIAS*) were novel and considered specific to the  
84 populations included in this study (Table 2, Figure 2). Variants of these loci were  
85 prioritized by positional and expression quantitative trait loci (eQTL) mapping with  
86 FUMA, identifying four lead variants linked to other 310 variants and 31 genes  
87 (Supplementary Tables 2-4). A gene-based association test revealed a significant  
88 association in *BAZ2B* and in previously known risk loci: *LZTFL1*, *XCRI*, *FYCO1*, *CCR9*,  
89 and *IFNAR2* (Supplementary Table 5).

90 Located within the gene *BAZ2B*, rs13003835 is an intronic variant associated with an  
91 increased risk of COVID-19 hospitalization (Odds Ratio [OR]=1.20, 95% Confidence  
92 Interval [CI]=1.12-1.27,  $p=3.66 \times 10^{-8}$ ). This association was not previously reported in  
93 any GWAS of COVID-19 published to date. According to gnomAD v3.1.2, the T allele  
94 at rs13003835 has a frequency of 43% in admixed AMR groups while allele frequency  
95 (AF) is lower in the EUR populations (16%) and in the global sample (29%). Local  
96 ancestry inference (LAI) reported by gnomAD shows that within the Native-American

97 component, the risk allele T is the major allele, whereas it is the minor allele within the  
98 AFR and EUR LAI components. The T allele frequency in the SCOURGE Latin-  
99 American controls is consistent with gnomAD (Table 2). Interestingly, rs13003835 did  
100 not reach a significant association ( $p=0.972$ ) in the COVID-19 HGI trans-ancestry meta-  
101 analysis including the five population groups<sup>1</sup>. Based on our mapping strategy (see  
102 Methods), *PLA2R1*, *LY75*, *WDSUB1*, and *CD302* were other prioritized genes in this  
103 locus.

104 The other novel risk locus is led by rs77599934, an intronic rare variant located in  
105 chromosome 11 within *DDIAS* and associated with risk of COVID-19 hospitalization  
106 (OR=2.27, 95%CI=1.70-3.04,  $p=2.26 \times 10^{-8}$ ). *PRCP* gene was an additional prioritized  
107 gene at this locus. rs77599934 showed an AF of 1.1% for the G allele in the non-  
108 hospitalized controls (Table 2), in line with the recorded gnomAD AF of 1% in admixed  
109 AMR groups (0.02% in EUR populations and 2.6% in the global sample). Examining the  
110 LAI, the G allele occurs at 1.1% frequency in the AFR component while it is almost  
111 absent in the Native-American and EUR. This variant was not included in the COVID-  
112 19 HGI B2 trans-ancestry meta-analysis nor in the GWAS summary statistics for the EUR  
113 populations.

114 We observed a suggestive association with rs2601183 in chromosome 15, which is  
115 located between *ZNF774* and *IQGAP1* (allele G OR=1.20, 95%CI=1.12-1.29,  $p=6.11 \times 10^{-8}$ ,  
116 see Supplementary Table 2) and has not yet been reported. This sentinel variant is in  
117 perfect LD ( $r^2=1$ ) with rs601183, an eQTL of *ZNF774* in the lung.

118 The GWAS meta-analysis pinpointed two variants at known loci, *LZTFL1* and *FOXP4*.  
119 The SNP rs35731912 was previously associated with COVID-19 severity in EUR  
120 populations<sup>18</sup>, and it was mapped to *LZTFL1*. As for rs2477820, while it is a novel risk  
121 variant within gene *FOXP4*, it has a moderate LD ( $r^2=0.295$ ) with rs2496644, which has

122 been linked to COVID-19 hospitalization<sup>19</sup>. This is consistent with the effects of LD in  
123 tag-SNPs when conducting GWAS in diverse populations.

## 124 **Functional mapping of novel risk variants**

### 125 *Bayesian fine mapping*

126 We performed different approaches to narrow down the prioritized loci to a set of most  
127 probable genes driving the associations. First, we computed credible sets at the 95%  
128 confidence for causal variants and annotated them with VEP and the V2G aggregate  
129 scoring (Supplementary Table 6, Supplementary Figure 3). The 95% confidence credible  
130 set from the region of chromosome 2 around rs13003835 included 76 variants. However,  
131 the approach was unable to converge allocating variants in a 95% confidence credible set  
132 for the region in chromosome 11.

### 133 *Colocalization of eQTLs*

134 To determine if the novel genetic risk loci were associated with gene expression in  
135 relevant tissues (whole blood, lung, lymphocytes, and oesophagus mucosa), we computed  
136 the posterior probabilities of colocalization for overlapping variants allocated to the 95%  
137 confidence credible set. We used the GTEx v8 tissues as the main expression dataset,  
138 although it is important to consider that the eQTL associations were carried out mainly  
139 on individuals of EUR ancestries. To confirm the colocalization in other ancestries, we  
140 performed secondary analyses on three expression datasets computed on admixed AMR,  
141 leveraging data from individuals with high African GIA, high Native-American ancestry,  
142 and from a pooled cohort (Methods). Results are shown in the supplementary Table 7.

143 Five genes (*LY75*, *BAZ2B*, *CD302*, *WDSUB1*, and *PLA2R1*) were the candidates for  
144 eQTL colocalization in the associated region in chromosome 2. However, *LY75* emerged  
145 as the most likely causal gene for this locus since the colocalization in whole blood was

146 supported with a posterior probability for H4 (PPH4) of 0.941 and with robust results  
147 (supplementary Figure 4). Moreover, this also allowed to prioritize rs12692550 as the  
148 most probable causal variant for both traits at this locus with a PP\_SNP\_H4 of 0.74.  
149 Colocalization with gene expression data from admixed AMR validated this finding.  
150 *LY75* also had evidence of colocalization in lung (PPH4=0.887) and the esophagus  
151 mucosa (PPH4=0.758). However, we could not prioritize a single causal variant in these  
152 two other tissues and sensitivity analyses revealed a weak support.

153 *CD302* and *BAZ2B* were the second and third most likely genes that could drive the  
154 association, respectively, according to the colocalization evidence. *CD302* was the most  
155 probable according to the high AFR genetic ancestries dataset (supplementary Figure 5).

156 Despite the chromosome 11 region failing to colocalize with gene expression associations  
157 for any of the tissues, the lead variant rs77599934 is in moderate-to-strong LD ( $r^2=0.776$ )  
158 with rs60606421, which is an eQTL associated to a reduced expression of *DDIAS* in the  
159 lung (supplementary Figure 6). The highest PPH4 for *DDIAS* was in the high AFR genetic  
160 ancestry expression dataset (0.71).

#### 161 *Transcriptome-wide association study (TWAS)*

162 Five novel genes, namely *SLC25A37*, *SMARCC1*, *CAMP*, *TYW3*, and *S100A12*  
163 (supplementary Table 8) were found significantly associated in the cross-tissue TWAS.  
164 To our knowledge, these genes have not been reported previously in any COVID-19  
165 TWAS or GWAS analyses published to date. In the single tissue analyses, *ATP5O* and  
166 *CXCR6* were significantly associated in lung, *CCR9* was significantly associated in whole  
167 blood, and *IFNAR2* and *SLC25A37* were associated in lymphocytes.

168 Likewise, we carried out the TWAS analyses using the models trained in the admixed  
169 populations. However, no significant gene-pairs were detected. The 50 genes with the  
170 lowest p-values are shown in the supplementary Table 9.

### 171 **Sensitivity analyses for population specificity of associated loci**

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

176 First, we combined the SCOURGE Latin American GWAS results with the HGI B2 ALL  
177 analysis (supplementary Table 10). We refer to this analysis as the SC-HGI<sub>ALL</sub> meta-  
178 analysis. Out of the 40 genome-wide significant loci associated with COVID-19  
179 hospitalization in the last HGI release<sup>1</sup>, this study replicated 39 and the association was  
180 stronger than in the original study in 29 of those (supplementary Table 11). However, the  
181 variant rs13003835 located in *BAZ2B* was not associated with COVID-19 hospitalization  
182 (OR=1.00, 95%CI=0.98-1.03, p=0.644). The direction of the effect was opposite between  
183 AMR and the AFR, EUR, and EAS populations. Results for the variant rs77599934 (in  
184 *DDIAS*) could not be evaluated in the meta-analysis as it was absent from the HGI B2  
185 ALL results.

186 In this cross-ancestry meta-analysis, we replicated two associations that were not found  
187 in HGIv7 albeit they were sentinel variants in the latest GenOMICC meta-analysis<sup>2</sup>. We  
188 found an association at the *CASC20* locus led by the variant rs2876034 (OR=0.95,  
189 95%CI=0.93-0.97, p=2.83x10<sup>-8</sup>). This variant is in strong LD with the sentinel variant of  
190 that study (rs2326788, r<sup>2</sup>=0.92), which was associated with critical COVID-19<sup>2</sup>. Besides,  
191 this meta-analysis identified the variant rs66833742 near *ZBTB7A* associated with



192 COVID-19 hospitalization (OR=0.94, 95%CI=0.92-0.96,  $p=2.50 \times 10^{-8}$ ). Notably,  
193 rs66833742 or its perfect proxy rs67602344 ( $r^2=1$ ) are also associated with upregulation  
194 of *ZBTB7A* in whole blood and in esophagus mucosa. This variant was previously  
195 associated with COVID-19 hospitalization<sup>2</sup>.

196 In a second analysis, we also explored the associations across the defined admixed AMR,  
197 EUR, and AFR ancestral sources by combining through meta-analysis the SCOURGE  
198 Latin American GWAS results with the HGI studies in EUR, AFR, and admixed AMR,  
199 and excluding those from EAS and SAS (Supplementary Table 12). We refer to this as  
200 the SC-HGI<sub>3POP</sub> meta-analysis. The association at rs13003835 (*BAZ2B*, OR=1.01,  
201 95%CI=0.98-1.03,  $p=0.605$ ) was not replicated and rs77599934 near *DDIAS* could not  
202 be assessed, but the association at the *ZBTB7A* locus was confirmed (rs66833742,  
203 OR=0.94, 95%CI=0.92-0.96,  $p=1.89 \times 10^{-8}$ ). The variant rs76564172 located near  
204 *CREBBP* also reached statistical significance (OR=1.31, 95% CI=1.25-1.38,  $p=9.64 \times 10^{-9}$ ).  
205 The sentinel variant of the region linked to *CREBBP* (in the trans-ancestry meta-  
206 analysis) was also subjected to the same Bayesian fine mapping procedure  
207 (supplementary Table 6) and colocalization with eQTLs under the GTEx v8 MASHR  
208 models in lung, esophagus mucosa, whole blood, and transformed lymphocytes. Eight  
209 variants were included in the credible set for the region in chromosome 16 (meta-analysis  
210 SC-HGI<sub>3POP</sub>), although *CREBBP* did not colocalize in any of the tissues.

211 Finally, we evaluated the weight of AMR and AFR GIA proportions in the novel  
212 associations found at chromosomes 2 and 11 in the AMR GWAS meta-analysis. For that,  
213 the SCOURGE Latin American participants were classified into four groups depending  
214 on the AFR/AMR GIA by quantiles 1 and 3 (see Methods). The four groups tested were:  
215 large AFR GIA (>19%,  $N_{\text{ctrls}}=622$ ,  $N_{\text{cases}}=256$ ); low AFR GIA (<2.8%,  $N_{\text{ctrls}}=298$ ,  
216  $N_{\text{cases}}=580$ ); large AMR GIA (>56%,  $N_{\text{ctrls}}=166$ ,  $N_{\text{cases}}=712$ ); and low AMR GIA

217 (<18%,  $N_{\text{ctrls}}=651$ ,  $N_{\text{cases}}=227$ ). The variant rs13003835 at chromosome 2 was not  
218 significantly associated in any of the four groups. Contrarily, the rs77599934 at  
219 chromosome 11 was significantly associated among the individuals with larger AFR GIA  
220 (OR=3.33, 95%CI=1.70-6.52,  $p=5.00 \times 10^{-4}$ ).

### 221 **Polygenic risk score models**

222 Using the 49 variants associated with disease severity that are shared across populations  
223 according to the HGIv7, we constructed a polygenic risk score (PGS) model to assess its  
224 generalizability in the admixed AMR (Supplementary Table 13). First, we calculated the  
225 PGS for the SCOURGE Latin Americans and explored the association with COVID-19  
226 hospitalization under a logistic regression model. The PGS model was associated with a  
227 1.48-fold increase in COVID-19 hospitalization risk per every PGS standard deviation. It  
228 also contributed to explain a slightly larger variance ( $R^2=1.07\%$ ) than the baseline model.  
229 Subsequently, we divided the individuals into PGS deciles and percentiles to assess their  
230 risk stratification. The median percentile among controls was 40, while in cases it was  
231 63. Those in the top PGS decile exhibited a 5.90-fold (95% CI=3.29-10.60,  $p=2.79 \times 10^{-9}$ )  
232 greater risk compared to individuals in the lowest decile, whereas the effects for the rest  
233 of the comparisons were much milder.

234 We also examined the distribution of PGS scores across a 5-level severity scale to further  
235 determine if there was any correspondence between clinical severity and genetic risk.  
236 Median PGS scores were lower in the asymptomatic and mild groups, whereas higher  
237 median scores were observed in the moderate, severe, and critical patients (Figure 3). We  
238 fitted a multinomial model using the asymptomatic class as reference and calculated the  
239 OR for each category (Supplementary Table 13), observing that the disease genetic risk  
240 was similar among asymptomatic, mild, and moderate patients. Given that the PGS was

241 built with variants associated with critical disease and/or hospitalization and that the  
242 categories severe and critical correspond to hospitalized patients, these results underscore  
243 the ability of cross-ancestry PGS for risk stratification even in an admixed population.

244 Finally, to explore whether the risk variants that we deemed to be specific to the admixed  
245 AMR population enhanced the prediction of COVID-19 hospitalization, we incorporated  
246 the novel lead SNPs from our meta-analysis (rs13003835, rs2477820, and rs77599934)  
247 into the PGS model. Their inclusion in the model contributed to explain a larger variance  
248 ( $R^2=1.74\%$ ) than the model without them. This result, however, should be taken with  
249 caution given the risk of overfitting due to the use of the same subjects both for the  
250 derivation and testing of the variants.

251

## 252 **DISCUSSION**

253 We have conducted the largest GWAS meta-analysis of COVID-19 hospitalization in  
254 admixed AMR to date. While the genetic risk basis discovered for COVID-19 is largely  
255 shared among populations, trans-ancestry meta-analyses on this disease have primarily  
256 included EUR samples. This dominance of studies in Europeans, and the subsequent bias  
257 in sample sizes, have the potential to mask population-specific genetic risks. We have  
258 found two risk loci adjacent to *DDIAS* and *BAZ2B*, first discovered in Latin-American  
259 populations and not yet detected in other population groups. Interestingly, the sentinel  
260 variant rs77599934 in the *DDIAS* gene is a rare coding variant (~1% for allele G) that has  
261 not been analysed on the cross-ancestry meta-analysis completed so far. Its absence in  
262 EUR-centric GWAS meta-analyses likely results from its low allele frequency on that  
263 group (G allele is 0.02% in EUR).

264 Fine mapping of the region harbouring *DDIAS* did not reveal further information about  
265 which gene could be the more prone to be causal, or about the functional consequences  
266 of the risk variant. However, *DDIAS*, known as damage-induced apoptosis suppressor  
267 gene, is itself a plausible candidate gene, as the activity has been linked to DNA damage  
268 repair mechanisms. Depletion of *DDIAS* leads to an increase of ATM phosphorylation  
269 and the formation of p53-binding protein (53BP1) foci, a known biomarker of DNA  
270 double-strand breaks, indicating its potential role in double-strand break repair<sup>20</sup>.  
271 Similarly, knocking down *DDIAS* results in elevated levels of phosphorylated nuclear  
272 histone 2AX $\gamma$ , further emphasizing its role in DNA damage<sup>21</sup>. Interestingly, SARS-CoV-  
273 2 infection also triggers ATM kinase phosphorylation and the accumulation of DNA  
274 damage by inhibiting repair mechanisms<sup>22</sup>. This same study reported the activation of the  
275 pro-inflammatory pathway p38/MAPK by the virus, a response prompted as well by  
276 knocking-down *DDIAS*<sup>21</sup>. This gene has been found to interact with STAT3, regulating  
277 IL-6<sup>23</sup> and thus mediating inflammatory processes. While it has been primarily associated  
278 with cancer, particularly lung cancer<sup>24</sup>, our findings suggest that *DDIAS* gene may be  
279 indeed involved in viral response and inflammation through DNA damage repair. The  
280 sentinel variant was in strong LD with an eQTL that reduced gene expression of *DDIAS*  
281 in lung. Thus, one hypothesis could be that reduced expression of *DDIAS* could  
282 potentially facilitate SARS-CoV-2 infection. Exploration of the group with higher  
283 relative AFR GIA suggests that this association was partly driven by the AFR genetic  
284 admixture. This result aligns with the fact that this variant is nearly absent from the AMR  
285 and EUR components as reported by GnomAD LAI allele frequencies. Another  
286 prioritized gene from this region was *PRCP*, an angiotensinase that has been linked to  
287 hypertension and for which a hypothesis on its role on COVID-19 progression has been  
288 raised<sup>25,26</sup>.

289 The risk region found in chromosome 2 prioritized more than one gene. The lead variant  
290 rs13003835 is located within *BAZ2B*. *BAZ2B* encodes one of the regulatory subunits of  
291 the Imitation switch (ISWI) chromatin remodelers<sup>27</sup> constituting the BRF-1/BRF-5  
292 complexes with SMARCA1 and SMARCA5, respectively, and the association signal  
293 colocalized with eQTLs in whole blood. The gene *LY75* (encoding the lymphocyte  
294 antigen 75) also colocalized with eQTLs in whole blood, esophagus mucosa, and lung  
295 tissues. Lymphocyte antigen 75 is involved in immune processes through antigen  
296 presentation in dendritic cells and endocytosis<sup>28</sup>, and has been associated with  
297 inflammatory diseases, representing also a compelling candidate for the region. Increased  
298 expression of *LY75* has been detected within hours after the infection by SARS-CoV-  
299 2<sup>29,30</sup>. This variant was not associated within any of the extreme-GIA groups. Yet, local  
300 ancestry AF from gnomAD v3.1.2 reported a 1.51 times higher frequency of the risk allele  
301 (T) in the AMR component. Lastly, the signal of *CD302* colocalized in individuals with  
302 high AFR ancestral admixture in whole blood. This gene is located in the vicinity of *LY75*  
303 and both conform the readthrough *LY75-CD302*.

304 A third novel risk region was observed in chromosome 15, between the genes *IQGAPI*  
305 and *ZNF774*, although not reaching genome-wide significance.

306 Secondary analyses revealed five TWAS associated genes, some of which have been  
307 already linked to severe COVID-19. In a comprehensive multi-tissue gene expression  
308 profiling study<sup>31</sup>, decreased expression of *CAMP* and *S100A8/S100A9* genes in COVID-  
309 19 severe patients was observed, while another study detected the upregulation of  
310 *SCL25A37* among severe COVID-19 patients<sup>32</sup>. SMARCC1 is a subunit of the SWI/SNF  
311 chromatin remodelling complex that has been identified as pro-viral for SARS-CoV-2  
312 and other coronavirus strains through a genome-wide screen<sup>33</sup>. This complex is crucial  
313 for *ACE2* expression and the viral entry in the cell<sup>34</sup>.

314 To confirm the specificity to the admixed AMR populations of these novel risk variants  
315 for COVID-19 hospitalization, we performed two cross-ancestry meta-analyses including  
316 the SCOURGE Latin-American cohort GWAS findings. We found that the two novel risk  
317 variants did not associate with COVID-19 hospitalization outside the population-specific  
318 meta-analysis, supporting a potential ancestry-specific role of these genes on COVID-19  
319 severity and highlighting the importance of complementing trans-ancestry meta-analyses  
320 with group-specific analyses. Notably, this analysis did not replicate the association at the  
321 *DSTYK* locus, which was associated with severe COVID-19 in Brazilian individuals with  
322 higher European admixture<sup>35</sup>. This lack of replication supports the initial hypothesis of  
323 that study suggesting that the risk haplotype derived from European populations, as we  
324 have reduced the weight of this ancestral contribution in our study by excluding those  
325 individuals.

326 Moreover, these cross-ancestry meta-analyses pointed to three loci that were not genome-  
327 wide significant in the HGI v7 ALL meta-analysis: a novel locus at *CREBBP*, and two  
328 loci at *ZBTB7A* and *CASC20* that were reported in another meta-analysis. *CREBBP* and  
329 *ZBTB7A* achieved a stronger significance when considering only EUR, AFR, and  
330 admixed AMR GIA groups. According to a recent study, elevated levels of the *ZBTB7A*  
331 gene promote a quasi-homeostatic state between coronaviruses and host cells, preventing  
332 cell death by regulating oxidative stress pathways<sup>36</sup>. This gene is involved in several  
333 signalling pathways, such as B and T cell differentiation<sup>37</sup>. On a separate note, *CREBBP*  
334 encodes the CREB binding protein (CBP), involved in transcription activation, that is  
335 known to positively regulate the type I interferon response through virus-induced  
336 phosphorylation of IRF-3<sup>38</sup>. Besides, the CREBP/CBP interaction has been implicated in  
337 SARS-CoV-2 infection<sup>39</sup> via the cAMP/PKA pathway. In fact, cells with suppressed

338 *CREBBP* gene expression exhibit reduced replication of the so called Delta and Omicron  
339 SARS-CoV-2 variants<sup>39</sup>.

340 The cross-ancestry PGS model effectively stratified individuals based on their genetic  
341 risk and demonstrated consistency with the clinical severity classification of the patients.  
342 The inclusion of the population-specific variants in the PGS model slightly improved the  
343 predictive value of the PGS. However, it is important to confirm this last finding in an  
344 external admixed AMR cohort to address potential overfitting arising from using the same  
345 individuals both for the discovery of the associations and for testing the model.

346 This study is subject to limitations, mostly concerning the sample recruitment and  
347 composition. The SCOURGE Latino-American sample size is small and the GWAS is  
348 underpowered. Another limitation is the difference in case-control recruitment across  
349 sampling regions that, yet controlled for, may reduce the ability to observe significant  
350 associations driven by different compositions of the populations. In this sense, the  
351 identified risk loci might not replicate in a cohort lacking any of the parental population  
352 sources from the three-way admixture. Likewise, we could not explicitly control for  
353 socio-environmental factors that could have affected COVID-19 spread and  
354 hospitalization rates, although genetic principal components are known to capture non-  
355 genetic factors. Finally, we must acknowledge the lack of a replication cohort. We have  
356 used all the available GWAS data for COVID-19 hospitalization in admixed AMR in this  
357 meta-analysis due to the low number of studies conducted. Therefore, we had no studies  
358 to replicate or validate the results. These concerns may be addressed in the future by  
359 including more AMR GWAS studies in the meta-analysis, both by involving diverse  
360 populations in study designs and by supporting research from countries in Latin-America.

361 This study provides novel insights into the genetic basis of COVID-19 severity,  
362 emphasizing the importance of considering host genetic factors through using non-

363 European populations, especially of admixed sources. Such complementary efforts can  
364 pin down variants with population-specific effects and increase our knowledge on the  
365 host genetic factors of severe COVID-19.

366

## 367 **Materials and methods**

### 368 **GWAS in Latin Americans from SCOURGE**

#### 369 *The SCOURGE Latin American cohort*

370 A total of 3,729 of COVID-19 positive cases were recruited across five countries from  
371 Latin America (Mexico, Brazil, Colombia, Paraguay, and Ecuador) by 13 participating  
372 centres (supplementary Table 1) from March 2020 to July 2021. In addition, we included  
373 1,082 COVID-19 positive individuals recruited between March and December 2020 in  
374 Spain who either had evidence of origin from a Latin American country or showed  
375 inferred genetic admixture between AMR, EUR, and AFR (with < 0.05% SAS/EAS).  
376 These individuals were excluded from a previous SCOURGE study that focused on  
377 participants with European genetic ancestries<sup>17</sup>. We used hospitalization as a proxy for  
378 disease severity and defined as cases those COVID-19 positive patients that underwent  
379 hospitalization as a consequence of the infection and used as controls those that did not  
380 need hospitalization due to COVID-19.

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



386 Samples and data were processed following normalized procedures. The REDCap  
387 electronic data capture tool<sup>40,41</sup>, hosted at Centro de Investigación Biomédica en Red  
388 (CIBER) from the Instituto de Salud Carlos III (ISCIII), was used to collect and manage  
389 demographic, epidemiological, and clinical variables. Subjects were diagnosed for  
390 COVID-19 based on quantitative PCR tests (79.3%), or according to clinical (2.2%) or  
391 laboratory procedures (antibody tests: 16.3%; other microbiological tests: 2.2%).

#### 392 *SNP array genotyping*

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

396 Samples were genotyped with the Axiom Spain Biobank Array (Thermo Fisher  
397 Scientific) following the manufacturer's instructions in the Santiago de Compostela Node  
398 of the National Genotyping Center (CeGen-ISCIII; <http://www.usc.es/cegen>). This array  
399 contains probes for genotyping a total of 757,836 SNPs. Clustering and genotype calling  
400 were performed using the Axiom Analysis Suite v4.0.3.3 software.

#### 401 *Quality control steps and variant imputation*

402 A quality control (QC) procedure using PLINK 1.9<sup>42</sup> was applied to both samples and the  
403 genotyped SNPs. We excluded variants with a minor allele frequency (MAF) <1%, a call  
404 rate <98%, and markers strongly deviating from Hardy-Weinberg equilibrium  
405 expectations ( $p < 1 \times 10^{-6}$ ) with mid-p adjustment. We also explored the excess of  
406 heterozygosity to discard potential cross-sample contaminations. Samples missing >2%  
407 of the variants were filtered out. Subsequently, we kept the autosomal SNPs and removed  
408 high LD regions and conducted LD-pruning (windows of 1,000 SNPs, with step size of  
409 80 and  $r^2$  threshold of 0.1) to assess kinship and estimate the global ancestral proportions.

410 Kinship was evaluated based on IBD values, removing one individual from each pair with  
411  $PI\_HAT > 0.25$  that showed a Z0, Z1, and Z2 coherent pattern (according to the theoretical  
412 expected values for each relatedness level). Genetic principal components (PCs) were  
413 calculated with PLINK with the subset of LD pruned variants.

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

#### 418 *Genetic admixture estimation*

419 Global genetic inferred ancestry (GIA), referred to the genetic similarity to the used  
420 reference individuals, was estimated with the ADMIXTURE<sup>43</sup> v1.3 software following a  
421 two-step procedure. First, we randomly sampled 79 European (EUR) and 79 African  
422 (AFR) samples from The 1000 Genomes Project (1KGP)<sup>44</sup> and merged them with the 79  
423 Native American (AMR) samples from Mao et al.<sup>45</sup> keeping the biallelic SNPs. LD-  
424 pruned variants were selected from this merge using the same parameters as in the QC.  
425 We then run an unsupervised analysis with  $K=3$  to redefine and homogenize the clusters  
426 and to compose a refined reference for the analyses, by applying a threshold of  $\geq 95\%$   
427 belonging to a particular cluster. As a result of this, 20 AFR, 18 EUR, and 38 AMR  
428 individuals were removed. The same LD-pruned variants data from the remaining  
429 individuals were merged with the SCOURGE Latin American cohort to perform a  
430 supervised clustering and estimated admixture proportions. A total of 471 samples from  
431 the SCOURGE cohort with  $>80\%$  estimated European GIA were removed to reduce the  
432 weight of the European ancestral component, leaving a total of 3,512 admixed American  
433 (AMR) subjects for downstream analyses.

#### 434 *Association analysis*

435 Results for the SCOURGE Latin Americans GWAS were obtained testing for COVID-  
436 19 hospitalization as a surrogate of severity. To accommodate the continuum of GIA in  
437 the cohort, we opted for a joint testing of all the individuals as a single study using a  
438 mixed regression model, as this approach has demonstrated a greater power and to  
439 sufficiently control population structure<sup>46</sup>. The SCOURGE cohort consisted of 3,512  
440 COVID-19 positive patients: cases (n=1,625) were defined as hospitalized COVID-19  
441 patients and controls (n=1,887) as non-hospitalized COVID-19 positive patients.

442 Logistic mixed regression models were fitted using the SAIGEgds<sup>47</sup> package in R, which  
443 implements the two-step mixed SAIGE<sup>48</sup> model methodology and the SPA test. Baseline  
444 covariables included sex, age, and the first 10 PCs. To account for a potential  
445 heterogeneity in the recruitment and hospitalization criteria across the participating  
446 countries, we adjusted the models by groups of the recruitment areas classified in six  
447 categories: Brazil, Colombia, Ecuador, Mexico, Paraguay, and Spain. This dataset has not  
448 been used in any previously GWAS of COVID-19 published to date.

#### 449 **Meta-analysis of Latin-American populations**

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

457 The meta-analysis was performed using an inverse-variance weighting method in  
458 METAL<sup>49</sup>. Average allele frequency was calculated and variants with low imputation  
459 quality ( $R_{sq} < 0.3$ ) were filtered out, leaving 10,121,172 variants for meta-analysis.

460 Heterogeneity between studies was evaluated with the Cochran's-Q test. The inflation of  
461 results was assessed based on a genomic control ( $\lambda$ ).

## 462 **Definition of the genetic risk loci and putative functional impact**

### 463 *Definition of lead variant and novel loci*

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

### 473 *Annotation and initial mapping*

474 Functional annotation was done with FUMA<sup>50</sup> for those variants with a  $p\text{-value} < 5 \times 10^{-8}$   
475 or in moderate-to-strong LD ( $r^2 > 0.6$ ) with the lead variants, where the LD was calculated  
476 from the 1KGP AMR panel. Genetic risk loci were defined by collapsing LD-blocks  
477 within 250 kb. Then, genes, scaled CADD v1.4 scores, and RegulomeDB v1.1 scores  
478 were annotated for the resulting variants with ANNOVAR in FUMA<sup>50</sup>. Gene-based  
479 analysis was also performed using MAGMA<sup>51</sup> as implemented in FUMA, under the SNP-  
480 wide mean model using the 1KGP AMR reference panel. Significance was set at a

481 threshold  $p < 2.66 \times 10^{-6}$  (which assumes that variants can be mapped to a total of 18,817  
482 genes).

483 FUMA allowed us to perform an initial gene mapping by two approaches: (1) positional  
484 mapping, which assigns variants to genes by physical distance using 10-kb windows; and  
485 (2) eQTL mapping based on GTEx v.8 data from whole blood, lung, lymphocytes, and  
486 oesophagus mucosa tissues, establishing a False Discovery Rate (FDR) of 0.05 to declare  
487 significance for variant-gene pairs.

488 Subsequently, to assign the variants to the most likely gene driving the association, we  
489 refined the candidate genes by fine mapping the discovered regions and implementing  
490 functional mapping.

491 To conduct a Bayesian fine mapping, credible sets for the genetic loci considered novel  
492 findings were calculated on the results from each of the three meta-analyses to identify a  
493 subset of variants most likely containing the causal variant at 95% confidence level,  
494 assuming that there is a single causal variant and that it has been tested. We used  
495 *corrcoverage* (<https://cran.rstudio.com/web/packages/corrcoverage/index.html>) for R to  
496 calculate the posterior probabilities of the variant being causal for all variants with an  
497  $r^2 > 0.1$  with the leading SNP and within 1 Mb except for the novel variant in chromosome  
498 19, for which we used a window of 0.5 Mb. Variants were added to the credible set until  
499 the sum of the posterior probabilities was  $\geq 0.95$ . VEP  
500 (<https://www.ensembl.org/info/docs/tools/vep/index.html>) and the V2G aggregate  
501 scoring from Open Targets Genetics (<https://genetics.opentargets.org>) were used to  
502 annotate the biological function of the variants contained in the fine-mapped credible sets

503 *Colocalization analysis*

504 We also conducted colocalization analyses to identify the putative causal genes that could  
505 act through the regulation of gene expression. FUMA's eQTL mapping enabled the  
506 identification of genes whose expression was associated with the variants in whole blood,  
507 lung, lymphocytes, and oesophagus mucosa tissues. We combined this information with  
508 the VEP and V2G aggregate scoring to prioritize genes. For the fine-mapping regions, we  
509 included the variants within the calculated credible sets. In the cases where the fine  
510 mapping was unsuccessful, we considered variants within a 0.2 Mb window of the lead  
511 variant.

512 For each prioritized gene, we then run COLOC<sup>52</sup> to assess the evidence of colocalization  
513 between association signals and the eQTLs in each tissue, when at least one variant  
514 overlapped between them. COLOC estimates the posterior probability of two traits  
515 sharing the same causal variant in a locus. Prior probabilities of a variant being associated  
516 to COVID-19 phenotype ( $p_1$ ) and gene expression ( $p_2$ ) were set at  $1 \times 10^{-4}$ , while  $pp_2$  was  
517 set at  $1 \times 10^{-6}$  as they are robust thresholds<sup>53</sup>. A posterior probability of colocalization  
518 ( $PP_4$ )  $> 0.75$  and a ratio  $PP_4/PP_3 > 3$  were used as the criteria to support evidence of  
519 colocalization. Additionally, a threshold of  $PP_4.SNP > 0.5$  was chosen for causal variant  
520 prioritization. In cases where colocalization of a single variant failed, we computed the  
521 95% credible sets. The eQTL data was retrieved from GTEx v8 and only significant  
522 variant-gene pairs were considered in the analyses.

523 Colocalization in whole-blood was also performed using the recent published gene  
524 expression datasets derived from a cohort of African Americans, Puerto Ricans, and  
525 Mexican Americans (GALA II-SAGE)<sup>54</sup>. We used the results from the pooled cohort for  
526 the three discovered loci, and from the AFRHp5 (African genetic ancestry  $> 50\%$ ) and  
527 IAMHp5 (Native American genetic ancestry  $> 50\%$ ) cohorts for the risk loci in  
528 chromosomes 2 and 11. Results are shown in the Supplementary Table 10.

529 Sensitivity plots are shown in supplementary Figures 4 and 5.

### 530 **Transcription-wide association studies**

531 Transcriptome-wide association studies (TWAS) were conducted using the pretrained  
532 prediction models with MASHR-computed effect sizes on GTEx v8 datasets<sup>55,56</sup>. Results  
533 from the Latin-American meta-analysis were harmonized and integrated with the  
534 prediction models through S-PrediXcan<sup>57</sup> for lung, whole blood, lymphocytes and  
535 oesophagus mucosa tissues. Statistical significance was set at  $p\text{-value} < 0.05$  divided by  
536 the number of genes that were tested for each tissue. Subsequently, we leveraged results  
537 for all 49 tissues and run a multi-tissue TWAS to improve power for association, as  
538 demonstrated recently<sup>58</sup>. TWAS was also conducted with the MASHR models for whole-  
539 blood in the pooled admixed AMR from the GALA and SAGE studies<sup>54</sup>.

### 540 **Assessment of population specificity of associated loci**

541 We conducted two additional meta-analyses as a sensitivity analysis to determine the  
542 population specificity of the discovered risk loci. This methodology enabled the  
543 comparison of effects and the significance of associations in the novel risk loci between  
544 the results from analyses that included or excluded other population groups.

545 The first meta-analysis comprised the five populations analysed within HGI (B2-ALL).  
546 Additionally, to evaluate the three GIA components within the SCOURGE Latin-  
547 American cohort<sup>59</sup>, we conducted a meta-analysis of the admixed AMR, EUR, and AFR  
548 cohorts (B2). All summary statistics were retrieved from the HGI repository. We applied  
549 the same meta-analysis methodology and filters as in the admixed AMR meta-analysis.  
550 Novel variants from these meta-analyses were fine-mapped and colocalized with gene  
551 expression.

552 The effect of GIA was studied to determine whether any of the estimated genetic ancestry  
553 components interact with the associations at these loci. Individuals in the SCOURGE  
554 Latin American cohort were classified into large ( $\% \text{ GIA}_i > 3^{\text{rd}}$  quartile of the distribution)  
555 and small GIA ( $\% \text{ GIA}_i < 1^{\text{st}}$  quartile of the distribution). Quantiles 1 and 3 for the AFR  
556 component were 2.8% and 19%, and for the AMR component were 18% and 56%. Within  
557 each group, we tested the association of the two sentinel variants in chromosomes 2 and  
558 11 (accounting for baseline covariables).

### 559 **Trans-ethnic Polygenic Risk Score**

560 A polygenic risk score (PGS) for critical COVID-19 was derived combining the variants  
561 associated with hospitalization or disease severity that have been discovered to date. We  
562 curated a list of lead variants that were: 1) associated to either severe disease or  
563 hospitalization in the latest HGIv7 release<sup>1</sup> (using the hospitalization weights); or 2)  
564 associated to severe disease in the latest GenOMICC meta-analysis<sup>2</sup> that were not  
565 reported in the latest HGI release. A total of 49 markers were used in the PGS model (see  
566 supplementary Table 13) since two variants were absent from our study.

567 Scores were calculated and normalized for the SCOURGE Latin-American cohort with  
568 PLINK 1.9. This cross-ancestry PGS was used as a predictor for hospitalization (COVID-  
569 19 positive that were hospitalized *vs.* COVID-19 positive that did not necessitate hospital  
570 admission) by fitting a logistic regression model. Prediction accuracy for the PGS was  
571 assessed by performing 500 bootstrap resamples of the increase in the pseudo-R-squared.  
572 We also divided the sample in deciles and percentiles to assess risk stratification. The  
573 models were fit for the dependent variable adjusting for sex, age, the first 10 PCs, and the  
574 sampling region (in the Admixed AMR cohort) with and without the PGS, and the partial  
575 pseudo-R<sup>2</sup> was computed and averaged among the resamples.



576 A clinical severity scale was used in a multinomial regression model to further evaluate  
577 the power of this cross-ancestry PGS for risk stratification. This severity strata were  
578 defined as follows: 0) asymptomatic; 1) mild, that is, with symptoms, but without  
579 pulmonary infiltrates or need of oxygen therapy; 2) moderate, that is, with pulmonary  
580 infiltrates affecting <50% of the lungs or need of supplemental oxygen therapy; 3) severe  
581 disease, that is with hospital admission and  $\text{PaO}_2 < 65$  mmHg or  $\text{SaO}_2 < 90\%$ ,  
582  $\text{PaO}_2/\text{FiO}_2 < 300$ ,  $\text{SaO}_2/\text{FiO}_2 < 440$ , dyspnea, respiratory frequency  $\geq 22$  bpm, and infiltrates  
583 affecting >50% of the lungs; and 4) critical disease, that is with an admission to the ICU  
584 or need of mechanical ventilation (invasive or non-invasive). We also included the  
585 admixed AMR-specific risk variants as predictors alongside the PRS to determine if they  
586 provided increased prediction ability.

#### 587 **Data availability**

588 Summary statistics from the SCOURGE Latin-American GWAS will be available at  
589 <https://github.com/CIBERER/Scourge-COVID19>.

#### 590 **Funding**

591 Instituto de Salud Carlos III (COV20\_00622 to A.C., COV20/00792 to M.B.,  
592 COV20\_00181 to C.A., COV20\_1144 to M.A.J.S. and A.F.R., PI20/00876 to C.F.);  
593 European Union (ERDF) ‘A way of making Europe’. Fundación Amancio Ortega, Banco  
594 de Santander (to A.C.), Estrella de Levante S.A. and Colabora Mujer Association (to  
595 E.G.-N.) and Obra Social La Caixa (to R.B.); Agencia Estatal de Investigación (RTC-  
596 2017-6471-1 to C.F.), Cabildo Insular de Tenerife (CGIEU0000219140 ‘Apuestas  
597 científicas del ITER para colaborar en la lucha contra la COVID-19’ to C.F.) and  
598 Fundación Canaria Instituto de Investigación Sanitaria de Canarias (PIFIISC20/57 to  
599 C.F.).

600 SD-DA was supported by a Xunta de Galicia predoctoral fellowship.

### 601 **Author contributions**

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

### 607 **Acknowledgements**

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

619

### 620 **References**

621 1. Initiative, T. C.-19 H. G. & Ganna, A. A second update on mapping the human  
622 genetic architecture of COVID-19. 2022.12.24.22283874 Preprint at  
623 <https://doi.org/10.1101/2022.12.24.22283874> (2023).

- 624 2. GWAS and meta-analysis identifies 49 genetic variants underlying critical  
625 COVID-19 | Nature. <https://www.nature.com/articles/s41586-023-06034-3>.
- 626 3. Niemi, M. E. K. *et al.* Mapping the human genetic architecture of COVID-19.  
627 *Nature* **600**, 472–477 (2021).
- 628 4. Niemi, M. E. K. *et al.* Mapping the human genetic architecture of COVID-19.  
629 *Nature* **600**, 472–477 (2021).
- 630 5. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**,  
631 161–164 (2016).
- 632 6. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human  
633 Genetic Studies. *Cell* **177**, 26–31 (2019).
- 634 7. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies:  
635 advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 91  
636 (2014).
- 637 8. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations.  
638 *Nat. Rev. Genet.* **11**, 356–366 (2010).
- 639 9. Kwok, A. J., Mentzer, A. & Knight, J. C. Host genetics and infectious disease:  
640 new tools, insights and translational opportunities. *Nat. Rev. Genet.* **22**, 137–153 (2021).
- 641 10. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and  
642 infectious disease in human populations. *Nat. Rev. Genet.* **15**, 379–393 (2014).
- 643 11. Namkoong, H. *et al.* DOCK2 is involved in the host genetics and biology of  
644 severe COVID-19. *Nature* **609**, 754–760 (2022).
- 645 12. Bastard, P. *et al.* A loss-of-function IFNAR1 allele in Polynesia underlies severe  
646 viral diseases in homozygotes. *J. Exp. Med.* **219**, e20220028 (2022).
- 647 13. Duncan, C. J. A. *et al.* Life-threatening viral disease in a novel form of  
648 autosomal recessive IFNAR2 deficiency in the Arctic. *J. Exp. Med.* **219**, e20212427  
649 (2022).
- 650 14. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse  
651 Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–  
652 603 (2019).
- 653 15. Mester, R. *et al.* Impact of cross-ancestry genetic architecture on GWAS in  
654 admixed populations. 2023.01.20.524946 Preprint at  
655 <https://doi.org/10.1101/2023.01.20.524946> (2023).
- 656 16. Tractor uses local ancestry to enable the inclusion of admixed individuals in  
657 GWAS and to boost power | Nature Genetics. [https://www.nature.com/articles/s41588-](https://www.nature.com/articles/s41588-020-00766-y)  
658 [020-00766-y](https://www.nature.com/articles/s41588-020-00766-y).
- 659 17. Cruz, R. *et al.* Novel genes and sex differences in COVID-19 severity. *Hum.*  
660 *Mol. Genet.* **31**, 3789–3806 (2022).

- 661 18. Degenhardt, F. *et al.* Detailed stratified GWAS analysis for severe COVID-19 in  
662 four European populations. *Hum. Mol. Genet.* **31**, 3945–3966 (2022).
- 663 19. Whole-genome sequencing reveals host factors underlying critical COVID-19 |  
664 Nature. <https://www.nature.com/articles/s41586-022-04576-6>.
- 665 20. Evolution-based screening enables genome-wide prioritization and discovery of  
666 DNA repair genes | PNAS. <https://www.pnas.org/doi/full/10.1073/pnas.1906559116>.
- 667 21. Human Noxin is an anti-apoptotic protein in response to DNA damage of A549  
668 non-small cell lung carcinoma - Won - 2014 - International Journal of Cancer - Wiley  
669 Online Library. <https://onlinelibrary.wiley.com/doi/10.1002/ijc.28600>.
- 670 22. Gioia, U. *et al.* SARS-CoV-2 infection induces DNA damage, through CHK1  
671 degradation and impaired 53BP1 recruitment, and cellular senescence. *Nat. Cell Biol.*  
672 **25**, 550–564 (2023).
- 673 23. Im, J.-Y. *et al.* DDIAS promotes STAT3 activation by preventing STAT3  
674 recruitment to PTPRM in lung cancer cells. *Oncogenesis* **9**, 1–11 (2020).
- 675 24. Im, J.-Y., Kang, M.-J., Kim, B.-K. & Won, M. DDIAS, DNA damage-induced  
676 apoptosis suppressor, is a potential therapeutic target in cancer. *Exp. Mol. Med.* 1–7  
677 (2023) doi:10.1038/s12276-023-00974-6.
- 678 25. Angeli, F. *et al.* The spike effect of acute respiratory syndrome coronavirus 2  
679 and coronavirus disease 2019 vaccines on blood pressure. *Eur. J. Intern. Med.* **109**, 12–  
680 21 (2023).
- 681 26. Silva-Aguiar, R. P. *et al.* Role of the renin-angiotensin system in the  
682 development of severe COVID-19 in hypertensive patients. *Am. J. Physiol.-Lung Cell.*  
683 *Mol. Physiol.* **319**, L596–L602 (2020).
- 684 27. Li, Y. *et al.* The emerging role of ISWI chromatin remodeling complexes in  
685 cancer. *J. Exp. Clin. Cancer Res.* **40**, 346 (2021).
- 686 28. The Dendritic Cell Receptor for Endocytosis, Dec-205, Can Recycle and  
687 Enhance Antigen Presentation via Major Histocompatibility Complex Class II–Positive  
688 Lysosomal Compartments | Journal of Cell Biology | Rockefeller University Press.  
689 [https://rupress.org/jcb/article/151/3/673/21295/The-Dendritic-Cell-Receptor-for-](https://rupress.org/jcb/article/151/3/673/21295/The-Dendritic-Cell-Receptor-for-Endocytosis-Dec)  
690 [Endocytosis-Dec.](https://rupress.org/jcb/article/151/3/673/21295/The-Dendritic-Cell-Receptor-for-Endocytosis-Dec)
- 691 29. Sims, A. C. *et al.* Release of Severe Acute Respiratory Syndrome Coronavirus  
692 Nuclear Import Block Enhances Host Transcription in Human Lung Cells. *J. Virol.* **87**,  
693 3885–3902 (2013).
- 694 30. A Network Integration Approach to Predict Conserved Regulators Related to  
695 Pathogenicity of Influenza and SARS-CoV Respiratory Viruses | PLOS ONE.  
696 <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0069374>.
- 697 31. Gómez-Carballa, A. *et al.* A multi-tissue study of immune gene expression  
698 profiling highlights the key role of the nasal epithelium in COVID-19 severity. *Environ.*  
699 *Res.* **210**, 112890 (2022).

- 700 32. Policard, M., Jain, S., Rego, S. & Dakshanamurthy, S. Immune characterization  
701 and profiles of SARS-CoV-2 infected patients reveals potential host therapeutic targets  
702 and SARS-CoV-2 oncogenesis mechanism. *Virus Res.* **301**, 198464 (2021).
- 703 33. Wei, J. *et al.* Genome-wide CRISPR Screens Reveal Host Factors Critical for  
704 SARS-CoV-2 Infection. *Cell* **184**, 76-91.e13 (2021).
- 705 34. Wei, J. *et al.* Pharmacological disruption of mSWI/SNF complex activity  
706 restricts SARS-CoV-2 infection. *Nat. Genet.* **55**, 471–483 (2023).
- 707 35. Pereira, A. C. *et al.* Genetic risk factors and COVID-19 severity in Brazil:  
708 results from BRACOVID study. *Hum. Mol. Genet.* **31**, 3021–3031 (2022).
- 709 36. Zhu, X. *et al.* ZBTB7A promotes virus-host homeostasis during human  
710 coronavirus 229E infection. *Cell Rep.* **41**, 111540 (2022).
- 711 37. Gupta, S. *et al.* Emerging role of ZBTB7A as an oncogenic driver and  
712 transcriptional repressor. *Cancer Lett.* **483**, 22–34 (2020).
- 713 38. Yoneyama, M. *et al.* Direct triggering of the type I interferon system by virus  
714 infection: activation of a transcription factor complex containing IRF-3 and CBP/p300.  
715 *EMBO J.* **17**, 1087–1095 (1998).
- 716 39. Yang, Q. *et al.* SARS-CoV-2 infection activates CREB/CBP in cellular cyclic  
717 AMP-dependent pathways. *J. Med. Virol.* **95**, e28383 (2023).
- 718 40. Harris, P. A. *et al.* Research electronic data capture (REDCap)—A metadata-  
719 driven methodology and workflow process for providing translational research  
720 informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
- 721 41. Harris, P. A. *et al.* The REDCap consortium: Building an international  
722 community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
- 723 42. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and  
724 Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 725 43. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of  
726 ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- 727 44. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–  
728 74 (2015).
- 729 45. Mao, X. *et al.* A Genomewide Admixture Mapping Panel for Hispanic/Latino  
730 Populations. *Am. J. Hum. Genet.* **80**, 1171–1178 (2007).
- 731 46. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery  
732 for complex traits. *Nature* **570**, 514–518 (2019).
- 733 47. Zheng, X. & Davis, J. W. SAIGEgds—an efficient statistical tool for large-scale  
734 PheWAS with mixed models. *Bioinformatics* **37**, 728–730 (2021).
- 735 48. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample  
736 relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

- 737 49. METAL: fast and efficient meta-analysis of genomewide association scans |  
738 Bioinformatics | Oxford Academic.  
739 <https://academic.oup.com/bioinformatics/article/26/17/2190/198154>.
- 740 50. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional  
741 mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826  
742 (2017).
- 743 51. MAGMA: Generalized Gene-Set Analysis of GWAS Data | PLOS  
744 Computational Biology.  
745 <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004219>.
- 746 52. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of  
747 Genetic Association Studies Using Summary Statistics. *PLOS Genet.* **10**, e1004383  
748 (2014).
- 749 53. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in  
750 colocalisation analyses. *PLOS Genet.* **16**, e1008720 (2020).
- 751 54. Kachuri, L. *et al.* Gene expression in African Americans, Puerto Ricans and  
752 Mexican Americans reveals ancestry-specific patterns of genetic architecture. *Nat.*  
753 *Genet.* **55**, 952–963 (2023).
- 754 55. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms  
755 at GWAS loci. *Genome Biol.* **22**, 49 (2021).
- 756 56. Barbeira, A. N. *et al.* GWAS and GTEx QTL integration. (2019)  
757 doi:10.5281/zenodo.3518299.
- 758 57. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific  
759 gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**,  
760 1825 (2018).
- 761 58. Barbeira, A. N. *et al.* Integrating predicted transcriptome from multiple tissues  
762 improves association detection. *PLOS Genet.* **15**, e1007889 (2019).
- 763 59. Genome-wide patterns of population structure and admixture among  
764 Hispanic/Latino populations | PNAS.  
765 [https://www.pnas.org/doi/10.1073/pnas.0914618107?url\\_ver=Z39.88-](https://www.pnas.org/doi/10.1073/pnas.0914618107?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%20pubmed)  
766 [2003&rfr\\_id=ori:rid:crossref.org&rfr\\_dat=cr\\_pub%20%20pubmed](https://www.pnas.org/doi/10.1073/pnas.0914618107?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%20pubmed).

767

768

769

770

771

772

773

774

775

776 Table 1. Demographic characteristics of the SCOURGE Latin-American cohort.

Variable	Non Hospitalized N = 1,887	Hospitalized N = 1,608
Age – mean years ± SD	39.1 ± 11.9	54.1 ± 14.5
Sex - N (%)		
Female (%)	1253 (66.4)	668 (41.5)
GIA* – % mean ±SD		
European	54.4 ± 16.2	39.4 ± 20.7
African	15.3 ± 12.7	9.1 ± 11.6
Native American	30.3 ± 19.8	51.46 ± 26.5
Comorbidities - N (%)		
Vascular/endocrinological	488 (25.9)	873 (54.3)
Cardiac	60 (3.2)	150 (9.3)
Nervous	15 (0.8)	61 (3.8)
Digestive	14 (0.7)	33 (2.0)
Onco-hematological	21 (1.1)	48 (3.00)
Respiratory	76 (4.0)	118 (7.3)

777 \*Global genetic inferred ancestry.

778 Table 2. Lead independent variants in the admixed AMR GWAS meta-analysis.

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

779 EA: effect allele; NEA: non-effect allele; EAF: effect allele frequency in the SCOURGE study.

780

781

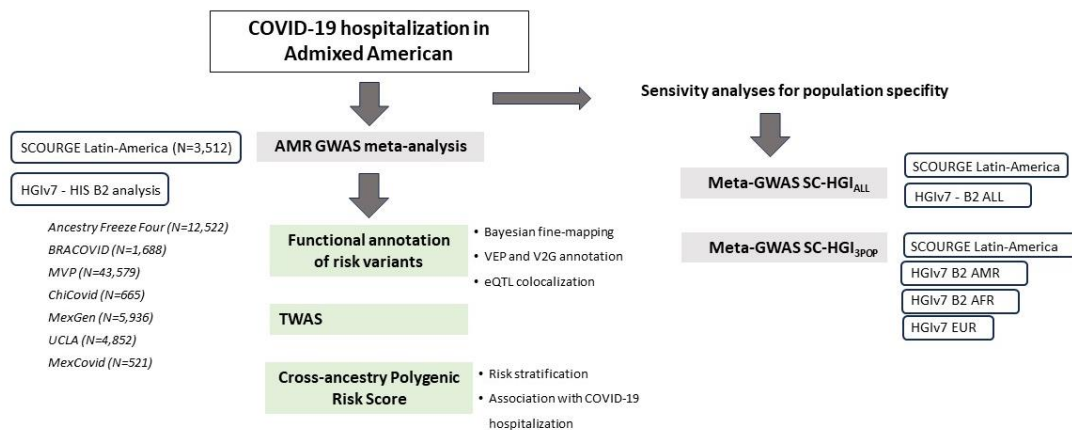
782

783 Table 3. Novel variants in the SC-HGI<sub>ALL</sub> and SC-HGI<sub>3POP</sub> meta-analyses (with respect  
784 to HGIv7). Independent signals after LD clumping.

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

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

786 **Figure 1. Flow chart of this study.**

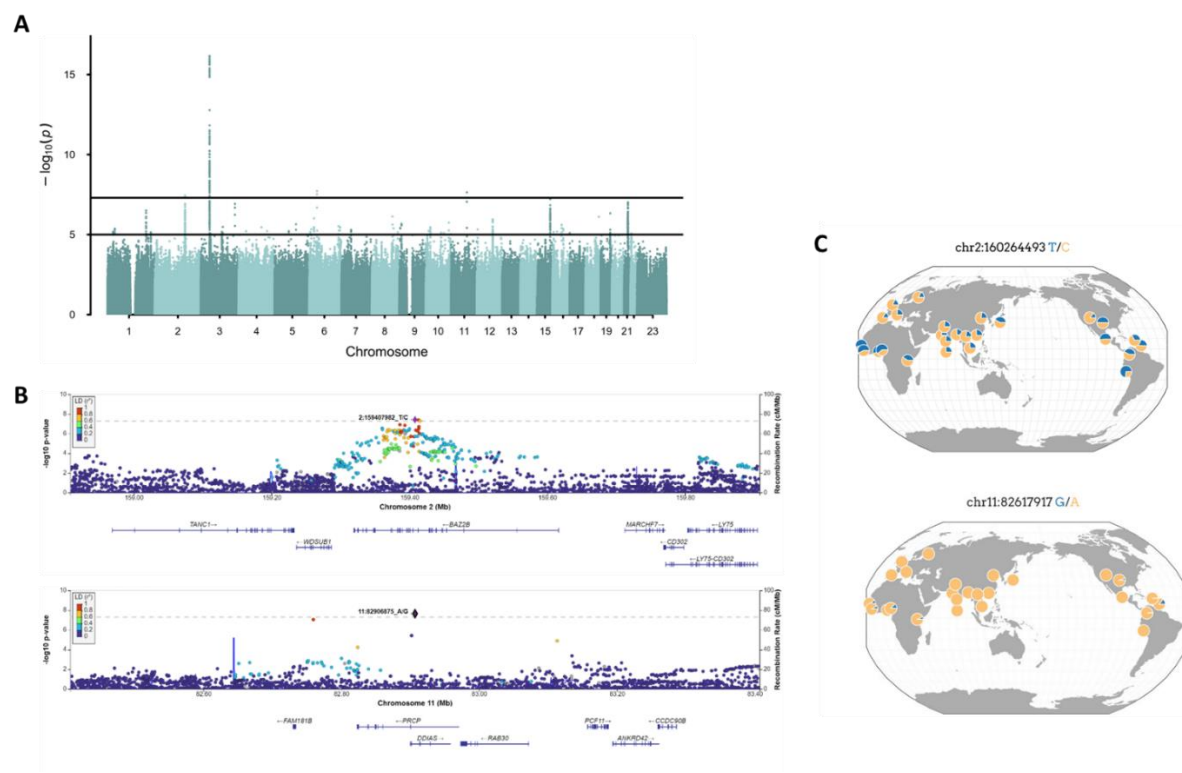


787

788 Figure 2. A) Manhattan plot for the admixed AMR GWAS meta-analysis. Probability  
789 thresholds at  $p=5 \times 10^{-8}$  and  $p=5 \times 10^{-5}$  are indicated by the horizontal lines. Genome-wide  
790 significant associations with COVID-19 hospitalizations were found in chromosome 2  
791 (within *BAZ2B*), chromosome 3 (within *LZTFL1*), chromosome 6 (within *FOXP4*), and  
792 chromosome 11 (within *DDIAS*). A Quantile-Quantile plot is shown in supplementary

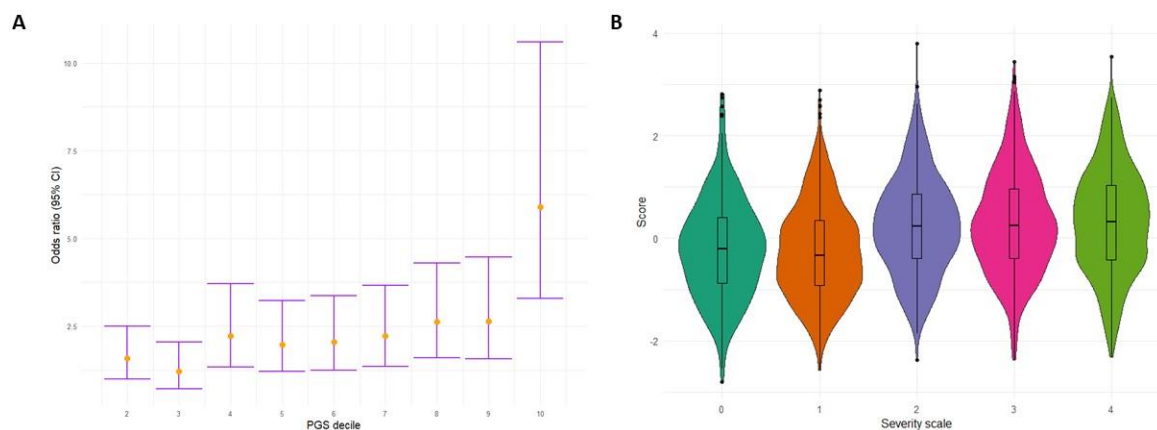


793 Figure 2. B) Regional association plots for rs1003835 at chromosome 2 and rs77599934  
794 at chromosome 11; C) Allele frequency distribution across The 1000 Genomes Project  
795 populations for the lead variants rs1003835 and rs77599934.



796

797 Figure 3. (A) Polygenic risk stratified by PGS deciles comparing each risk group against  
798 the lowest risk group (OR-95% CI); (B) Distribution of the PGS scores in each of the  
799 severity scale classes (0-Asymptomatic, 1-Mild disease, 2-Moderate disease, 3-Severe  
800 disease, 4-Critical disease).



801

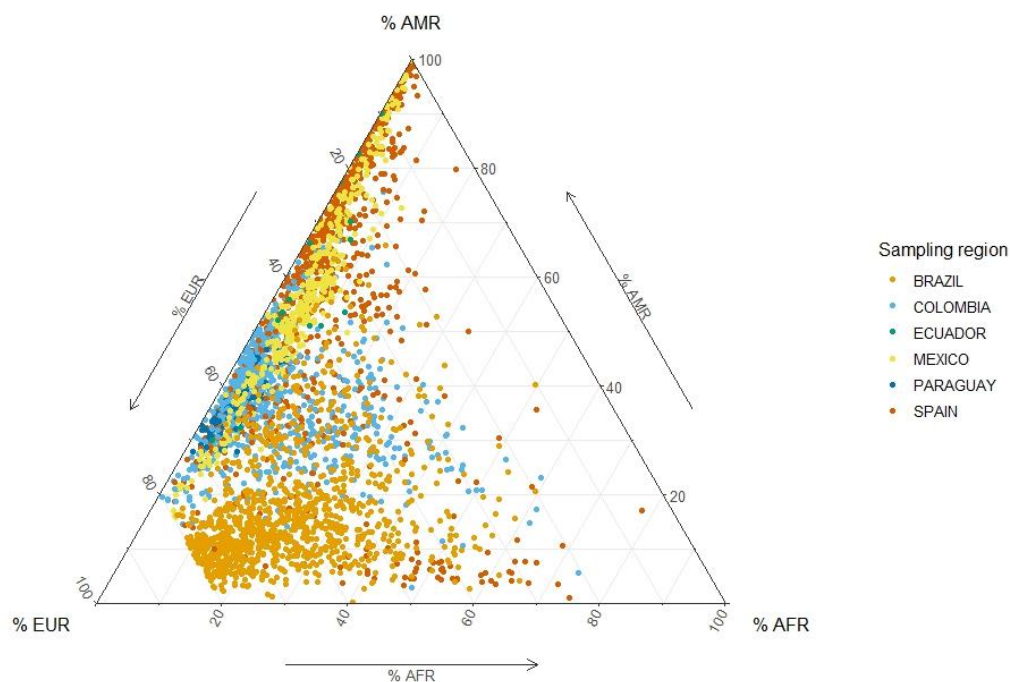
802

803 **Supplementary Material for: Novel risk loci for COVID-19 hospitalization among**  
804 **admixed American populations**

805 Supplementary Tables are provided in a separate excel file

806 **Supplementary figures**

807 **Supplementary Figure 1. Global Genetic Inferred Ancestry (GIA) composition in the**  
808 **SCOURGE Latin-American cohort.** European (EUR), African (AFR) and Native American  
809 (AMR) GIA was derived with ADMIXTURE from a reference panel composed of Aymaran,  
810 Mayan, Nahuatl, and Quechuan individuals of Native-American genetic ancestry and randomly  
811 selected samples from the EUR and AFR 1KGP populations. The colours represent the different  
812 geographical sampling regions from which the admixed American individuals from SCOURGE  
813 were recruited.



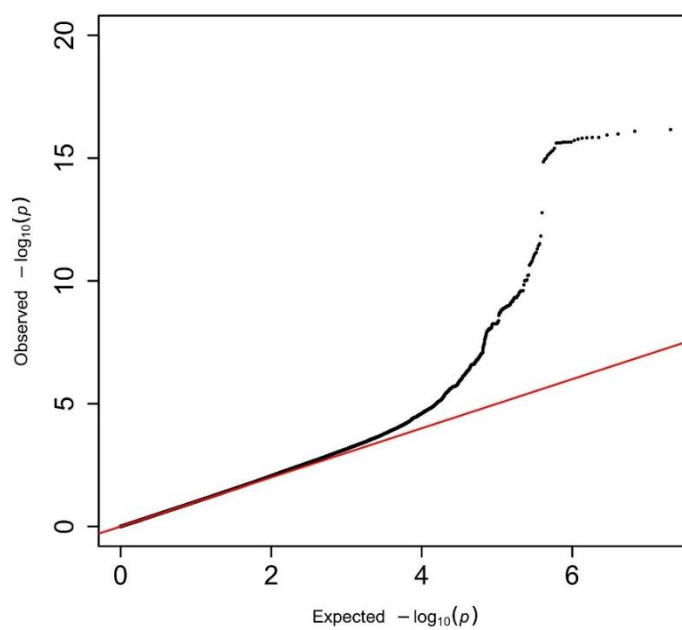
814

815

816 **Supplementary Figure 2. Quantile-Quantile plot for the AMR GWAS meta-analysis. A**

817 lambda inflation factor of 1.015 was obtained.

818



819

820

821

822

823

824

825

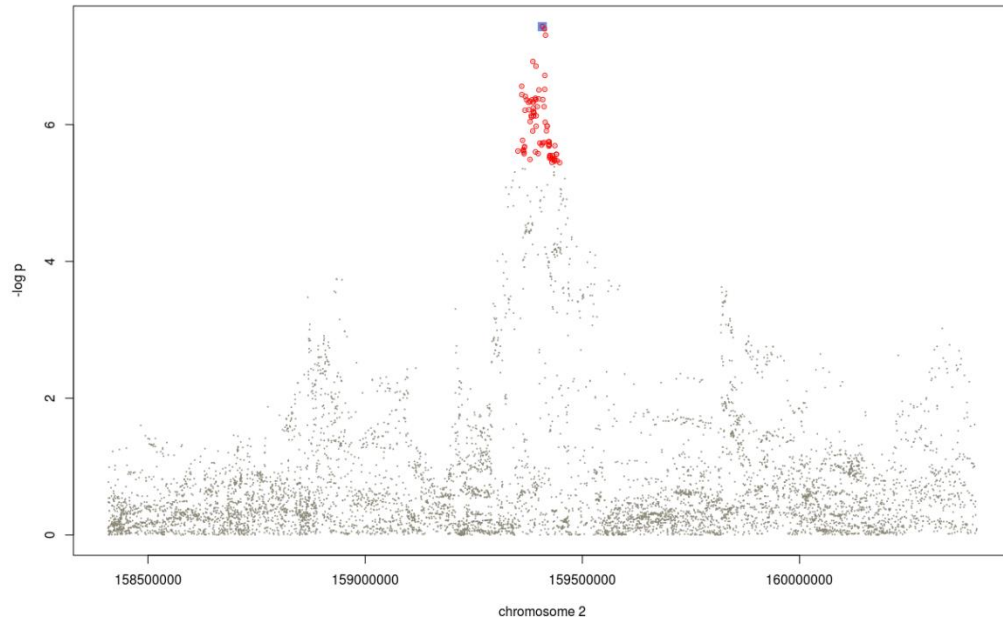
826

827

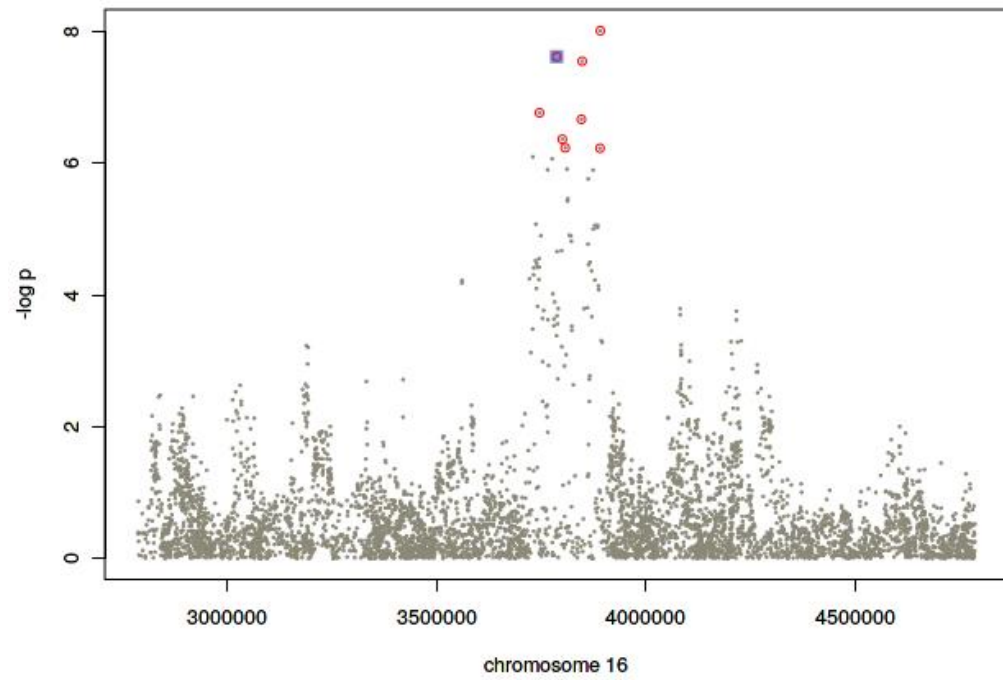
828

829

830 **Supplementary Figure 3. Regional association plots for the fine mapped loci in**  
831 **chromosomes 2 (upper panel) and 16 (lower panel).** Coloured in red, the variants allocated to  
832 the credible set at the 95% confidence according to the Bayesian fine mapping. In blue, the  
833 sentinel variant.



834

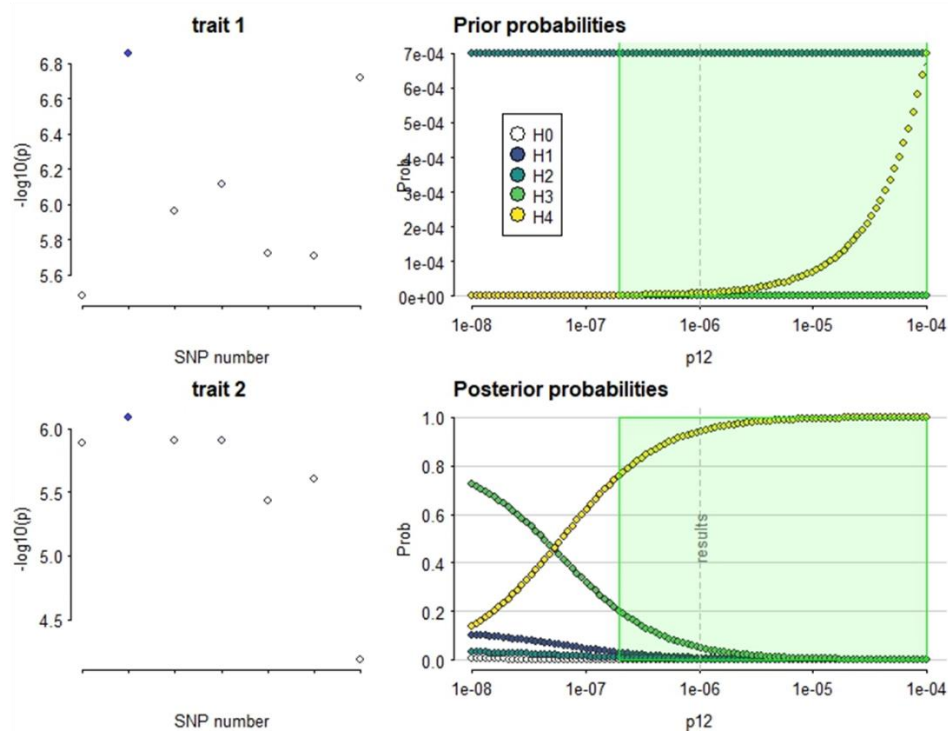


835

836

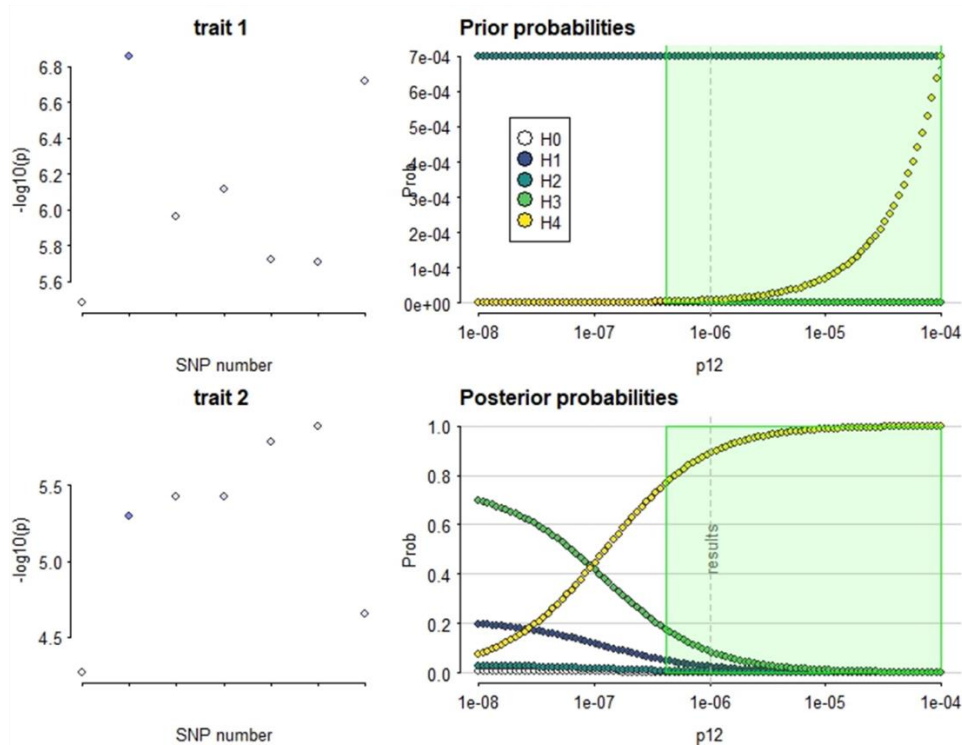
837 **Supplementary Figure 4. Sensitivity plots from COLOC with expression data from GTEx**  
838 **v8.** The range of  $p_{12}$  values (probability that a SNP is associated with both traits) for which the  
839 rule  $H_4 > 0.7$  is supported is shown in green in the right plots for each analysis. Plots in the left  
840 represent the variants included in the risk region common to both traits along their individual  
841 association  $-\log_{10}(p)$  for each trait, whereas the shading shows the posterior probability  
842 that the SNP is causal given  $H_4$  is true. Trait 1 corresponds to COVID-19 hospitalization, while  
843 trait 2 corresponds to gene expression in each analysis.  
844

### LY75 in whole blood



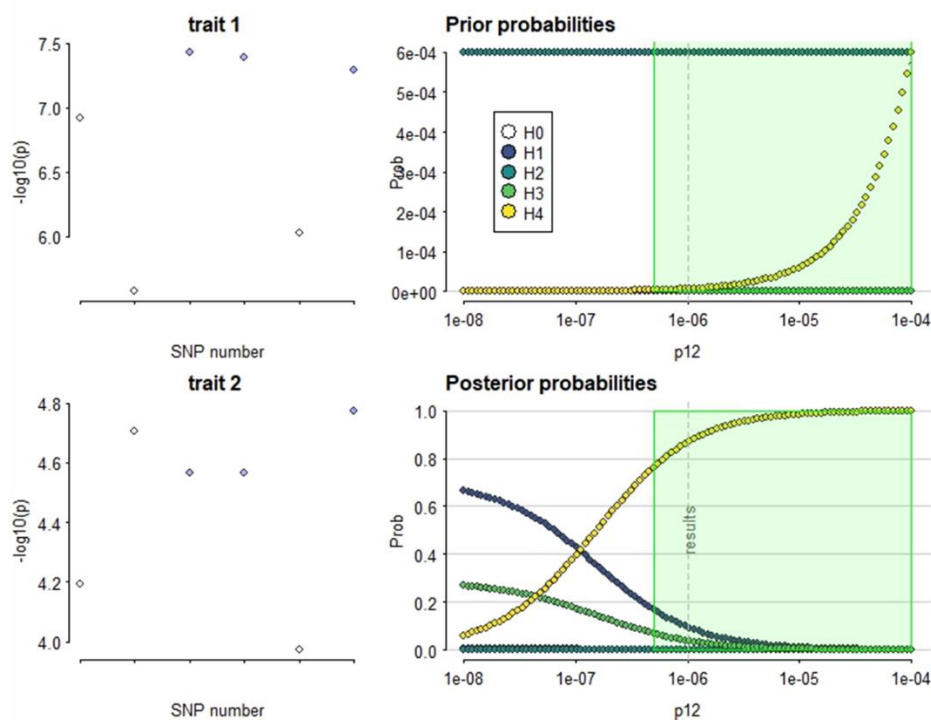
845

## LY75 in lung



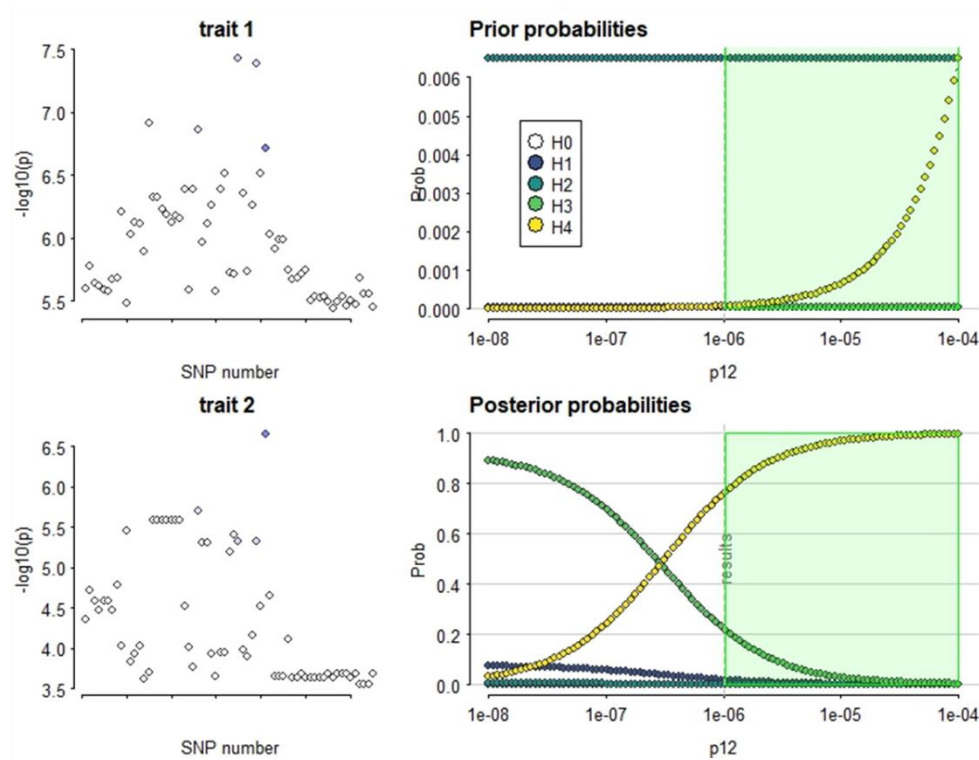
846

## BAZ2B in whole blood



847

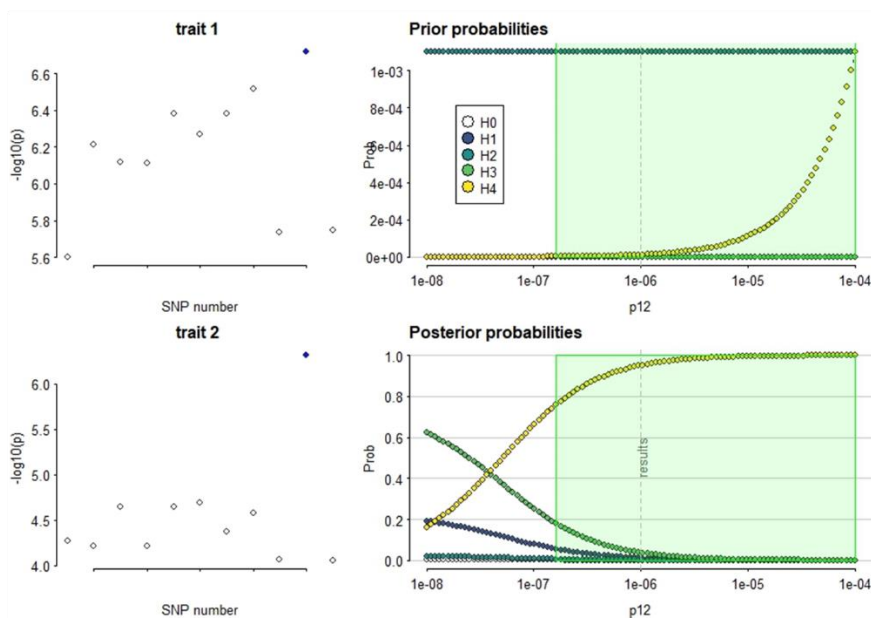
## LY75 in esophagus mucosa



850 **Supplementary Figure 5. Sensitivity plots from COLOC with whole blood**  
851 **expression data from the GALA and SAGE II studies in AMR individuals. *AFRhp5***  
852 **corresponds to the expression dataset computed in individuals with high African**  
853 **ancestries; *AMRhp5* corresponds to the expression dataset computed individuals with**  
854 **high AMR ancestries; *pooled* corresponds to the dataset computed with the total of**  
855 **individuals from the study. In the right, the plots show in green the range of p12 values**  
856 **(probability that a SNP is associated with both traits) for which the rule  $H_4 > 0.7$  is**  
857 **supported. Plots in the left represent the variants included in the risk region common to**  
858 **both traits along their individual association  $-\log_{10}(p\text{-values})$  for each trait, whereas the**  
859 **shading shows the posterior probability that the SNP is causal given  $H_4$  is true. Trait 1**  
860 **corresponds to COVID-19 hospitalization, while trait 2 corresponds to gene expression.**

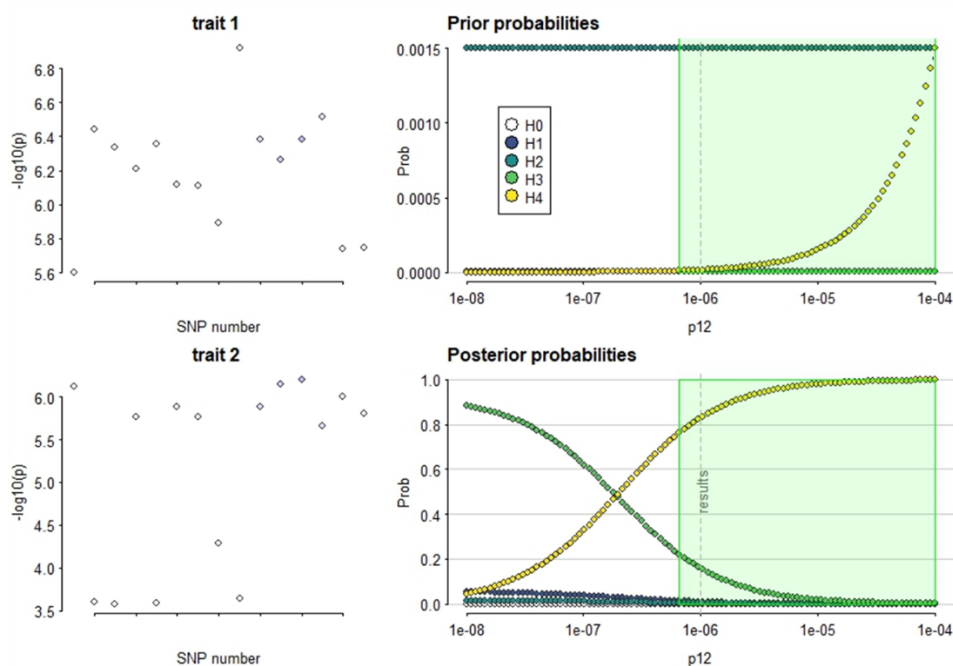


### CD302 in whole blood AFRhp5



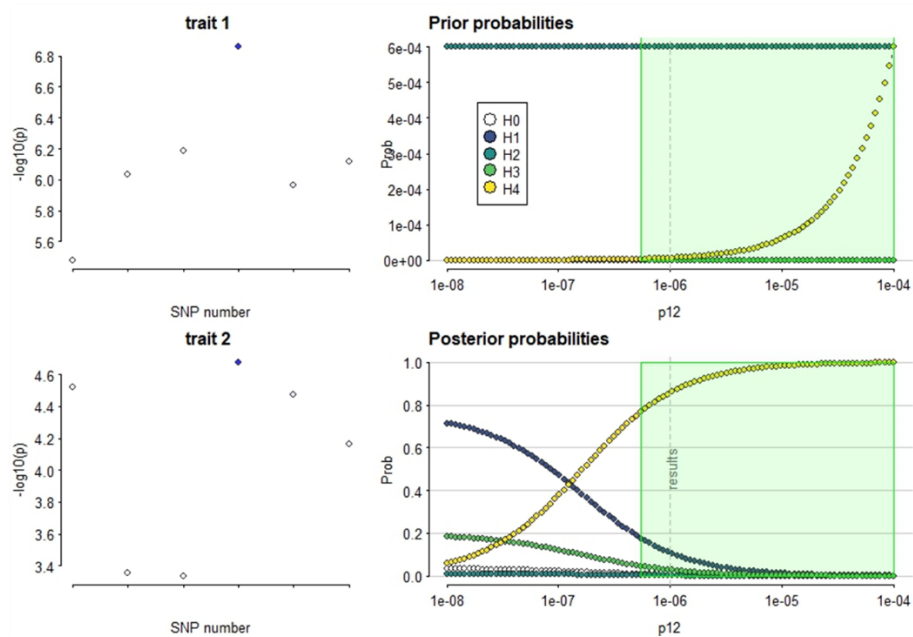
861

### BAZ2B in whole blood AFRhp5



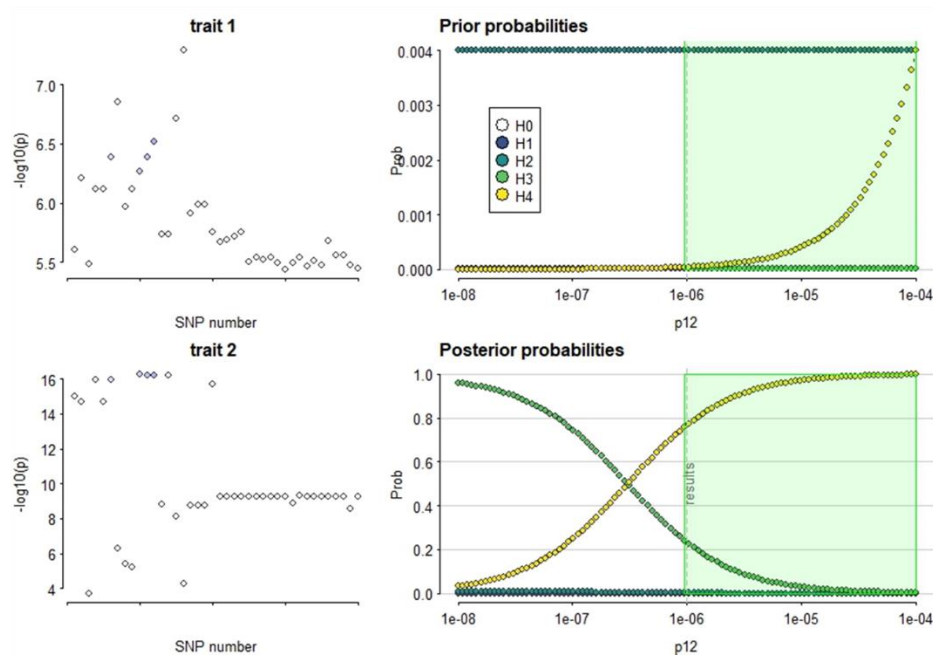
862

### WDSUB1 in whole blood AFRhp5



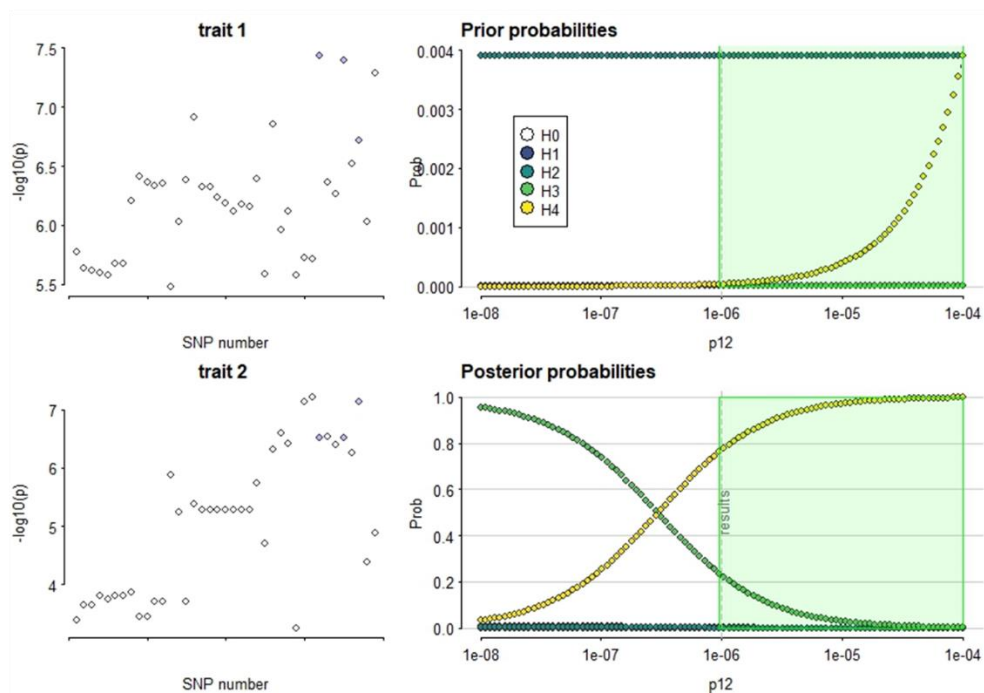
863

### LY75 in whole blood AFRhp5



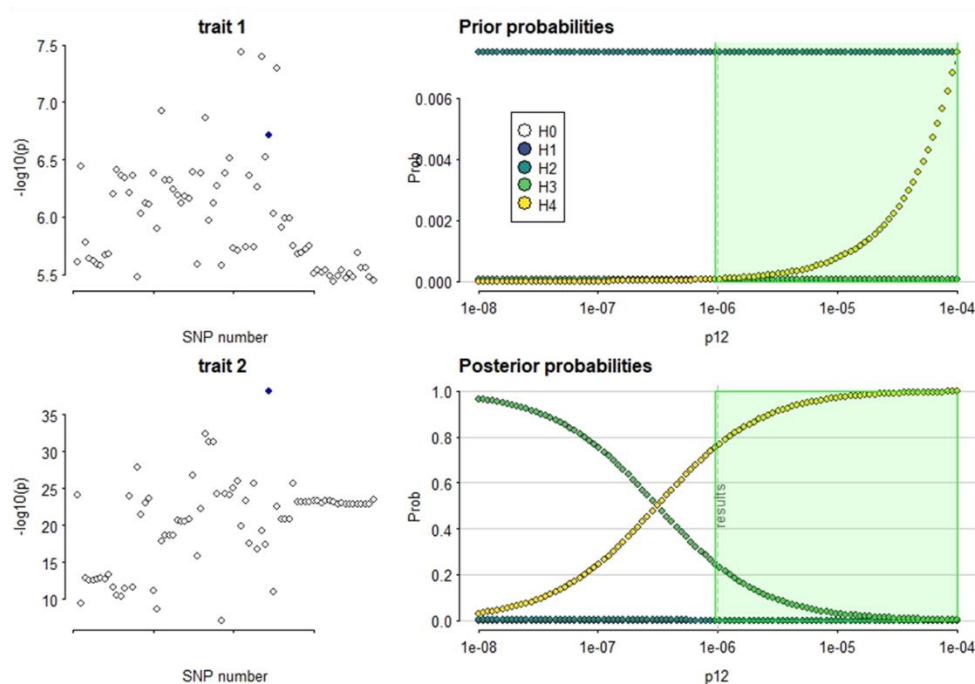
864

### LY75 in whole blood AMRhp5



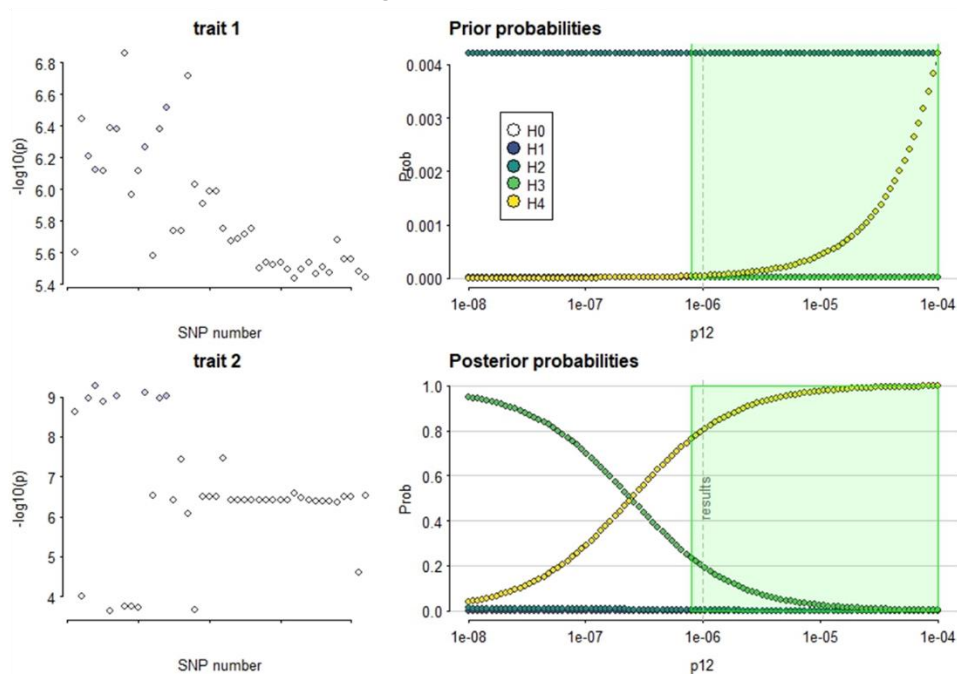
865

### LY75 in whole blood pooled



866

### CD302 in whole blood pooled



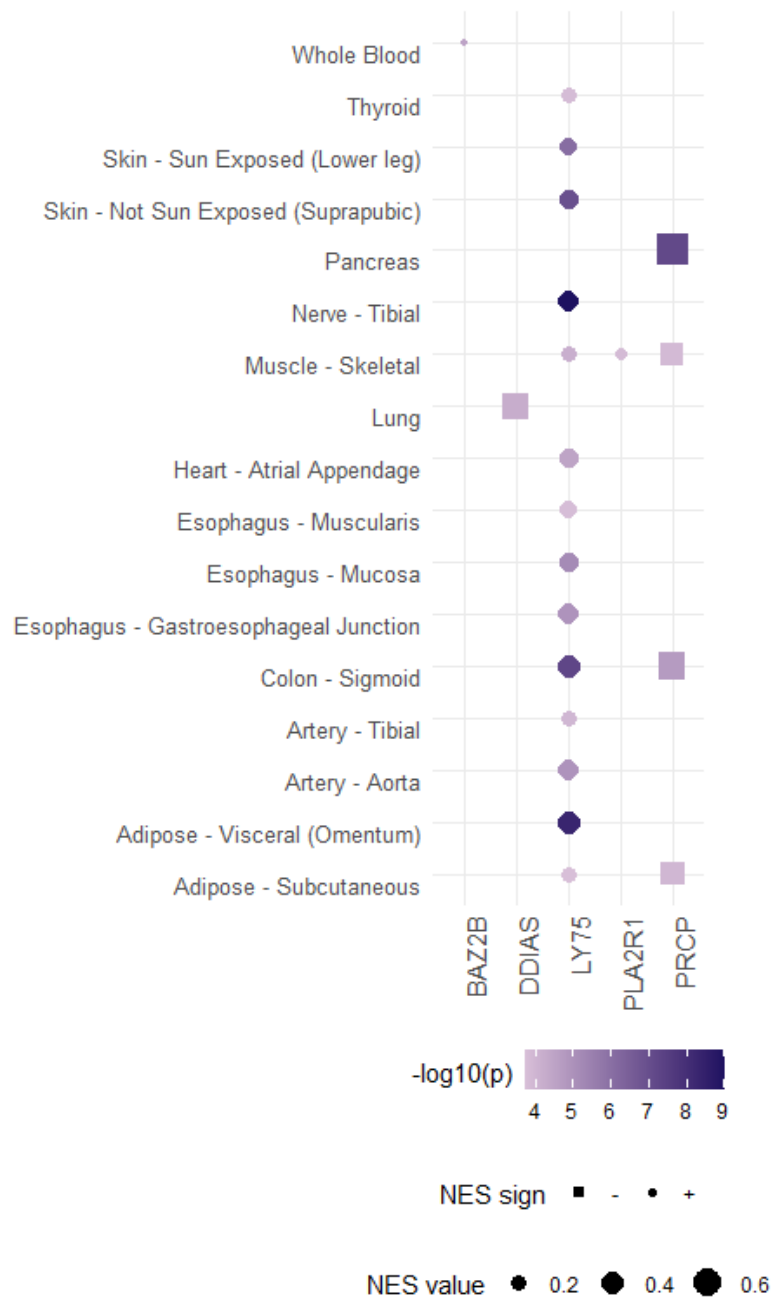
867

868

869

870 **Supplementary Figure 6. Gene-tissue pairs for which either rs1003835 or**  
871 **rs60606421 are significant eQTLs at FDR<0.05 (data retrieved from**  
872 **<https://gtexportal.org/home/snp/>). rs1003835 (chromosome 2) maps to *BAZ2B*, *LY75*,**  
873 **and *PLA2R* genes. As for the lead variant of chromosome 11, rs77599934, since it was**  
874 **not an eQTL, we used an LD proxy variant (rs60606421). *DDIAS* and *PRCP* genes map**  
875 **closely to this variant. NES and p-values correspond to the normalized effect size (and**  
876 **direction) of eQTL-gene associations and the p-value for the tissue, respectively.**

877



878

879

880

881