

1
2 **An atlas connecting shared genetic architecture of human diseases and**
3 **molecular phenotypes provides insight into COVID-19 susceptibility**
4

5 **Authors:**

6 Liuyang Wang¹, Thomas J. Balmat², Alejandro L. Antonia¹, Florica J. Constantine³,
7 Ricardo Henao³, Thomas W. Burke³, Andy Ingham², Micah T. McClain^{3,4,5}, Ephraim L.
8 Tsalik^{1,3,4,5}, Emily R. Ko^{3,6}, Geoffrey S. Ginsburg³, Mark R. DeLong², Xiling Shen⁷,
9 Christopher W. Woods^{3,4,5}, Elizabeth R. Hauser^{8,9}, and Dennis C. Ko^{1,5,10*}

10
11 **Affiliations**

12 ¹Department of Molecular Genetics and Microbiology, School of Medicine, Duke
13 University, Durham, NC 27710, USA

14 ² Duke Research Computing, Duke University, Durham, NC 27710, USA

15 ³Center for Applied Genomics and Precision Medicine, Department of Medicine, Duke
16 University, Durham, NC 27710, USA.

17 ⁴Durham Veterans Affairs Health Care System, Durham, NC 27705, USA.

18 ⁵Division of Infectious Diseases, Department of Medicine, Duke University Medical
19 Center, Durham, NC 27710, USA.

20 ⁶Department of Hospital Medicine, Duke Regional Hospital, Durham, NC, 27705, USA.

21 ⁷Department of Biomedical Engineering, Woo Center for Big Data and Precision Health,
22 Duke University, Durham, NC 27710, USA.

23 ⁸Duke Molecular Physiology Institute and Department of Biostatistics and Bioinformatics,
24 Duke University Medical Center Durham, NC 27710, USA

25 ⁹Cooperative Studies Program Epidemiology Center-Durham, Durham VA Health Care
26 System, Durham, NC 27705, USA

27 ¹⁰Lead contact

28 *To whom correspondence should be addressed: Dennis C. Ko, 0049 CARL Building
29 Box 3053, 213 Research Drive, Durham, NC 27710. 919-684-5834.

30 dennis.ko@duke.edu. @denniskoHiHOST

31

32

33 **Abstract**

34 While genome-wide associations studies (GWAS) have successfully elucidated the
35 genetic architecture of complex human traits and diseases, understanding mechanisms
36 that lead from genetic variation to pathophysiology remains an important challenge.
37 Methods are needed to systematically bridge this crucial gap to facilitate experimental
38 testing of hypotheses and translation to clinical utility. Here, we leveraged cross-
39 phenotype associations to identify traits with shared genetic architecture, using linkage
40 disequilibrium (LD) information to accurately capture shared SNPs by proxy, and
41 calculate significance of enrichment. This shared genetic architecture was examined
42 across differing biological scales through incorporating data from catalogs of clinical,
43 cellular, and molecular GWAS. We have created an interactive web database
44 (interactive Cross-Phenotype Analysis of GWAS database (iCPAGdb);
45 <http://cpag.oit.duke.edu>) to facilitate exploration and allow rapid analysis of user-
46 uploaded GWAS summary statistics. This database revealed well-known relationships
47 among phenotypes, as well as the generation of novel hypotheses to explain the
48 pathophysiology of common diseases. Application of iCPAGdb to a recent GWAS of
49 severe COVID-19 demonstrated unexpected overlap of GWAS signals between COVID-
50 19 and human diseases, including with idiopathic pulmonary fibrosis driven by the *DPP9*
51 locus. Transcriptomics from peripheral blood of COVID-19 patients demonstrated that
52 *DPP9* was induced in SARS-CoV-2 compared to healthy controls or those with bacterial
53 infection. Further investigation of cross-phenotype SNPs with severe COVID-19
54 demonstrated colocalization of the GWAS signal of the *ABO* locus with plasma protein
55 levels of a reported receptor of SARS-CoV-2, CD209 (DC-SIGN), pointing to a possible

56 mechanism whereby glycosylation of CD209 by *ABO* may regulate COVID-19 disease
57 severity. Thus, connecting genetically related traits across phenotypic scales links
58 human diseases to molecular and cellular measurements that can reveal mechanisms
59 and lead to novel biomarkers and therapeutic approaches.

60

61 **Keywords:** pleiotropy, cross-phenotype association, gout, LD-score, colocalization,
62 PheWAS, Hi-HOST, idiopathic pulmonary fibrosis, macular telangiectasia, rs2869462,
63 rs505922, rs12610495

64

65 **Introduction**

66 Genome-wide association studies (GWAS) have identified hundreds of
67 thousands of genomic regions that are associated with complex human traits and have
68 increased our understanding of the genetic architecture of human disease (Visscher et
69 al., 2017). While GWAS now utilize even millions of subjects through leveraging
70 electronic medical record data (Bycroft et al., 2018; McCarty et al., 2011), progress
71 towards understanding how identified genetic variants alter cellular function and
72 physiology remains elusive. More efficient mechanisms are needed for translating
73 knowledge of genetic disease risk and severity into insight of the underlying physiology.
74 Integrating analysis of GWAS across different scales of biological phenotypes
75 (molecular, cellular, and organismal) may provide novel insight into how genetic variants
76 influence complex traits.

77 Comparative analyses of GWAS have revealed that numerous, seemingly
78 unrelated traits are connected by shared underlying genetic variants (Visscher et al.,
79 2017). This phenomenon in which genetic variants affect multiple traits or diseases is
80 called pleiotropy. Several methods have been developed to study pleiotropic SNPs by
81 exploring the genetic relationship of multiple phenotypes. Broadly, these approaches
82 can be categorized into three major groups. The first method is genetic correlation,
83 which aims to quantify the similarity of the genetic effects on pairwise traits using GWAS
84 summary statistics such as LD-score regression (Bulik-Sullivan et al., 2015b) or from
85 individual genotype data with GCTA GREML (Lee et al., 2012). With large population
86 sizes, these methods can accurately partition variance into a shared genetic component
87 but do not reveal the genetic variants driving the genetic correlation. Genome-wide

88 cross-trait analysis (Zhu et al., 2018) has emerged as a means to follow up such results,
89 but these univariate meta-analyses of two traits requires genome wide summary
90 statistics for both traits, can suffer from effect size heterogeneity in combining results
91 from disparate traits, and cannot be easily applied to thousands of traits at once. The
92 second approach is colocalization, which estimates how well the GWAS signals from
93 two signals overlap in a given region while revealing plausibility of individual causal
94 variants (Giambartolomei et al., 2014). These two methods have successfully identified
95 novel genetic connections across distant traits as well as pleiotropic genomic regions
96 but have generally been used independently of each other. Finally, perhaps the most
97 intuitive approach, is quantifying cross-phenotype SNPs that are shared across multiple
98 phenotypes. In its simplest form, a phenome-wide association study takes a single SNP
99 and examines the significance of association across many traits, often from electronic
100 medical record (Denny et al., 2010). Valuable websites, including PhenoScanner
101 (Staley et al., 2016), GRASP (Leslie et al., 2014), and GeneATLAS (Canela-Xandri et
102 al., 2018) have integrated thousands of GWAS studies with billions of SNP-traits
103 associations and allow users to query individual SNPs across the phenome. However,
104 such PheWAS approaches do not leverage shared genetic architecture that extends
105 beyond individual SNPs and do not take advantage of LD information.

106 Motivated to simultaneously connect human phenotypes with shared genetic
107 architecture and to identify the precise loci driving this similarity, we previously
108 developed a method, CPAG (Cross-phenotype Analysis of GWAS), which estimated
109 phenotype similarity of NHGRI-EBI GWAS catalog traits based on shared genetic
110 associations (Wang et al. 2015). CPAG utilized cross-phenotype SNP associations to

111 cluster traits into groups that were consistent with pre-defined categories and
112 discovered novel pleiotropic SNPs connecting Crohn's disease and the fatty acid
113 palmitoleic acid. However, CPAG could not scale sufficiently to keep up with the
114 massive increase in the scope and scale of GWAS (facilitated through increasing use of
115 electronic medical record (EMR)-based GWAS of huge cohorts) and the deeper
116 phenotyping of molecular and cellular traits that can provide insight into mechanisms of
117 pathophysiology of disease. Here, we introduce iCPAGdb, a new cross-phenotype
118 analysis platform with improved identification of shared loci using pre-computed
119 ancestry-specific LD databases and a more efficient algorithm for capturing cross-
120 phenotype associations. These improvements facilitated integration of the NHGRI-EBI
121 GWAS catalog with large datasets of plasma and urine metabolites and cellular host-
122 pathogen traits. Such integration of pleiotropic analyses using GWAS datasets that
123 include intermediate traits across biological scales are crucial for moving from lists of
124 associated SNPs to understanding the pathophysiology of complex diseases. Finally,
125 iCPAGdb allows users to upload their own GWAS summary statistics via web interface
126 (<http://cpag.oit.duke.edu>) to identify and explore shared SNPs between their own
127 GWAS and a deep catalog of 4418 molecular, cellular, and disease phenotypes. Using
128 a GWAS of severe COVID-19 as the querying phenotype in iCPAGdb revealed shared
129 SNPs associated with idiopathic pulmonary fibrosis and plasma protein levels of CD209,
130 a possible receptor for SARS-CoV-2.

131

132 **Results**

133 ***iCPAGdb: An atlas for discovery of cross-phenotype associations***

134 We created iCPAGdb to facilitate exploration of cross-phenotype associations of
135 human phenotypes and discovery of shared genetics connecting traits that were
136 previously not known to be related. iCPAGdb utilizes 85639 SNP-trait associations ($p <$
137 5×10^{-8}) across 3793 traits from the NHGRI-EBI GWAS catalog, incorporates additional
138 GWAS datasets (see below and Table 1), and allows for uploading and analysis of user
139 GWAS summary statistics (Fig. 1A). In contrast, the original CPAG (published in 2015)
140 used only 14198 SNP-trait associations for 887 traits from the NHGRI-EBI GWAS
141 catalog.

142 Beyond this large expansion in traits and associations, we improved on the
143 original CPAG algorithm by clumping GWAS data from each study (Fig. S1), creating a
144 database of LD values based on 1000 Genomes (Genomes Project et al., 2015),
145 allowing selection of either European, African, or Asian LD structure, and efficiently
146 capturing cross-phenotype associations that are driven by LD proxy (Fig. 1B). For each
147 trait pair, iCPAGdb first selects the lead SNPs from all associated loci at a selected p -
148 value threshold ($p < 5 \times 10^{-8}$ was used for analysis of the NHGRI-EBI GWAS catalog;
149 Table S1; Fig. S2). These lead SNPs are compared across the trait pair to count directly
150 shared SNPs. For SNPs that are not directly shared, iCPAGdb then checks an LD
151 database for overlap by LD proxy. For all directly or indirectly shared SNPs, iCPAGdb
152 further forms them into bigger SNP blocks by recursively merging them until each SNP
153 block has no LD proxy with $R^2 \geq 0.4$ against all others. iCPAGdb improves memory
154 efficiency with built-in functions connecting to SQL GWAS and LD proxy databases and
155 improves computational efficiency and speed by utilizing multiple CPUs. For the
156 NHGRI-EBI GWAS Catalog, the growth of GWAS findings and improvements of

157 iCPAGdb over the previous version of CPAG led to a 27.7-fold increase in direct cross-
158 phenotype associations and a 47.7-fold increase in indirect cross-phenotype
159 associations, many of which would have been missed by the original CPAG algorithm
160 (Fig. 1C, D). Indeed, analyzing the 2013 NHGRI-EBI GWAS catalog with iCPAGdb had
161 little effect on direct associations but increased indirect associations by 76% (Fig. S3).

162 Results of iCPAGdb are consistent with results from the orthogonal approach of
163 genetic correlation by LD score regression (Bulik-Sullivan et al., 2015b). Comparing the
164 absolute values for genetic correlation of 24 phenotypes from (Bulik-Sullivan et al.,
165 2015a) against a similarity index quantifying the degree of shared SNPs in iCPAGdb
166 revealed that the two are significantly correlated ($p=3.52 \times 10^{-8}$; $R^2 = 0.14$) (Fig. 1E).
167 Nearly all phenotypes (64 of 70) that showed significant correlation by LD score
168 regression also demonstrated a significant excess of shared SNPs in iCPAGdb. The
169 output of iCPAGdb provides the SNPs driving the similarity between the two phenotypes,
170 facilitating follow-up studies. Interestingly, 61% of pairwise comparisons that had
171 significant overlap based on iCPAGdb did not have significant genetic correlation based
172 on LD-score regression. For example, LD-score regression did not detect significant
173 genetic correlation between LDL and HDL cholesterol measurements, but iCPAGdb
174 detected 92 shared SNPs, including 31 by direct overlap where the two phenotypes
175 have the same lead SNPs ($p=7.55 \times 10^{-195}$ by Fisher's exact test; $p=1.49 \times 10^{-190}$ after
176 Benjamini-Hochberg procedure. P-values from iCPAGdb in the remainder of the paper
177 are FDR-corrected for all pairwise comparisons using Benjamini-Hochberg procedure).

178

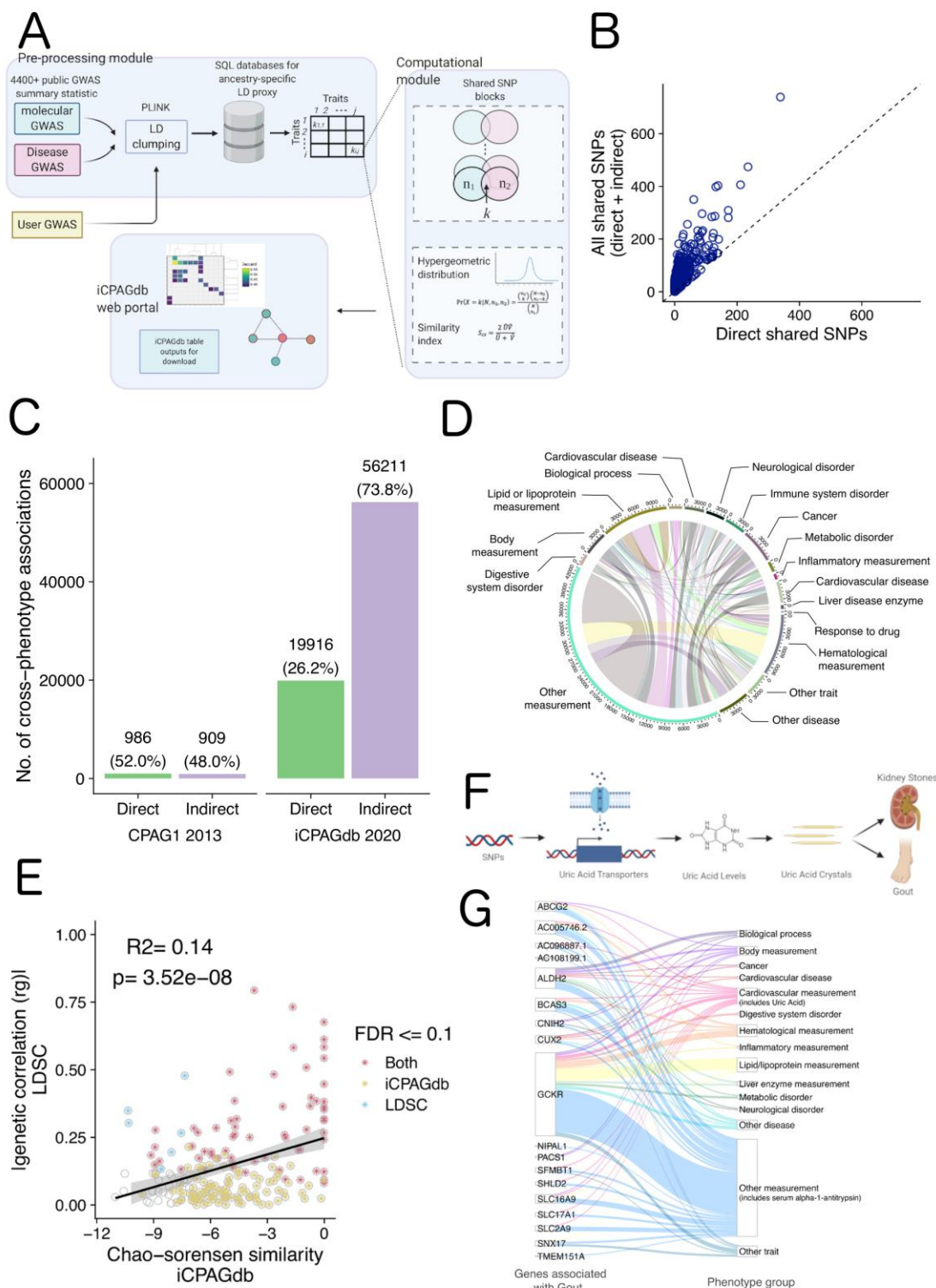
179

180 **Table 1. A summary of GWAS data in iCPAGdb. GWAS summary statistics were**
 181 **clumped to include only a lead SNP for each trait locus.**
 182

	Type	Traits/Diseases #	SNPs ($p < 5e-8$)	Trait-SNP associations #	Urls
NHGRI Catalog	Clinical GWAS	3793	63933	85639	https://www.ebi.ac.uk/gwas/
H2P2	Molecular/ cellular GWAS	79 (44 flow cytometric phenotypes + 35 cytokines)	17	3489 ($p < 1e-5$)	http://h2p2.oit.duke.edu
Blood Metabolites	Molecular GWAS	491 Blood (453 metabolites + 38 xenobiotics)	1441	2024	http://metabolomics.helmholtz-muenchen.de/gwas/
Urine Metabolites	Molecular GWAS	55 Urine	149	171	http://metabolomics.helmholtz-muenchen.de/gwas/
Sum		4418	65540	91323	

183

184



185

186 **Figure 1. An improved method for finding shared genetic architecture of human**187 **traits.**

188 **(A)** The overall framework of the iCPAGdb pipeline. GWAS summary statistics (from
189 published GWAS datasets or from user-uploaded GWAS) undergo LD clumping to
190 obtain a lead variant for each signal below a specified p-value threshold. These SNPs
191 are queried against an LD proxy database generated from 1000 Genomes African,
192 Asian, or European population to identify cross-phenotype associations through direct
193 overlap or LD proxy at $R^2 > 0.4$. Significance of overlap for each trait pair is calculated
194 using Fisher's exact test. Outputs can be visualized/downloaded from the iCPAGdb web
195 browser.

196 **(B)** Comparison of the number of shared SNPs for each NHGRI-EBI GWAS catalog
197 trait pair identified through direct overlap vs. both direct and indirect (LD-proxy) overlap.

198 **(C)** iCPAGdb detected more significant cross-phenotypes associations than CPAG1
199 at $FDR < 0.1$. Expansion of the NHGRI-EBI GWAS catalog and improvements in
200 capturing by LD proxy in iCPAGdb fueled a large increase in detected cross-phenotype
201 associations across human traits. Comparisons between CPAG1 and iCPAGdb on the
202 same 2013 dataset are in Fig. S3.

203 **(D)** Circle plot of cross-phenotype associations detected by iCPAGdb in the NHGRI-
204 EBI GWAS catalog. After excluding compound phenotypes (phenotypes described by
205 NHGRI-EBI GWAS catalog as > 1 comma-separated phenotype in their ontology), a
206 total of 1709 traits involved in a total of 53314 cross-phenotype associations were left.
207 These were categorized into 17 EFO Parental groups. Inner ribbons link phenotypes
208 connected by cross-phenotype associations with the width of ribbon corresponding to
209 the number of cross-phenotype associations. The axis outside the circle represents the
210 cumulative number of associations for each group vs all other groups.

211 **(E)** Comparison of genetic correlation from LD score regression (LDSC) and the
212 Chao-Sorensen similarity index implemented in iCPAG demonstrates significant
213 correlation. The genetic correlation r_g of 24 diseases/trait were obtained from (Bulik-
214 Sullivan et al., 2015a). Since Chao-Sorensen values are bounded from 0 to 1 and r_g
215 ranges from -1 to 1, we used the absolute value of r_g here. Colored * indicates
216 significant trait-pair for LDSC, iCPAGdb, or both at false discovery rate of 0.1.

217 **(F)** A model demonstrating how SNPs regulate uric acid levels to impact the
218 development of kidney stones and gout.

219 **(G)** Riverplot of gout cross-phenotype associations generated from iCPAGdb output
220 shows causal connections, comorbid outcomes, and regulators of disease. Mapped
221 genes for SNPs associated with gout are shown on the left and connected to other
222 NHGRI-EBI GWAS phenotypes grouped by EFO on the right.

223

224 ***GWAS of varying phenotypic scales reveals shared genetic architecture***
225 ***connecting molecular and cellular traits with human disease***

226 In a previous study (Wang et al., 2015), we defined 4 categories of cross-
227 phenotype associations: 1) SNP similarity between an intermediate trait/risk factor and
228 disease, 2) SNP similarity between a disease and a consequence of disease, 3) SNP
229 similarity between two traits affected by the same gene/pathway, and 4) SNP similarity
230 between two traits affected by the same gene having effects in different tissues or on
231 different pathways. Of these categories, perhaps the most clinically useful is the first
232 category—shared SNPs that connect an intermediate trait to a disease may reveal how
233 molecular or cellular phenotypes mediate some aspect of the pathophysiology of

234 disease. While the NHGRI-EBI GWAS catalog is comprised primarily of case-control
235 GWAS of disease, we detected numerous known shared associations linking a human
236 disease with levels of a metabolite. Metabolites are the substrates, intermediates, and
237 products of cellular metabolism and are routinely already used as biomarkers, such as
238 measuring glucose in diabetes management.

239 Cross-phenotype associations involving the metabolite uric acid and gout, an
240 inflammatory arthritis driven by excess levels of uric acid (Bodofsky et al., 2020), are
241 illustrative of iCPAGdb's usefulness. GWAS studies have been conducted on risk of
242 gout (Chen et al., 2018; Lai et al., 2012; Lee et al., 2019; Li et al., 2015; Matsuo et al.,
243 2016; Nakayama et al., 2017; Nakayama et al., 2020; Sulem et al., 2011) as well as uric
244 acid or urate levels (Boocock et al., 2020; Dehghan et al., 2008; Doring et al., 2008;
245 Kamatani et al., 2010; Kottgen et al., 2013; Li et al., 2007; Tin et al., 2019; Tin et al.,
246 2011). Notably, of 31 GWAS loci for gout and 123 GWAS loci for serum uric acid levels
247 at $p < 5 \times 10^{-8}$, 13 loci overlap, including 9 loci identified only by LD proxy (nearly 6000-fold
248 enrichment; $p = 5.9 \times 10^{-43}$). These loci are spread across 7 chromosomes and include
249 several solute carrier (SLC) and ATP-binding class (ABC) transporters that control urate
250 absorption and secretion. Some of the loci are in close proximity but are counted
251 separately by iCPAGdb, as could occur if different GWAS studies locate nearby peaks
252 that fall below our $R^2 > 0.4$ threshold or if multiple causal signals are located in the same
253 region. These data provide genetic evidence for the well-known causal role of excess
254 uric acid in gout and further reveal multiple genes that may serve as therapeutic targets.
255 Inhibitors of renal uric acid reabsorption through URAT1 (*SLC22A12*) are commonly
256 used in treating gout (Dong et al., 2019), but additional transporters implicated through

257 human genetics may also prove to be useful drug targets. Beyond uric acid levels,
258 GWAS of kidney stones (Howles et al., 2019; Oddsson et al., 2015; Thorleifsson et al.,
259 2009), a second manifestation of elevated uric acid levels, also share associated SNPs
260 with gout (3 shared loci, all identified by proxy on chromosomes 2, 4, and 17; $p=5.2 \times 10^{-9}$).
261 Finally, gout shares 2 loci (out of 5 from (Setoh et al., 2015; Suhre et al., 2017)) with
262 levels of serum alpha-1-antitrypsin, an anti-inflammatory endogenous protease inhibitor
263 ($p=9.3 \times 10^{-7}$), providing a human genetic rationale for the use of alpha-1-antitrypsin-
264 based therapeutics in acute gouty flares (as has been demonstrated to be efficacious in
265 mice (Joosten et al., 2016)). Thus, examining the gout cross-phenotype associations
266 revealed causal connections, comorbid conditions with shared etiology, and factors that
267 modulate inflammation in the disease (Fig. 1F, G).

268 Shared genetic associations reveal other well-known molecular and cellular
269 disease relationships such as LDL cholesterol levels with cardiovascular disease
270 (1.24×10^{-81}) and Alzheimer's disease ($p=4.8 \times 10^{-17}$) as well as glucose with type II
271 diabetes mellitus ($p=1.5 \times 10^{-40}$). Other cross-phenotype associations highlight genetic
272 variation that can extend our knowledge. For example, cross-phenotype associations
273 were found between malaria (Band et al., 2013; Jallow et al., 2009; Malaria Genomic
274 Epidemiology, 2019; Malaria Genomic Epidemiology et al., 2015; Ravenhall et al., 2018;
275 Timmann et al., 2012) and red blood cell distribution width (Astle et al., 2016; Chen et
276 al., 2020; Chen et al., 2013; Fatumo et al., 2019; Kichaev et al., 2019) ($p=1.3 \times 10^{-9}$). This
277 overlap is driven by well-known genetic variation in the beta-hemoglobin gene (*HBB*)
278 and *ABO* blood type affecting malaria risk but also by genetic variation in *ATP2B4* which
279 encodes a calcium transporter. To the best of our knowledge, whether size of red blood

280 cells impacts susceptibility to malaria parasites has not been examined. These cross-
281 phenotype associations demonstrate the promise of this approach for revealing novel
282 relationships that can be mined through iCPAGdb.

283

284 ***Expansion of iCPAGdb to additional datasets of molecular and cellular traits***

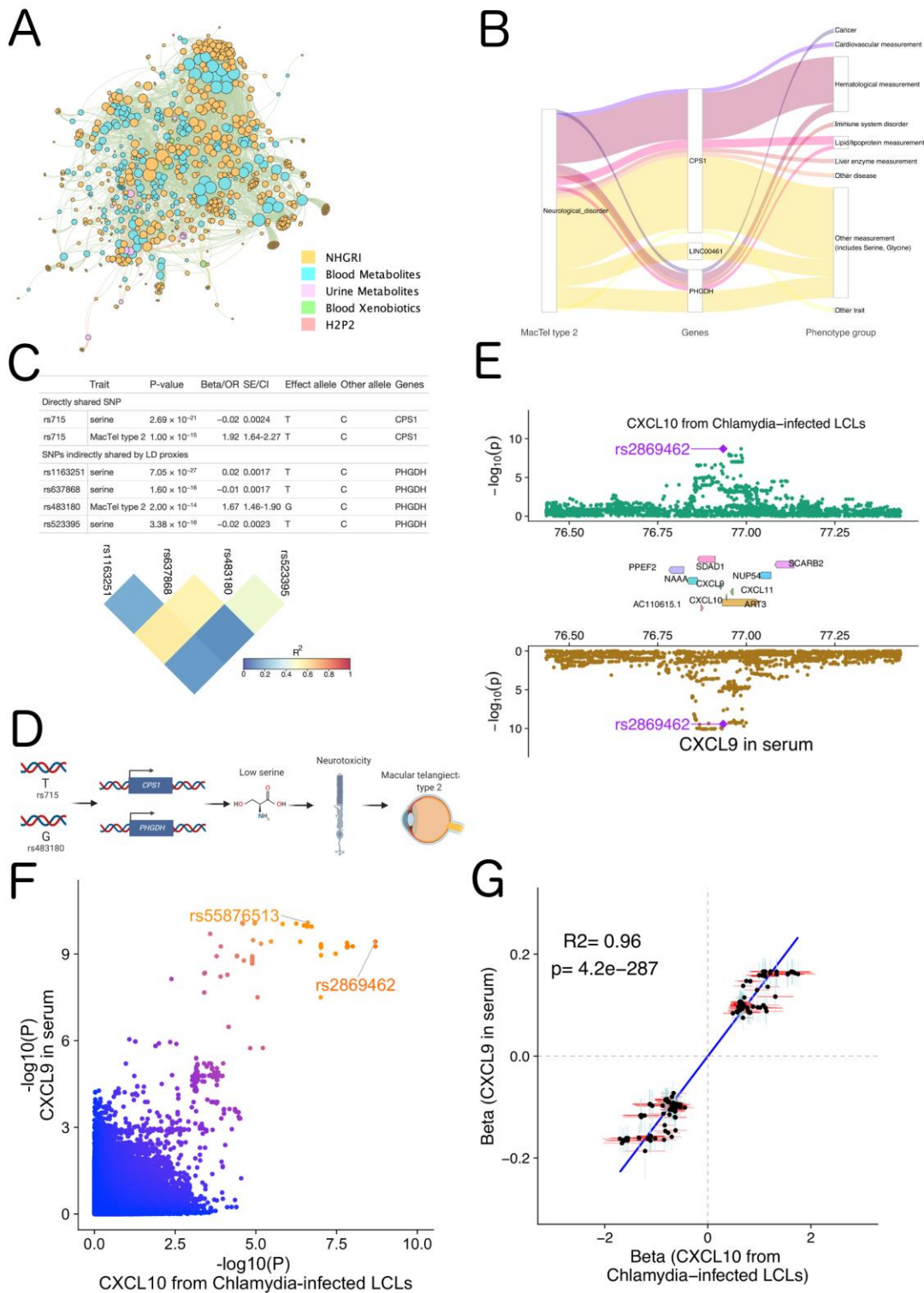
285 The above examples of clinically relevant cross-phenotype associations involving
286 metabolite and cellular phenotypes motivated expansion of iCPAGdb to additional
287 datasets. We used three datasets to provide molecular and cellular traits to our analysis:
288 491 metabolites and xenobiotics in blood (Shin et al., 2014) and 55 metabolites in urine
289 (Raffler et al., 2015), both from the Metabolomics GWAS Server
290 (<http://metabolomics.helmholtz-muenchen.de/gwas/index.php>), and 79 cellular host-
291 pathogen interaction traits from our dataset of cellular host-pathogen interaction GWAS,
292 H2P2 (Wang et al., 2018). iCPAGdb revealed many connections between these
293 molecular/cellular datasets and the NHGRI-EBI GWAS catalog (Fig. 2A; Table S2).

294 Cross-phenotype associations with macular telangiectasia (MacTel) type 2, a
295 disease characterized by loss of central vision due to alterations in blood vessels in the
296 macula of the eye, confirmed the importance of the amino acid serine (Fig. 2B). A
297 GWAS of MacTel type 2 uncovered 3 genome-wide significant loci and the authors
298 noted that two of these loci were involved in serine/glycine metabolism, with the alleles
299 associated with low glycine and serine conferring increased risk of MacTel type 2 (Scerri
300 et al., 2017). The authors speculated that the low serine levels could lead to high levels
301 of ammonia and glutamate causing neurotoxicity and stress-induced angiogenesis
302 (Scerri et al., 2017). Gantner et al. have since provided evidence that low serine levels

303 result in elevated levels of deoxysphingolipids to trigger cell death in photoreceptors
304 (Gantner et al., 2019). iCPAGdb rediscovered the connection of two loci being shared
305 between serine in serum (measured by (Shin et al., 2014)) and risk of MacTel (Fig. 2C,
306 D; $p=4.0 \times 10^{-7}$; 99,010-fold enrichment). iCPAGdb also revealed 7 other serum
307 metabolites including glycine that shared an association with rs715 but not with the
308 second MacTel locus. While serine was not part of the urine metabolomics dataset,
309 iCPAGdb did detect overlap of glycine in urine and MacTel type 2 ($p=0.01$).

310 We also included host-pathogen traits from H2P2, a cellular GWAS we
311 previously carried out using 528 lymphoblastoid cell lines (LCLs) exposed to 7 different
312 pathogens (Wang et al., 2018). Notably, unlike the metabolomics datasets, H2P2
313 identified SNPs associated with traits at baseline and in response to stimuli. Further, as
314 pathogens have likely been drivers of human evolution (Fumagalli et al., 2011; Pittman
315 et al., 2016), comparing H2P2 to human disease GWAS may reveal unintended
316 consequences of past pandemics on the human genome. Previously, we reported
317 colocalization of a locus regulating CXCL10 levels following *Chlamydia trachomatis*
318 infection (rs2869462) and risk of inflammatory bowel disease (Wang et al., 2018).
319 iCPAGdb revealed shared genetic variants for this H2P2 phenotype and blood levels of
320 CXCL9 (MIG) (Ahola-Olli et al., 2017) (Fig. 2E; $p=0.04$). P-values for the two
321 associations are strongly correlated (Fig. 2F), and the effect size for SNPs associated
322 with both chemokines are significantly positive correlated (Fig. 2G). We utilized COLOC,
323 which uses a Bayesian framework to determine whether GWAS signals in the same
324 region are likely due to the same causal variant (Giambartolomei et al., 2014). The
325 posterior probability that both CXCL10 protein levels from cells and CXCL9 levels in

326 blood share the same causal variant is 0.90 (Table 2), with rs2869462 identified as the
327 most likely causal SNP (Table S3). The genes encoding these two chemokines are
328 adjacent to each other on chromosome 4, and this result points to variants regulating
329 expression of both genes that will make it challenging to disentangle their effects in
330 disease.
331



332

333 **Figure 2. iCPAGdb integrates GWAS of different scales to reveal biological**

334 **insight.**

335 **(A)** Multi-dataset network of cross-phenotype associations detected by iCPAGdb.
336 Phenotypes that demonstrated significant overlap ($FDR \leq 0.1$) are color-coded in the
337 indicated colors.

338 **(B)** Riverplot of macular telangiectasia type 2 (MacTel type 2) cross-phenotype
339 associations generated from iCPAGdb shows causal connections, comorbid outcomes,
340 and regulators of disease.

341 **(C)** Cross-phenotype associations connecting MacTel type 2 and serine. One locus
342 demonstrated direct SNP overlap (rs715). A second locus demonstrated indirect overlap
343 based on 4 SNPs in LD as visualized in the heatmap color-coded by LD.

344 **(D)** A model for how SNPs regulate serine levels to impact pathogenesis of MacTel
345 type 2 based on iCPAGdb and prior work described in the text.

346 **(E)** Regional Miami colocalization plot demonstrates a genetic locus that impacts
347 both CXCL10 level in lymphoblastoid cell lines following *Chlamydia trachomatis*
348 infection and CXCL9 (MIG) levels in whole blood.

349 **(F)** Comparison of $-\log_{10}(p \text{ value})$ for GWAS of CXCL10 following *Chlamydia*
350 *trachomatis* infection and levels of CXCL9 (MIG) in whole blood. The lead SNP in the
351 region for each phenotype is marked.

352 **(G)** Scatter plot demonstrates a highly positive correlation of the effect coefficients of
353 cellular CXCL10 after *Chlamydia trachomatis* infection and of SNPs associated with
354 blood CXCL9 levels. Each dot represents a SNP which has $p \text{ value} < 0.01$ for both
355 phenotypes. A total of 413 SNPs from a 4-mb window surrounding the leading SNP
356 rs2869462 was selected. The blue vertical or red horizontal bar shows the standard
357 error of the beta value for each SNP.

358 **Table 2. COLOC analysis output.** PP3 is the posterior probability for the model where
 359 the two traits have independent causal variants. PP4 is the posterior probability for the
 360 model where the two traits share a single causal variant.
 361

Trait1	Trait2	Locus	SNP #	PP3	PP4	PP3+PP4	PP4/PP3	Lead causal SNP
CXCL10 level after <i>Chlamydia</i> infection	Blood CXCL9 levels	CXCL10	1533	0.101	0.899	1.00	8.91	rs2869462
COVID-19	Plasma CD209 antigen level	ABO	56	0.0159	0.984	1.00	61.72	rs505922
COVID-19	idiopathic pulmonary fibrosis	DPP9	1233	0.00216	0.994	0.996	459.63	rs12610495

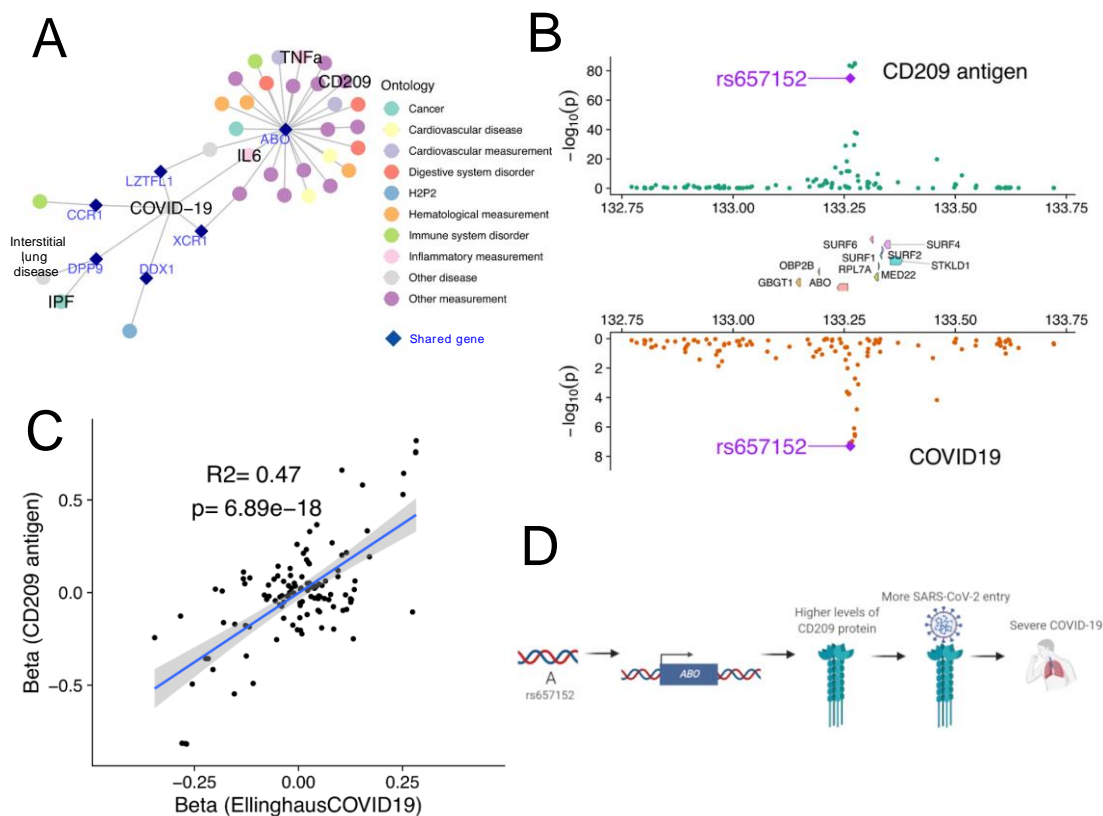
362

363 ***Application of iCPAGdb to COVID-19 reveals susceptibility due to ABO may occur***
 364 ***through regulation of CD209***

365 We applied iCPAGdb to a recently published GWAS of severe COVID-19 with
 366 respiratory failure (Ellinghaus et al., 2020). While this study focused on two genome-
 367 wide significant associations at the ABO locus and in a cluster of chemokine receptors
 368 and other genes on Chromosome 3, we relaxed the p-value threshold for iCPAGdb to
 369 1×10^{-5} , resulting in 24 suggestive loci after LD clumping. Not surprisingly, iCPAGdb
 370 revealed that the genome-wide significant association near the blood type locus *ABO* is
 371 in LD with multiple other SNPs in this region associated with other human diseases and
 372 traits (Fig. 3A; Table S4). This included the classic association with malaria resistance
 373 (Timmann et al., 2012), but also less well known associations with duodenal ulcer
 374 (Tanikawa et al., 2012), pancreatic cancer (Amundadottir et al., 2009), and heart failure
 375 (Shah et al., 2020). Multiple studies have now reported the association of the *ABO* locus
 376 with risk of COVID-19 (Ellinghaus et al., 2020; Zhao et al., 2020). The causal effect on
 377 COVID-19 may involve A and B antigens on blood cells, antibodies against A and B

378 antigens, the enzymatic activity of the ABO glycosyltransferase on possibly other
379 glycoproteins, or even other genes in the region. Insight into these possible
380 mechanisms was revealed by iCPAGdb, which identified association of this locus with
381 levels of 8 individual proteins in the NHGRI-EBI GWAS catalog. These proteins, all
382 encoded on different chromosomes than ABO, include IL-6, TNF- α , CD209 (DC-SIGN),
383 Tie-1, mannose-binding protein C, FGF23, and clotting factors (factor VIII and vWF). In
384 each of these cases, the association of the locus to both molecular trait and disease
385 provides a plausible causal chain from SNP to cis-effect on *ABO* to trans effect on a
386 protein to severe COVID-19 disease. For example, association with VWF and Factor
387 VIII may indicate ABO affects COVID-19 through regulation of thrombosis, as patients
388 with severe COVID-19 can have thromboembolic complications as part of a hyper-
389 inflammatory state (Wool and Miller, 2020). In fact, both VWF and factor VIII are targets
390 of glycosylation by ABO (Canis et al., 2018; Matsui et al., 1992; Sodetz et al., 1979) and
391 levels of these proteins are reported to be regulated by ABO (Albanez et al., 2016;
392 Gallinaro et al., 2008; Murray et al., 2020; Shima et al., 1995; Song et al., 2015). Further,
393 regulation of levels of IL-6 and TNF- α suggest possible regulation of inflammation, as
394 “cytokine storm” plays an important role during severe COVID-19 (Mangalmurti and
395 Hunter, 2020). Most interestingly, the *ABO* locus is associated with both COVID-19 and
396 CD209 ($p=0.008$). A preprint recently confirmed this association across populations, and
397 these authors speculated that ABO may affect CD209 levels to regulate SARS-CoV-2
398 entry (Katz et al., 2020). Indeed, there has since been evidence from two preprints that
399 CD209 can bind to SARS-CoV-2 and can act as a receptor for entry into immune cells
400 (Amraie et al., 2020; Chen et al., 2021).

401 The “A” allele of rs657152 associated with increased risk of COVID-19 with
 402 respiratory failure is also associated with increased levels of CD209 (Fig. 3B). We
 403 performed colocalization analysis of the GWAS signals for COVID-19 (Ellinghaus et al.,
 404 2020) and CD209 protein levels (Suhre et al., 2017). This analysis indicated the two are
 405 likely driven by the same causal variants (Fig. 3C; COLOC posterior probability PP4 =
 406 0.98 with the lead causal SNP of rs505922; Table S5). Thus, iCPAGdb and subsequent
 407 colocalization analysis support a model where *ABO* regulates CD209 protein levels to
 408 impact COVID-19 risk, though much future experimental and clinical studies will be
 409 required to fully test this hypothesis (Fig. 3D). The pleiotropic effects of *ABO* on levels of
 410 multiple proteins will make defining the mechanism challenging.



412 **Figure 3. Cross-phenotype association of *ABO* reveals a possible role for CD209**
413 **in severe COVID-19.**

414 **(A)** A network of genetic associations involving severe COVID-19. Each node represents
415 either a disease/trait (filled circles) or a gene (dark blue diamond). The *ABO* locus was
416 associated with multiple other diseases and levels of specific proteins, while DPP9 connects
417 COVID-19 only with IPF and interstitial lung disease (idiopathic interstitial pneumonia).

418 **(B)** Regional Miami colocalization plot demonstrates the *ABO* locus impacts both
419 CD209 protein levels and risk of severe COVID-19.

420 **(C)** A significant positive correlation for effect size of SNPs in the *ABO* locus on
421 CD209 protein levels and risk of severe COVID-19.

422 **(D)** Model of how *ABO* may affect CD209 and severe COVID-19.

423 ***Application of iCPAGdb to COVID-19 reveals a role for DPP9 in regulation of both***
424 ***COVID-19 and idiopathic pulmonary fibrosis***

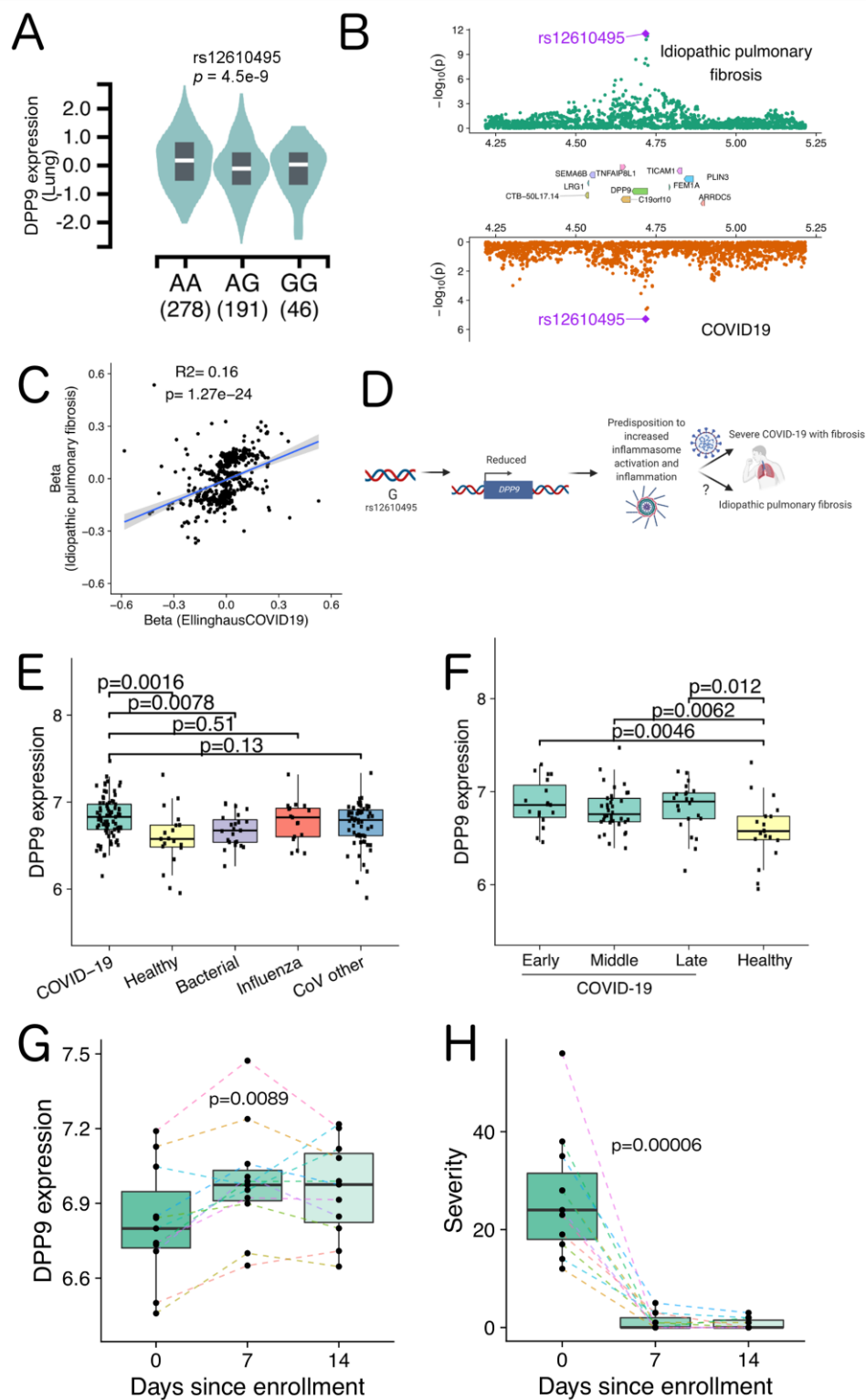
425 Beyond *ABO*, a locus in the dipeptidyl peptidase 9 (*DPP9*) gene associated at
426 $p < 1 \times 10^{-5}$ with severe COVID-19 was identified as being shared with a GWAS of fibrotic
427 idiopathic interstitial pneumonia (Fingerlin et al., 2013) and a recent GWAS of the most
428 severe form of that group of diseases, idiopathic pulmonary fibrosis (IPF) (Allen et al.,
429 2020). rs12610495 was the lead variant for each of these GWAS studies as well as the
430 suggestive peak for severe COVID-19 ($p = 5.2 \times 10^{-6}$; (Ellinghaus et al., 2020)). Much
431 evidence has already accumulated that pulmonary fibrosis is a hallmark of severe
432 COVID-19 (Ojo et al., 2020; Shi et al., 2020). While the association of rs12610495 with
433 COVID-19 did not reach genome-wide significance in Ellinghaus et al. 2020, this SNP is
434 in LD with the lead variant from a recent GWAS of critically ill COVID-19 patients that
435 does surpass genome-wide significance ($p = 3.98 \times 10^{-12}$; (Pairo-Castineira et al., 2020));

436 $R^2=0.95$ in 1000 Genomes European populations). Thus, iCPAGdb alerted us to the
437 importance of a suggestive COVID-19 susceptibility locus that has since been validated
438 in an independent cohort.

439 We determined that rs12610495 is an eQTL in lung tissue for the gene for *DPP9*
440 (and no other genes in the region) in GTEx ($p=4.5 \times 10^{-9}$; (Bao et al., 2015)), with the “G”
441 allele being associated with lower expression (Fig. 4A). Interestingly, *DPP9* is a
442 protease in the same family as *DPP4*, the receptor for MERS-coronavirus (Raj et al.,
443 2013). Additionally, *DPP9* is an inhibitor of inflammasome activation by NLRP3 (Okondo
444 et al., 2017; Okondo et al., 2018; Zhong et al., 2018). Colocalization analysis confirmed
445 the signals from severe COVID-19 and IPF are likely driven by the same causal variant
446 (Fig. 4B; COLOC posterior probability PP4 = 0.994, lead SNP rs12610495; Table S6).
447 Based on these data and the known biology, we developed alternative hypotheses for
448 how this SNP might be regulating risk of severe COVID-19: *DPP9* may be acting as a
449 previously unrecognized receptor for SARS-CoV-2 or it may be inhibiting inflammation
450 during COVID-19 infection. Based on the directionality of effect of rs12610495 on *DPP9*
451 gene expression, the “G” allele should lead to lower *DPP9* expression and less entry if
452 the receptor model is correct. However, the “G” allele is instead associated with
453 increased risk of severe COVID-19 (Fig. 4C). Alternatively, the “G” allele could lead to
454 lower *DPP9* to increase inflammasome activation in lung tissue, a model consistent with
455 “G” increasing risk of severe COVID-19 and this allele also increasing risk of idiopathic
456 pulmonary fibrosis (Fig. 4D).

457 To further examine the role of *DPP9* in COVID-19, we analyzed transcriptomics
458 of peripheral blood from COVID-19 patients (McClain et al., 2020). Levels of *DPP9*

459 expression across 46 COVID-19 patients were compared to individuals with seasonal
460 coronavirus, influenza, bacterial pneumonia, and healthy controls. *DPP9* levels were
461 significantly increased in COVID-19 patients compared to the other groups (fold-change
462 = 1.15, $p = 0.003$ adjusted by Benjamini-Hochberg method). Comparing COVID-19 data
463 vs. each comparator individually revealed that *DPP9* levels were elevated vs. healthy
464 controls ($p = 0.0016$) and bacterial infection ($p = 0.0078$) but not influenza or other
465 coronavirus infection (Fig. 4E). This data supports a role for *DPP9* in the host response
466 to viral infections. In examining all samples in the cohort, increased *DPP9* was observed
467 both early and late in COVID-19 infection (Fig. 4F). However, eleven subjects that did
468 not require hospitalization had repeated measurements at day 0 (initial enrollment into
469 the study), day 7, and day 14 that revealed changes in *DPP9* expression as infection
470 resolved. While *DPP9* expression increased from day 0 compared to 7 days and 14
471 days (Fig. 4G; $p = 0.0089$), symptom severity dramatically improved over this period
472 (Fig. 4H; $p = 0.00006$). We speculate that *DPP9* may be induced to effectively turn off
473 the inflammatory response to SARS-CoV-2 to minimize tissue damage and fibrosis.
474 Combined with our human genetic data, these findings suggest that insufficient
475 induction of *DPP9* expression could predispose to severe COVID-19.



476

477 **Figure 4. Cross-phenotype analysis and COVID-19 patient transcriptomics reveals**478 **a role for *DPP9* in severe COVID-19.**

- 479 **(A)** Lung eQTL data from GTEx shows rs12610495 “G” allele is associated with
480 reduced expression of *DPP9*.
- 481 **(B)** Regional Miami colocalization plot demonstrates the *DPP9* locus impacts both
482 idiopathic pulmonary fibrosis and risk of severe COVID-19.
- 483 **(C)** A significant positive correlation for effect size of SNPs in the *DPP9* locus on
484 idiopathic pulmonary fibrosis and risk of severe COVID-19.
- 485 **(D)** Model of how *DPP9* may affect idiopathic pulmonary fibrosis and risk of severe
486 COVID-19.
- 487 **(E)** *DPP9* expression in peripheral blood is significantly higher in COVID-19 patients
488 compared to healthy and bacteria-infected patients. The p values were calculated using
489 the Wilcoxon rank-sum test.
- 490 **(F)** COVID-19 patients demonstrate significantly higher *DPP9* expression compared
491 to healthy controls during early (days 1-10), middle (days 11-20) and late (21+ days)
492 stages of SARS-CoV-2 infection. The p values were calculated using the Wilcoxon rank-
493 sum test.
- 494 **(G)** *DPP9* demonstrates increased expression during recovery from COVID-19. A
495 total of 11 patients were measured sequentially at enrollment (day 0), day 7, and day
496 14. The colored dash line connects measurements from the same patient across time
497 points. P value was calculated using Friedman test.
- 498 **(H)** Decreased symptom severity scores of COVID-19 patients over time. The eleven
499 subjects in G were assessed for symptom severity at day 0, 7 and 14. The colored dash
500 line connects measurements from the same patient across time points. P value was
501 calculated using Friedman test.

502 ***Searching the iCPAGdb web server with user-provided GWAS summary statistics***

503 As the above examples demonstrate, iCPAGdb analysis can rapidly generate
504 hypotheses connecting molecular and cellular traits to human disease. The website
505 allows quick access to the pre-calculated cross-phenotype associations results
506 described in this manuscript. Users can also upload their own GWAS summary
507 statistics for comparing against all 4414 GWAS traits in the iCPAGdb website,
508 facilitating the discovery of new cross-phenotype relationships. Total time for uploading,
509 clumping of summary statistics, and calculation of cross-phenotype associations is
510 typically <2 minutes.

511

512 **Discussion**

513 The expansion of GWAS studies to more molecular, cellular, and human disease
514 traits requires the development and implementation of new tools to facilitate drawing
515 meaningful connections between phenotypes and understanding the molecular
516 mechanisms that explain this shared genetic architecture. Our work demonstrates that
517 leveraging available GWAS summary statistics and efficient algorithms of integrating
518 pleiotropic information using ancestry-specific LD structure can rapidly reveal cross-
519 phenotype associations across different phenotypic scales, which can be applied in
520 real-time to better understand ongoing health crises such as the SARS-CoV2 pandemic.

521 In examining cross-phenotype connections, it is important to carefully examine
522 the overlapping SNPs provided as part of the iCPAGdb output to determine 1) the
523 genome location where the variants are located, as some may be adjacent/overlapping
524 loci in weak LD and not truly distinct, and 2) how well identified GWAS signals from two

525 traits overlap. Indeed, we view iCPAGdb as the first step in a pipeline for gaining greater
526 understanding of any GWAS that then moves to colocalization analysis (see Fig. 2E, 3B,
527 4B; Table S3, S5, S6), to further dissect GWAS signals in the same region. Making
528 summary statistics more readily available for all GWAS, especially earlier studies in
529 NHGRI-EBI GWAS, would facilitate these validation studies. Finally, functional studies
530 in model systems and clinical studies are needed to test the proposed hypothesis and
531 deeply understand the underlying mechanisms.

532 While the current web implementation of iCPAGdb uses NHGRI-EBI GWAS
533 catalog (Welter et al., 2014), H2P2 (Wang et al., 2018), and metabolomics GWAS
534 datasets (Raffler et al., 2015; Shin et al., 2014), additional datasets of molecular,
535 cellular, and disease GWAS can be easily added. Analysis of user-uploaded GWAS
536 may be the most useful application of iCPAGdb and will lead to discovery of new
537 connections among human phenotypes to encourage experimental and clinical follow-
538 up studies. Our studies of COVID-19 provide a test case for this and revealed possible
539 mechanisms underlying the associations of severe COVID-19 with *ABO* and *DPP9*.

540 While our work highlights shared genetic architecture regulating *ABO*, protein
541 abundance, and COVID-19, much work remains to be done to understand the
542 mechanisms underlying these connections. The *ABO* locus controls abundance of many
543 proteins. Some of these proteins, such as VWF and Factor VIII, have already been
544 shown to be regulated by glycosylation of *ABO* (Canis et al., 2018; Matsui et al., 1992;
545 Sodetz et al., 1979). For CD209, *ABO* is a pQTL, but it is unknown whether CD209
546 protein abundance is regulated by *ABO* glycosylation. CD209 has a predicted N-linked
547 glycosylation site (N80) and glycosylation has been observed by mass spectrometry

548 (<http://glycositeatlas.biomarkercenter.org/glycosites/33001/>). Whether human genetic
549 variation also impacts CD209 glycosylation is also an unanswered question. Previous
550 studies have examined protein glycosylation as a GWAS trait, resulting in 16 genome-
551 wide significant loci (Huffman et al., 2011; Lauc et al., 2010; Sharapov et al., 2019), 15
552 of which have been recently replicated (Sharapov et al., 2020). However, these studies
553 quantified total plasma N-glycans released from proteins and did not specifically
554 quantify glycosylation and glycoforms for individual proteins. Future GWAS quantifying
555 individual glycosylated protein isoforms, as well as other post-translational modifications,
556 may therefore be valuable.

557 The shared underlying genetic risk factors for IPF and COVID-19 suggest that
558 *DPP9* may have a common role in pathogenesis in these diseases. iCPAGdb was able
559 to identify this connection in the first published COVID19 GWAS despite the *DPP9* allele
560 being below genome-wide significance in that cohort, demonstrating the utility of
561 iCPAGdb in expanding the power of GWAS studies on emerging and understudied
562 diseases. We speculate that characteristics of inflammasome-mediated responses,
563 normally suppressed by high expression of *DPP9*, may predispose to fibrosis. The
564 shared genetic architecture also suggests that therapeutic approaches targeting fibrosis
565 may be beneficial in both conditions. Pirfenidone and Nintedanib are anti-fibrotic FDA-
566 approved drugs used to treat IPF, and our findings support the idea that these drugs
567 may prove beneficial in COVID-19 (Ferrara et al., 2020; George et al., 2020; Seifirad,
568 2020). As our examination of COVID-19 demonstrates, iCPAGdb is a powerful
569 hypothesis engine that will lead to a deeper understanding of the genetic underpinnings
570 of human disease risk, severity, and drug response.

571 **Materials and Method**

572 ***Collection of GWAS summary statistics***

573 Publicly available GWAS summary statistics were downloaded from the following
574 sources: 3793 traits from NHGRI-EBI GWAS Catalog (version 1.02, downloaded on
575 2020/08/05), 79 traits from H2P2 cellular GWAS (Wang et al., 2018), 587 traits from
576 human blood circulating metabolites and urine metabolites GWAS (Raffler et al., 2015;
577 Shin et al., 2014). NHGRI-EBI GWAS catalog traits included annotation by Experimental
578 Factor Ontology (EFO). All GWAS data were harmonized to genome coordinates of
579 HG19. In total, we collected 4,225 GWAS traits, and 104,247 Trait-SNPs pairs at a p
580 value threshold of 5×10^{-8} . A detailed list of trait-SNP pairs at varying p-value
581 threshold can be found in Table 1.

582 Severe COVID-19 GWAS summary statistics were downloaded from the GRASP
583 website (<https://grasp.nhlbi.nih.gov/Covid19GWASResults.aspx>) (download date
584 2020/07/15). Genome coordinates were converted from GRCh38 to HG19 using UCSC
585 liftOver. GWAS summary statistics of IPF were kindly provided by Allen et al. 2020 after
586 requesting access <https://github.com/genomicsITER/PFgenetics>.

587

588 ***LD clumping***

589 GWAS summary statistics were individually pre-processed by LD clumping using *PLINK*
590 v1.9 (Chang et al., 2015), based on genotypes from European populations from the
591 1000 Genome project. The general PLINK command was “--clump-p1 1e-5 --clump-p2 1
592 --clump-r2 0.4 --clump-kb 1000”. For NHGRI/EBI GWAS catalog, the index SNPs were
593 selected using the genome-wide significant p value threshold of 5×10^{-8} (--clump-p1

594 5e-8). For molecular and cellular GWAS, we used a varying p-value cutoff from
595 1×10^{-3} to 1×10^{-5} for --clump-p1 parameter to choose the index SNPs.

596 For uploaded GWAS data, iCPAGdb calls on PLINK automatically to perform LD
597 clumping. Users can define the p value for --clump-p1 to select the index SNPs and
598 choose proper LD structure (European, African, or Asian) based on the ancestry of the
599 GWAS.

600

601 ***LD proxy calculation***

602 To maximize phenotypic associations due to indirect associations, pairwise LD R^2
603 values were computed for each leading SNP against its surrounding SNPs using the
604 genotypes from the 1000 Genome project (Phase 3 genotypes). Prior to calculation, all
605 SNPs with minor allele frequency less than 0.01 and missingness > 0.1 were removed.
606 R^2 of pairwise SNPs within 10,000 bp windows were then calculated, and only LD
607 proxies with $R^2 > 0.4$ were retained in further analysis. The PLINK parameters for
608 calculating LD was "--ld-window-kb 1000 --ld-window 10000 --keep-allele-order --r2 in-
609 phase with-freqs gz".

610 Since GWAS may be performed on diverse populations from different ancestry or
611 continents, we calculated ancestry-specific LD proxies for European, African, and Asian
612 populations separately. European population included 503 samples from 5 populations
613 (CEU, TSI, FIN, GBR, IBS), African included 661 samples from 7 populations (YRI,
614 LWK, GWD, MSL, ESN, ASW, ACB), and Asian population included 504 samples from
615 5 populations (CHB, JPT, CHS, CDX, and KHV). We filtered genotypes for each
616 ancestry population by minor allele frequency more than 0.01 and retained only biallelic

617 SNPs. SNPs which have same genome coordinates were merged using “—merge-
 618 equal-pos”. For duplicated SNPs with same variant rsID, we kept only the first variant by
 619 using “--rm-dup force-first” using PLINK 2.0,

620

621 ***Cross-phenotype SNP analysis***

622 Cross-phenotype SNPs were used to quantify the similarity of different traits. Cross-
 623 phenotype loci were identified as leading SNPs and/or their LD proxies having
 624 statistically significant associations with more than one trait/disease. If two traits shared
 625 a common leading SNP, we termed this “direct association”. If a leading SNP was
 626 associated with one trait, while its LD proxy SNPs were associated with another trait, we
 627 called this “indirect association”. If any shared SNP was in LD with another SNP with $R^2 >$
 628 0.4, these SNPs were merged into a SNP block until no further LD was found across
 629 shared SNP/LD pairs.

630 The significant association among each trait pair were using the hypergeometric
 631 distribution.

$$632 \quad p = \frac{\binom{n_2}{k} \binom{N_e - n_1 - n_2}{n_1}}{\binom{N_e}{n_1}}$$

633 Where N_e is the effective number of independent SNPs in the selected population, the
 634 n_1 and n_2 are the number of independent SNPs associated with trait 1 and trait2, and k
 635 is the number of independent SNPs blocks. The effective number of independent SNPs
 636 for European, African and Asian population were obtained from Table 4 from (Li et al.,
 637 2012).

638 The significance of associations for all trait pairs was further corrected for all possible
639 pairwise comparisons using the Benjamini-Hochberg and Bonferroni methods for
640 multiple test correction. A false discovery rate of 0.1 was chosen to identify significantly
641 correlated trait pairs.

642

643 ***Comparison to LDSC***

644 Bulik-Sullivan et al. (Bulik-Sullivan et al., 2015b) developed an innovative and unbiased
645 method, LDSC, to estimate genetic correlation using GWAS summary statistic for all
646 measured SNPs. Their model calculated the LD scores for a variant against all other
647 variants in a 1 centimorgan window and hypothesized that SNPs with higher LD scores
648 are tagged to a risk-conferring variant, and the genetic correlation among traits can be
649 calculated by normalizing genetic covariance of SNP heritability. With this method, they
650 estimated 276 genetic correlations for 24 diseases/traits based on full GWAS summary
651 statistic (Bulik-Sullivan et al., 2015a). To evaluate the power of iCPAGdb, we calculated
652 the genetic associations on the same 24 GWAS traits. For each trait pair, only SNPs
653 associated with each trait passing genome wide significant threshold (5×10^{-8}) were
654 used by iCPAGdb. We quantified the strength of cross-phenotype similarity for each trait
655 pair using Chao-Sorensen similarity index. Since the p values from (Bulik-Sullivan et al.,
656 2015a) were not corrected by multiple test correction, we calculated the p values for
657 r_g using R “p.adjust” function with a total number of 276 comparisons.

658

659 ***Colocalization analysis***

660 To evaluate whether the associations of GWAS trait pairs identified by iCPAGdb were
661 due to sharing the same causal variants, we performed colocalization analysis using R
662 COLOC packages (Giambartolomei et al., 2014). COLOC uses a Bayesian framework
663 to estimate the posterior probability that two GWAS traits share two independent causal
664 signals (PP3) or shares a single casual variant (PP4) in the selected genome region.
665 For each trait pair evaluated by COLOC, SNPs within 200 kb window from the lead SNP
666 were included. Since COLOC requires minor allele frequency (MAF) for each SNP in
667 both GWAS studies, when MAF was not available, we calculated the MAF using
668 European populations from the 1000 Genome Project. We ran COLOC “*coloc.abf*”
669 function using the default prior parameters, $p1 = 1 \times 10^{-4}$, $p2 = 1 \times 10^{-4}$, and $p12 = 1 \times 10^{-5}$.
670 We also ran built-in “*sensitivity*” function to evaluate the robustness of predefined priors,
671 and all tests suggested that default prior parameters are robust, therefore, we ran all
672 colocalization analyses with default priors values.

673

674 ***COVID-19 transcriptomic analysis***

675 As described in (McClain et al., 2020), samples were collected as part of the Molecular
676 and Epidemiological Study of Suspected Infection (MESSI) which was conducted at
677 Duke University Health System (DUHS) and the Durham Veterans Affairs Health Care
678 System (DVAHCS). The study was approved by each institution’s IRB. Informed
679 consent was obtained from all subjects or their legally authorized representatives, and
680 informed consent were collected for all subjects. SARS-CoV-2 RT-PCR testing was
681 used to confirm infection status. A total of 46 subjects were analyzed, 14 of which were
682 assayed at more than 1 timepoint. In total, 77 samples were assayed. Subjects were

683 divided into early (≤ 10 days), middle (11-21 days), and late (> 21 days) stage based on
684 duration of symptoms. Participant self-reported symptoms were recorded at each
685 timepoint for 39 symptom categories. Each symptom was scored on a scale of 0–4, with
686 0 indicating not present, 1 mild, 2 moderate, 3 severe, and 4 very severe symptoms.
687 Daily symptom severity (sum of symptom scores for all symptoms) was determined for
688 each timepoint. At enrollment (Day 0), date of symptom onset was determined, and an
689 initial symptom survey recorded maximum score for each symptom category between
690 symptom onset and study enrollment. Total RNA was extracted from peripheral whole
691 blood, and cDNA libraries prepared using NuGEN Universal Plus mRNA-seq with
692 AnyDeplete Globin reduction were sequenced on the Illumina NovaSeq 6000, as
693 described (McClain et al., 2020). In brief, STAR v 2.7.1 (Dobin et al., 2013) was used to
694 align the short reads and generate the count matrix. The count matrix was further
695 normalized using TMM method (Robinson and Oshlack, 2010) and log₂ transformed.
696 Associations were performed with generalized linear models (LIMMA, (Ritchie et al.,
697 2015)) and corrected for multiple testing using the Benjamini-Hochberg method
698 (Benjamini and Hochberg, 1995). Analysis of *DPP9* was carried out in R, and p values
699 were calculated using the Wilcoxon rank-sum test.

700

701 ***iCPAGdb software and website implementation***

702 iCPAGdb is comprised of two core parts, the back-end computation and the front-end
703 web browser. The back-end was written in python v3.6 with utilization of SQLite. SQLite
704 tables were constructed for harmonized GWAS datasets and LD tables for different
705 populations and are accessed using python sqlite3 package. The GWAS table stores

706 clumped GWAS summary statistic, including trait name, trait sources, SNPs' rsID, beta
707 values, standard error/standard deviation of beta, effective allele, and p values. The
708 ancestry-specific LD proxy tables contain pairwise SNPs' rsID and R^2 values ($R^2 \geq 0.4$)
709 for different populations. All SQLite tables were indexed on unique combinations of SNP
710 and trait or SNP pairs for LD proxy tables, which greatly reduces the searching time. To
711 further increase calculation speed, the core cross-phenotype analysis part of iCPAGdb
712 is parallelized by utilizing multiple threads.

713 Primary software components for the web portion of iCPAGdb are the R statistical
714 programming language (Team, 2020), the R package Shiny (v1.5.0) for interaction of
715 web pages with R scripts (Cheng et al., 2020), Shiny Server as a 24/7 multi-user
716 platform to make Shiny apps publicly accessible (RStudio, 2020), the database
717 environment SQLite for efficient querying of GWAS and CPAG results (Hipp, 2020),
718 and the R package RSQLite to execute SQL queries from within R scripts (Muller et al.,
719 2020). The results of a CPAG execution are read by the R script, processed, and
720 presented to the viewer in various tables and graphs on a web page. The iCPAGdb
721 website is currently loaded with associations across more than 4400 public GWAS
722 datasets that can be browsed and searched in "Review" mode. The user requests an
723 existing CPAG result set from which a corresponding table and heatmap are generated
724 and displayed. Various filtering and graph construction controls are available for
725 iterative sub-setting of data and selection of significance measure and number of top
726 significant phenotype pairs to plot. The "Download" button enables the researcher to
727 make a local copy of records appearing in the currently displayed results table.
728 Important packages used in this mode are DT for construction of and interaction with

729 tables and ggplot2, plotly, and heatmaply for basic plotting, interactive plotting (hover
730 labels), and heatmap generation, respectively. The web browser also allows users to
731 upload their own GWAS summary data, and iCPAGdb will automatically perform LD
732 clumping based on selected population and generate an atlas of connections for the
733 user's GWAS against > 4400 GWAS traits in the database. In this "Upload" mode, the
734 user browses files on a local computer, selects a properly formatted GWAS result file of
735 interest (containing, for a single phenotype, SNP rsIDs and GWAS p-values), specifies
736 format and column configuration, then uploads the file. Next, CPAG computation
737 parameter values, including iCPAGdb GWAS set to be crossed with, significance
738 thresholds for filtering, and linkage disequilibrium (LD) population are specified. When
739 "Compute CPAG" is pressed, the R script composes a system level command to
740 execute the CPAG (Python) function. The future() function of the R future package
741 (Bengtsson, 2020) combined with a delaying pipe from the promises package execute
742 CPAG operations asynchronously, waiting on completion before resuming R script
743 execution. Typical run time for a single uploaded GWAS that is already clumped to lead
744 variants is <30 seconds. For GWAS summary statistics including all SNPs in a study,
745 run time is typically < 2 minutes. The results are available with downloadable tables and
746 figures. Additional information on webapp is in Supplemental Note.

747

748 Web resources:

749 iCPAGdb: <http://cpag.oit.duke.edu>750 NHGRI GWAS Catalog: <https://www.ebi.ac.uk/gwas/>751 H2P2 cellular GWAS: <http://h2p2.oit.duke.edu>

752 Human metabolite GWAS summary statistics: <http://metabolomics.helmholtz->
753 [muenchen.de/gwas/index.php?task=download](http://metabolomics.helmholtz-muenchen.de/gwas/index.php?task=download)

754 COVID-19 GWAS summary statistics from Ellinghaus et al. (2020):
755 <https://grasp.nih.gov/Covid19GWASResults.aspx>

756 IPF GWAS: download link was obtained by applying for access following the
757 collaborative protocol from <https://github.com/genomicsITER/PFgenetics>

758

759 Tools for visualization:

760 R packages:

761 ggplot2: <https://cran.r-project.org/web/packages/ggplot2/>

762 gggene: <https://cran.r-project.org/web/packages/gggenes/index.html>

763 tidygraph: <https://cran.r-project.org/web/packages/tidygraph/>

764 ggnetwork: <https://cran.r-project.org/web/packages/ggnetwork/>

765 circlize: <https://cran.r-project.org/web/packages/circlize/>

766 ggpubr: <https://cran.r-project.org/web/packages/ggpubr/>

767 DT: <https://cran.r-project.org/web/packages/DT>

768 plotly: <https://cran.r-project.org/web/packages/plotly/>

769 heatmaply: <https://cran.r-project.org/web/packages/heatmaply/>

770 promises: <https://CRAN.R-project.org/package=promises>

771

772 Further information and requests for resources should be directed to and will be fulfilled
773 by the Lead Contact, Dennis C. Ko (dennis.ko@duke.edu). All iCPAGdb output
774 described in this manuscript are available for browsing from <http://cpag.oit.duke.edu>.

775 Supplemental files also contain iCPAGdb output and COLOC analysis results. Code is
776 available at GitHub <https://github.com/tbalmat/iCPAGdb>.

777

778 **Acknowledgements**

779 LW, TJB, AI, MRD, ERH, and DCK were supported by NIH R21AI133305. LW, ALA,
780 and DCK were supported by NIH R01AI118903. FJC, RH, TWB, MTM, XS, ELT, ERK,
781 and CWW were supported by DARPA/USMRAA W911NF1920111. We thank Benjamin
782 Schott, Jeffrey Bourgeois, and Kyle Gibbs for thoughtful discussions. We thank Dr.
783 Richard J. Allen and colleagues for sharing idiopathic pulmonary fibrosis GWAS
784 summary statistics.

785

786 **Author Contributions**

787 LW and DCK conceived of the study. LW, TJB, ERH, AI, MRD, ERH, and DCK
788 developed iCPAGdb. LW, TJB, FJC and RH carried out computational analysis. LW,
789 ALA, and DCK analyzed iCPAGdb results. MTM, FJC, RH, TWB, XS, GSG, ELT, ERK,
790 and CWW carried out the COVID-19 transcriptomics study and helped design
791 subsequent analysis carried out by LW. All authors contributed to the manuscript.

792

793 **Competing interests**

794 The author(s) declare no competing interests.

795

796 **References**

- 797 Ahola-Olli, A.V., Wurtz, P., Havulinna, A.S., Aalto, K., Pitkanen, N., Lehtimaki, T.,
798 Kahonen, M., Lyytikainen, L.P., Raitoharju, E., Seppala, I., *et al.* (2017). Genome-wide
799 Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines
800 and Growth Factors. *Am J Hum Genet* *100*, 40-50.
- 801 Albanez, S., Ogiwara, K., Michels, A., Hopman, W., Grabell, J., James, P., and Lillicrap,
802 D. (2016). Aging and ABO blood type influence von Willebrand factor and factor VIII
803 levels through interrelated mechanisms. *J Thromb Haemost* *14*, 953-963.
- 804 Allen, R.J., Guillen-Guio, B., Oldham, J.M., Ma, S.F., Dressen, A., Paynton, M.L.,
805 Kraven, L.M., Obeidat, M., Li, X., Ng, M., *et al.* (2020). Genome-Wide Association Study
806 of Susceptibility to Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* *201*, 564-
807 574.
- 808 Amraie, R., Napoleon, M.A., Yin, W., Berrigan, J., Suder, E., Zhao, G., Olejnik, J.,
809 Gummuluru, S., Muhlberger, E., Chitalia, V., *et al.* (2020). CD209L/L-SIGN and
810 CD209/DC-SIGN act as receptors for SARS-CoV-2 and are differentially expressed in
811 lung and kidney epithelial and endothelial cells. *bioRxiv*.
- 812 Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R.Z., Fuchs, C.S., Petersen, G.M.,
813 Arslan, A.A., Bueno-de-Mesquita, H.B., Gross, M., Helzlsouer, K., Jacobs, E.J., *et al.*
814 (2009). Genome-wide association study identifies variants in the ABO locus associated
815 with susceptibility to pancreatic cancer. *Nat Genet* *41*, 986-990.
- 816 Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman,
817 H., Riveros-Mckay, F., Kostadima, M.A., *et al.* (2016). The Allelic Landscape of Human
818 Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* *167*, 1415-1429
819 e1419.
- 820 Band, G., Le, Q.S., Jostins, L., Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F.,
821 Bojang, K., Pinder, M., Sirugo, G., *et al.* (2013). Imputation-based meta-analysis of
822 severe malaria in three African populations. *PLoS Genet* *9*, e1003509.
- 823 Bao, Y., Xu, S., Jing, X., Meng, L., and Qin, Z. (2015). De novo assembly and
824 characterization of *Oryza officinalis* leaf transcriptome by using RNA-seq. *Biomed Res*
825 *Int* *2015*, 982065.
- 826 Bengtsson, H. (2020). A unifying framework for parallel and distributed processing in r
827 using futures.
- 828 Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical
829 and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* *57*.
- 830 Bodofsky, S., Merriman, T.R., Thomas, T.J., and Schlesinger, N. (2020). Advances in
831 our understanding of gout as an auto-inflammatory disease. *Semin Arthritis Rheum* *50*,
832 1089-1100.
- 833 Boocock, J., Leask, M., Okada, Y., Asian Genetic Epidemiology Network, C., Matsuo,
834 H., Kawamura, Y., Shi, Y., Li, C., Mount, D.B., Mandal, A.K., *et al.* (2020). Genomic

- 835 dissection of 43 serum urate-associated loci provides multiple insights into molecular
836 mechanisms of urate control. *Human molecular genetics* 29, 923-943.
- 837 Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R.,
838 ReproGen, C., Psychiatric Genomics, C., Genetic Consortium for Anorexia Nervosa of
839 the Wellcome Trust Case Control, C., Duncan, L., *et al.* (2015a). An atlas of genetic
840 correlations across human diseases and traits. *Nat Genet* 47, 1236-1241.
- 841 Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia
842 Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L.,
843 and Neale, B.M. (2015b). LD Score regression distinguishes confounding from
844 polygenicity in genome-wide association studies. *Nat Genet* 47, 291-295.
- 845 Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A.,
846 Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The UK Biobank resource with
847 deep phenotyping and genomic data. *Nature* 562, 203-209.
- 848 Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations
849 in UK Biobank. *Nat Genet* 50, 1593-1599.
- 850 Canis, K., Anzengruber, J., Garenaux, E., Feichtinger, M., Benamara, K., Scheiflinger,
851 F., Savoy, L.A., Reipert, B.M., and Malisaukas, M. (2018). In-depth comparison of N-
852 glycosylation of human plasma-derived factor VIII and different recombinant products:
853 from structure to clinical implications. *J Thromb Haemost.*
- 854 Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015).
855 Second-generation PLINK: rising to the challenge of larger and richer datasets.
856 *GigaScience* 4, 7.
- 857 Chen, C.J., Tseng, C.C., Yen, J.H., Chang, J.G., Chou, W.C., Chu, H.W., Chang, S.J.,
858 and Liao, W.T. (2018). ABCG2 contributes to the development of gout and
859 hyperuricemia in a genome-wide association study. *Sci Rep* 8, 3137.
- 860 Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi,
861 B., Jiang, T., Akbari, P., Vuckovic, D., *et al.* (2020). Trans-ethnic and Ancestry-Specific
862 Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* 182, 1198-
863 1213 e1114.
- 864 Chen, Q., Gao, R., and Jia, L. (2021). Enhancement of the peroxidase-like activity of
865 aptamers modified gold nanoclusters by bacteria for colorimetric detection of
866 *Salmonella typhimurium*. *Talanta* 221, 121476.
- 867 Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y.,
868 Yanek, L.R., Keating, B., *et al.* (2013). Genome-wide association analysis of red blood
869 cell traits in African Americans: the COGENT Network. *Human molecular genetics* 22,
870 2529-2538.
- 871 Cheng, W., Cheng, J., Allaire, J.J., Xie, Y., and McPherson, J. (2020). Shiny: Web
872 Application Framework for R.
- 873 Dehghan, A., Kottgen, A., Yang, Q., Hwang, S.J., Kao, W.L., Rivadeneira, F.,
874 Boerwinkle, E., Levy, D., Hofman, A., Astor, B.C., *et al.* (2008). Association of three

- 875 genetic loci with uric acid concentration and risk of gout: a genome-wide association
876 study. *Lancet* 372, 1953-1961.
- 877 Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry,
878 K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS:
879 demonstrating the feasibility of a phenome-wide scan to discover gene-disease
880 associations. *Bioinformatics* 26, 1205-1210.
- 881 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P.,
882 Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner.
883 *Bioinformatics* 29, 15-21.
- 884 Dong, Y., Zhao, T., Ai, W., Zalloum, W.A., Kang, D., Wu, T., Liu, X., and Zhan, P.
885 (2019). Novel urate transporter 1 (URAT1) inhibitors: a review of recent patent literature
886 (2016-2019). *Expert Opin Ther Pat* 29, 871-879.
- 887 Doring, A., Gieger, C., Mehta, D., Gohlke, H., Prokisch, H., Coassin, S., Fischer, G.,
888 Henke, K., Klopp, N., Kronenberg, F., *et al.* (2008). SLC2A9 influences uric acid
889 concentrations with pronounced sex-specific effects. *Nat Genet* 40, 430-436.
- 890 Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P.,
891 Fernandez, J., Prati, D., Baselli, G., Asselta, R., *et al.* (2020). Genomewide Association
892 Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med*.
- 893 Fatumo, S., Carstensen, T., Nashiru, O., Gurdasani, D., Sandhu, M., and Kaleebu, P.
894 (2019). Complimentary Methods for Multivariate Genome-Wide Association Study
895 Identify New Susceptibility Genes for Blood Cell Traits. *Front Genet* 10, 334.
- 896 Ferrara, F., Granata, G., Pelliccia, C., La Porta, R., and Vitiello, A. (2020). The added
897 value of pirfenidone to fight inflammation and fibrotic state induced by SARS-CoV-2 :
898 Anti-inflammatory and anti-fibrotic therapy could solve the lung complications of the
899 infection? *Eur J Clin Pharmacol* 76, 1615-1618.
- 900 Fingerlin, T.E., Murphy, E., Zhang, W., Peljto, A.L., Brown, K.K., Steele, M.P., Loyd,
901 J.E., Cosgrove, G.P., Lynch, D., Groshong, S., *et al.* (2013). Genome-wide association
902 study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* 45, 613-620.
- 903 Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L., and Nielsen, R.
904 (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main
905 selective pressure through human evolution. *PLoS Genet* 7, e1002355.
- 906 Gallinaro, L., Cattini, M.G., Sztukowska, M., Padrini, R., Sartorello, F., Pontara, E.,
907 Bertomoro, A., Daidone, V., Pagnan, A., and Casonato, A. (2008). A shorter von
908 Willebrand factor survival in O blood group subjects explains how ABO determinants
909 influence plasma von Willebrand factor. *Blood* 111, 3540-3545.
- 910 Gantner, M.L., Eade, K., Wallace, M., Handzlik, M.K., Fallon, R., Trombley, J., Bonelli,
911 R., Giles, S., Harkins-Perry, S., Heeren, T.F.C., *et al.* (2019). Serine and Lipid
912 Metabolism in Macular Disease and Peripheral Neuropathy. *N Engl J Med* 381, 1422-
913 1433.

- 914 Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M.,
915 Korb, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., *et al.* (2015). A global
916 reference for human genetic variation. *Nature* 526, 68-74.
- 917 George, P.M., Wells, A.U., and Jenkins, R.G. (2020). Pulmonary fibrosis and COVID-
918 19: the potential role for antifibrotic therapy. *Lancet Respir Med* 8, 807-815.
- 919 Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace,
920 C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic
921 association studies using summary statistics. *PLoS Genet* 10, e1004383.
- 922 Hipp, R.D. (2020). SQLite.
- 923 Howles, S.A., Wiberg, A., Goldsworthy, M., Bayliss, A.L., Gluck, A.K., Ng, M., Grout, E.,
924 Tanikawa, C., Kamatani, Y., Terao, C., *et al.* (2019). Genetic variants of calcium and
925 vitamin D metabolism in kidney stone disease. *Nat Commun* 10, 5175.
- 926 Huffman, J.E., Knezevic, A., Vitart, V., Kattla, J., Adamczyk, B., Novokmet, M., Igl, W.,
927 Pucic, M., Zgaga, L., Johannson, A., *et al.* (2011). Polymorphisms in B3GAT1, SLC9A9
928 and MGAT5 are associated with variation within the human plasma N-glycome of 3533
929 European adults. *Human molecular genetics* 20, 5000-5011.
- 930 Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K.,
931 Bojang, K.A., Conway, D.J., Pinder, M., *et al.* (2009). Genome-wide and fine-resolution
932 association analysis of malaria in West Africa. *Nat Genet* 41, 657-665.
- 933 Joosten, L.A., Crisan, T.O., Azam, T., Cleophas, M.C., Koenders, M.I., van de
934 Veerdonk, F.L., Netea, M.G., Kim, S., and Dinarello, C.A. (2016). Alpha-1-anti-trypsin-
935 Fc fusion protein ameliorates gouty arthritis by reducing release and extracellular
936 processing of IL-1beta and by the induction of endogenous IL-1Ra. *Annals of the*
937 *rheumatic diseases* 75, 1219-1227.
- 938 Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y.,
939 and Kamatani, N. (2010). Genome-wide association study of hematological and
940 biochemical traits in a Japanese population. *Nat Genet* 42, 210-215.
- 941 Katz, D.H., Tahir, U.A., Ngo, D., Benson, M.D., Bick, A.G., Pampana, A., Gao, Y.,
942 Keyes, M.J., Correa, A., Sinha, S., *et al.* (2020). Proteomic Profiling in Biracial Cohorts
943 Implicates DC-SIGN as a Mediator of Genetic Risk in COVID-19. medRxiv.
- 944 Kichaev, G., Bhatia, G., Loh, P.R., Gazal, S., Burch, K., Freund, M.K., Schoech, A.,
945 Pasaniuc, B., and Price, A.L. (2019). Leveraging Polygenic Functional Enrichment to
946 Improve GWAS Power. *Am J Hum Genet* 104, 65-75.
- 947 Kottgen, A., Albrecht, E., Teumer, A., Vitart, V., Krumsiek, J., Hundertmark, C., Pistis,
948 G., Ruggiero, D., O'Seaghdha, C.M., Haller, T., *et al.* (2013). Genome-wide association
949 analyses identify 18 new loci associated with serum urate concentrations. *Nat Genet* 45,
950 145-154.
- 951 Lai, H.M., Chen, C.J., Su, B.Y., Chen, Y.C., Yu, S.F., Yen, J.H., Hsieh, M.C., Cheng,
952 T.T., and Chang, S.J. (2012). Gout and type 2 diabetes have a mutual inter-dependent

- 953 effect on genetic risk factors and higher incidences. *Rheumatology (Oxford)* 51, 715-
954 720.
- 955 Lauc, G., Essafi, A., Huffman, J.E., Hayward, C., Knezevic, A., Kattla, J.J., Polasek, O.,
956 Gornik, O., Vitart, V., Abrahams, J.L., *et al.* (2010). Genomics meets glycomics-the first
957 GWAS study of human N-Glycome identifies HNF1alpha as a master regulator of
958 plasma protein fucosylation. *PLoS Genet* 6, e1001256.
- 959 Lee, M.G., Hsu, T.C., Chen, S.C., Lee, Y.C., Kuo, P.H., Yang, J.H., Chang, H.H., and
960 Lee, C.C. (2019). Integrative Genome-Wide Association Studies of eQTL and GWAS
961 Data for Gout Disease Susceptibility. *Sci Rep* 9, 4981.
- 962 Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation
963 of pleiotropy between complex diseases using single-nucleotide polymorphism-derived
964 genomic relationships and restricted maximum likelihood. *Bioinformatics* 28, 2540-2542.
- 965 Leslie, R., O'Donnell, C.J., and Johnson, A.D. (2014). GRASP: analysis of genotype-
966 phenotype results from 1390 genome-wide association studies and corresponding open
967 access database. *Bioinformatics* 30, i185-194.
- 968 Li, C., Li, Z., Liu, S., Wang, C., Han, L., Cui, L., Zhou, J., Zou, H., Liu, Z., Chen, J., *et al.*
969 (2015). Genome-wide association analysis identifies three new risk loci for gout arthritis
970 in Han Chinese. *Nat Commun* 6, 7041.
- 971 Li, M.X., Yeung, J.M., Cherny, S.S., and Sham, P.C. (2012). Evaluating the effective
972 numbers of independent tests and significant p-value thresholds in commercial
973 genotyping arrays and public imputation reference datasets. *Human genetics* 131, 747-
974 756.
- 975 Li, S., Sanna, S., Maschio, A., Busonero, F., Usala, G., Mulas, A., Lai, S., Dei, M., Orru,
976 M., Albai, G., *et al.* (2007). The GLUT9 gene is associated with serum uric acid levels in
977 Sardinia and Chianti cohorts. *PLoS Genet* 3, e194.
- 978 Malaria Genomic Epidemiology, N. (2019). Insights into malaria susceptibility using
979 genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat Commun*
980 10, 5732.
- 981 Malaria Genomic Epidemiology, N., Band, G., Rockett, K.A., Spencer, C.C., and
982 Kwiatkowski, D.P. (2015). A novel locus of resistance to severe malaria in a region of
983 ancient balancing selection. *Nature* 526, 253-257.
- 984 Mangalmurti, N., and Hunter, C.A. (2020). Cytokine Storms: Understanding COVID-19.
985 *Immunity* 53, 19-25.
- 986 Matsui, T., Titani, K., and Mizuochi, T. (1992). Structures of the asparagine-linked
987 oligosaccharide chains of human von Willebrand factor. Occurrence of blood group A,
988 B, and H(O) structures. *J Biol Chem* 267, 8723-8731.
- 989 Matsuo, H., Yamamoto, K., Nakaoka, H., Nakayama, A., Sakiyama, M., Chiba, T.,
990 Takahashi, A., Nakamura, T., Nakashima, H., Takada, Y., *et al.* (2016). Genome-wide
991 association study of clinically defined gout identifies multiple risk loci and its association
992 with clinical subtypes. *Annals of the rheumatic diseases* 75, 652-659.

- 993 McCarty, C.A., Chisholm, R.L., Chute, C.G., Kullo, I.J., Jarvik, G.P., Larson, E.B., Li, R.,
994 Masys, D.R., Ritchie, M.D., Roden, D.M., *et al.* (2011). The eMERGE Network: a
995 consortium of biorepositories linked to electronic medical records data for conducting
996 genomic studies. *BMC Med Genomics* 4, 13.
- 997 McClain, M.T., Constantine, F.J., Henao, R., Liu, Y., Tsalik, E.L., Burke, T.W.,
998 Steinbrink, J.M., Petzold, E., Nicholson, B.P., Rolfe, R., *et al.* (2020). Dysregulated
999 transcriptional responses to SARS-CoV-2 in the periphery support novel diagnostic
1000 approaches. medRxiv.
- 1001 Muller, K., Wickham, H., James, D.A., and Falcon, S. (2020). RSQLite: 'SQLite'
1002 Interface for R.
- 1003 Murray, G.P., Post, S.R., and Post, G.R. (2020). ABO blood group is a determinant of
1004 von Willebrand factor protein levels in human pulmonary endothelial cells. *J Clin Pathol*
1005 73, 347-349.
- 1006 Nakayama, A., Nakaoka, H., Yamamoto, K., Sakiyama, M., Shaukat, A., Toyoda, Y.,
1007 Okada, Y., Kamatani, Y., Nakamura, T., Takada, T., *et al.* (2017). GWAS of clinically
1008 defined gout and subtypes identifies multiple susceptibility loci that include urate
1009 transporter genes. *Annals of the rheumatic diseases* 76, 869-877.
- 1010 Nakayama, A., Nakatochi, M., Kawamura, Y., Yamamoto, K., Nakaoka, H., Shimizu, S.,
1011 Higashino, T., Koyama, T., Hishida, A., Kuriki, K., *et al.* (2020). Subtype-specific gout
1012 susceptibility loci and enrichment of selection pressure on ABCG2 and ALDH2 identified
1013 by subtype genome-wide meta-analyses of clinically defined gout patients. *Annals of the*
1014 *rheumatic diseases* 79, 657-665.
- 1015 Oddsson, A., Sulem, P., Helgason, H., Edvardsson, V.O., Thorleifsson, G.,
1016 Sveinbjornsson, G., Haraldsdottir, E., Eyjolfsson, G.I., Sigurdardottir, O., Olafsson, I., *et*
1017 *al.* (2015). Common and rare variants associated with kidney stones and biochemical
1018 traits. *Nat Commun* 6, 7975.
- 1019 Ojo, A.S., Balogun, S.A., Williams, O.T., and Ojo, O.S. (2020). Pulmonary Fibrosis in
1020 COVID-19 Survivors: Predictive Factors and Risk Reduction Strategies. *Pulm Med*
1021 2020, 6175964.
- 1022 Okondo, M.C., Johnson, D.C., Sridharan, R., Go, E.B., Chui, A.J., Wang, M.S.,
1023 Poplawski, S.E., Wu, W., Liu, Y., Lai, J.H., *et al.* (2017). DPP8 and DPP9 inhibition
1024 induces pro-caspase-1-dependent monocyte and macrophage pyroptosis. *Nat Chem*
1025 *Biol* 13, 46-53.
- 1026 Okondo, M.C., Rao, S.D., Taabazuing, C.Y., Chui, A.J., Poplawski, S.E., Johnson, D.C.,
1027 and Bachovchin, D.A. (2018). Inhibition of Dpp8/9 Activates the Nlrp1b Inflammasome.
1028 *Cell Chem Biol* 25, 262-267 e265.
- 1029 Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A.D., Rawlik, K., Pasko, D.,
1030 Walker, S., Parkinson, N., Fourman, M.H., Russell, C.D., *et al.* (2020). Genetic
1031 mechanisms of critical illness in Covid-19. *Nature*.

- 1032 Pittman, K.J., Glover, L.C., Wang, L., and Ko, D.C. (2016). The Legacy of Past
1033 Pandemics: Common Human Mutations That Protect against Infectious Disease. *PLoS*
1034 *Pathog* *12*, e1005680.
- 1035 Raffler, J., Friedrich, N., Arnold, M., Kacprowski, T., Rueedi, R., Altmaier, E., Bergmann,
1036 S., Budde, K., Gieger, C., Homuth, G., *et al.* (2015). Genome-Wide Association Study
1037 with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary
1038 Human Metabolic Individuality. *PLoS Genet* *11*, e1005487.
- 1039 Raj, V.S., Mou, H., Smits, S.L., Dekkers, D.H., Muller, M.A., Dijkman, R., Muth, D.,
1040 Demmers, J.A., Zaki, A., Fouchier, R.A., *et al.* (2013). Dipeptidyl peptidase 4 is a
1041 functional receptor for the emerging human coronavirus-EMC. *Nature* *495*, 251-254.
- 1042 Ravenhall, M., Campino, S., Sepulveda, N., Manjurano, A., Nadjm, B., Mtove, G.,
1043 Wangai, H., Maxwell, C., Olomi, R., Reyburn, H., *et al.* (2018). Novel genetic
1044 polymorphisms associated with severe malaria and under selective pressure in North-
1045 eastern Tanzania. *PLoS Genet* *14*, e1007172.
- 1046 Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015).
1047 limma powers differential expression analyses for RNA-sequencing and microarray
1048 studies. *Nucleic Acids Res* *43*, e47.
- 1049 Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential
1050 expression analysis of RNA-seq data. *Genome Biol* *11*, R25.
- 1051 RStudio (2020). Shiny Server: Put Shiny Web Apps Online.
- 1052 Scerri, T.S., Quagliari, A., Cai, C., Zernant, J., Matsunami, N., Baird, L., Schepke, L.,
1053 Bonelli, R., Yannuzzi, L.A., Friedlander, M., *et al.* (2017). Genome-wide analyses
1054 identify common variants associated with macular telangiectasia type 2. *Nat Genet* *49*,
1055 559-567.
- 1056 Seifirad, S. (2020). Pirfenidone: A novel hypothetical treatment for COVID-19. *Med*
1057 *Hypotheses* *144*, 110005.
- 1058 Setoh, K., Terao, C., Muro, S., Kawaguchi, T., Tabara, Y., Takahashi, M., Nakayama,
1059 T., Kosugi, S., Sekine, A., Yamada, R., *et al.* (2015). Three missense variants of
1060 metabolic syndrome-related genes are associated with alpha-1 antitrypsin levels. *Nat*
1061 *Commun* *6*, 7754.
- 1062 Shah, S., Henry, A., Roselli, C., Lin, H., Sveinbjornsson, G., Fatemifar, G., Hedman,
1063 A.K., Wilk, J.B., Morley, M.P., Chaffin, M.D., *et al.* (2020). Genome-wide association
1064 and Mendelian randomisation analysis provide insights into the pathogenesis of heart
1065 failure. *Nat Commun* *11*, 163.
- 1066 Sharapov, S.Z., Shadrina, A.S., Tsepilov, Y.A., Elgaeva, E.E., Tiys, E.S., Feoktistova,
1067 S.G., Zaytseva, O.O., Vuckovic, F., Cuadrat, R., Jager, S., *et al.* (2020). Replication of
1068 fifteen loci involved in human plasma protein N-glycosylation in 4,802 samples from four
1069 cohorts. *Glycobiology*.
- 1070 Sharapov, S.Z., Tsepilov, Y.A., Klaric, L., Mangino, M., Thareja, G., Shadrina, A.S.,
1071 Simurina, M., Dagostino, C., Dmitrieva, J., Vilaj, M., *et al.* (2019). Defining the genetic

- 1072 control of human blood plasma N-glycome using genome-wide association study.
1073 Human molecular genetics 28, 2062-2077.
- 1074 Shi, H., Han, X., Jiang, N., Cao, Y., Alwalid, O., Gu, J., Fan, Y., and Zheng, C. (2020).
1075 Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a
1076 descriptive study. The Lancet infectious diseases 20, 425-434.
- 1077 Shima, M., Fujimura, Y., Nishiyama, T., Tsujiuchi, T., Narita, N., Matsui, T., Titani, K.,
1078 Katayama, M., Yamamoto, F., and Yoshioka, A. (1995). ABO blood group genotype and
1079 plasma von Willebrand factor in normal individuals. Vox Sang 68, 236-240.
- 1080 Shin, S.Y., Fauman, E.B., Petersen, A.K., Krumsiek, J., Santos, R., Huang, J., Arnold,
1081 M., Erte, I., Forgetta, V., Yang, T.P., *et al.* (2014). An atlas of genetic influences on
1082 human blood metabolites. Nat Genet 46, 543-550.
- 1083 Sodetz, J.M., Paulson, J.C., and McKee, P.A. (1979). Carbohydrate composition and
1084 identification of blood group A, B, and H oligosaccharide structures on human Factor
1085 VIII/von Willebrand factor. J Biol Chem 254, 10754-10760.
- 1086 Song, J., Chen, F., Campos, M., Bolgiano, D., Houck, K., Chambless, L.E., Wu, K.K.,
1087 Folsom, A.R., Couper, D., Boerwinkle, E., *et al.* (2015). Quantitative Influence of ABO
1088 Blood Groups on Factor VIII and Its Ratio to von Willebrand Factor, Novel Observations
1089 from an ARIC Study of 11,673 Subjects. PLoS One 10, e0132626.
- 1090 Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S.,
1091 Freitag, D., Burgess, S., Danesh, J., *et al.* (2016). PhenoScanner: a database of human
1092 genotype-phenotype associations. Bioinformatics 32, 3207-3209.
- 1093 Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath, H.,
1094 Thareja, G., Wahl, A., DeLisle, R.K., *et al.* (2017). Connecting genetic risk to disease
1095 end points through the human blood plasma proteome. Nat Commun 8, 14357.
- 1096 Sulem, P., Gudbjartsson, D.F., Walters, G.B., Helgadóttir, H.T., Helgason, A.,
1097 Gudjonsson, S.A., Zanon, C., Besenbacher, S., Bjornsdóttir, G., Magnusson, O.T., *et al.*
1098 (2011). Identification of low-frequency variants associated with gout and serum uric acid
1099 levels. Nat Genet 43, 1127-1130.
- 1100 Tanikawa, C., Urabe, Y., Matsuo, K., Kubo, M., Takahashi, A., Ito, H., Tajima, K.,
1101 Kamatani, N., Nakamura, Y., and Matsuda, K. (2012). A genome-wide association study
1102 identifies two susceptibility loci for duodenal ulcer in the Japanese population. Nat
1103 Genet 44, 430-434, S431-432.
- 1104 Team, R.C. (2020). R: A language and environment for statistical computing. R
1105 Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org/>.
- 1106 Thorleifsson, G., Holm, H., Edvardsson, V., Walters, G.B., Styrkarsdóttir, U.,
1107 Gudbjartsson, D.F., Sulem, P., Halldorsson, B.V., de Vegt, F., d'Ancona, F.C., *et al.*
1108 (2009). Sequence variants in the CLDN14 gene associate with kidney stones and bone
1109 mineral density. Nat Genet 41, 926-930.

- 1110 Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., Sievertsen, J., Muntau,
1111 B., Ruge, G., Loag, W., *et al.* (2012). Genome-wide association study indicates two
1112 novel resistance loci for severe malaria. *Nature* *489*, 443-446.
- 1113 Tin, A., Marten, J., Halperin Kuhns, V.L., Li, Y., Wuttke, M., Kirsten, H., Sieber, K.B.,
1114 Qiu, C., Gorski, M., Yu, Z., *et al.* (2019). Target genes, variants, tissues and
1115 transcriptional pathways influencing human serum urate levels. *Nat Genet* *51*, 1459-
1116 1474.
- 1117 Tin, A., Woodward, O.M., Kao, W.H., Liu, C.T., Lu, X., Nalls, M.A., Shriner, D., Semmo,
1118 M., Akylbekova, E.L., Wyatt, S.B., *et al.* (2011). Genome-wide association study for
1119 serum urate concentrations and gout among African Americans identifies genomic risk
1120 loci and a novel URAT1 loss-of-function allele. *Human molecular genetics* *20*, 4056-
1121 4068.
- 1122 Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and
1123 Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am*
1124 *J Hum Genet* *101*, 5-22.
- 1125 Wang, L., Oehlers, S.H., Espenschied, S.T., Rawls, J.F., Tobin, D.M., and Ko, D.C.
1126 (2015). CPAG: software for leveraging pleiotropy in GWAS to reveal similarity between
1127 human traits links plasma fatty acids and intestinal inflammation. *Genome Biol* *16*, 190.
- 1128 Wang, L., Pittman, K.J., Barker, J.R., Salinas, R.E., Stanaway, I.B., Williams, G.D.,
1129 Carroll, R.J., Balmat, T., Ingham, A., Gopalakrishnan, A.M., *et al.* (2018). An Atlas of
1130 Genetic Variation Linking Pathogen-Induced Cellular Traits to Human Disease. *Cell*
1131 *Host Microbe* *24*, 308-323 e306.
- 1132 Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A.,
1133 Flicek, P., Manolio, T., Hindorff, L., *et al.* (2014). The NHGRI GWAS Catalog, a curated
1134 resource of SNP-trait associations. *Nucleic Acids Res* *42*, D1001-1006.
- 1135 Wool, G.D., and Miller, J.L. (2020). The Impact of COVID-19 Disease on Platelets and
1136 Coagulation. *Pathobiology*, 1-13.
- 1137 Zhao, J., Yang, Y., Huang, H., Li, D., Gu, D., Lu, X., Zhang, Z., Liu, L., Liu, T., Liu, Y., *et*
1138 *al.* (2020). Relationship between the ABO Blood Group and the COVID-19
1139 Susceptibility. *Clin Infect Dis*.
- 1140 Zhong, F.L., Robinson, K., Teo, D.E.T., Tan, K.Y., Lim, C., Harapas, C.R., Yu, C.H.,
1141 Xie, W.H., Sobota, R.M., Au, V.B., *et al.* (2018). Human DPP9 represses NLRP1
1142 inflammasome and protects against autoinflammatory diseases via both peptidase
1143 activity and FIIND domain binding. *J Biol Chem* *293*, 18864-18878.
- 1144 Zhu, Z., Anttila, V., Smoller, J.W., and Lee, P.H. (2018). Statistical power and utility of
1145 meta-analysis methods for cross-phenotype genome-wide association studies. *PLoS*
1146 *One* *13*, e0193256.
- 1147