Genome Medicine

SOFTWARE Open Access

An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility

Liuyang Wang[1], Thomas J. Balmat[2], Alejandro L. Antonia[1], Florica J. Constantine[3], Ricardo Henao[3], Thomas W. Burke[3], Andy Ingham[2], Micah T. McClain[3,4,5], Ephraim L. Tsalik[1,3,4,5], Emily R. Ko[3,6], Geoffrey S. Ginsburg[3], Mark R. DeLong[2], Xiling Shen[7], Christopher W. Woods[3,4,5], Elizabeth R. Hauser[8,9] and Dennis C. Ko[1,5*]

Abstract

Background: While genome-wide associations studies (GWAS) have successfully elucidated the genetic architecture of complex human traits and diseases, understanding mechanisms that lead from genetic variation to pathophysiology remains an important challenge. Methods are needed to systematically bridge this crucial gap to facilitate experimental testing of hypotheses and translation to clinical utility.

Results: Here, we leveraged cross-phenotype associations to identify traits with shared genetic architecture, using linkage disequilibrium (LD) information to accurately capture shared SNPs by proxy, and calculate significance of enrichment. This shared genetic architecture was examined across differing biological scales through incorporating data from catalogs of clinical, cellular, and molecular GWAS. We have created an interactive web database (interactive Cross-Phenotype Analysis of GWAS database (iCPAGdb)) to facilitate exploration and allow rapid analysis of user-uploaded GWAS summary statistics. This database revealed well-known relationships among phenotypes, as well as the generation of novel hypotheses to explain the pathophysiology of common diseases. Application of iCPAGdb to a recent GWAS of severe COVID-19 demonstrated unexpected overlap of GWAS signals between COVID-19 and human diseases, including with idiopathic pulmonary fibrosis driven by the *DPP9* locus. Transcriptomics from peripheral blood of COVID-19 patients demonstrated that *DPP9* was induced in SARS-CoV-2 compared to healthy controls or those with bacterial infection. Further investigation of cross-phenotype SNPs associated with both severe COVID-19 and other human traits demonstrated colocalization of the GWAS signal at the *ABO* locus with plasma protein levels of a reported receptor of SARS-CoV-2, CD209 (DC-SIGN). This finding points to a possible mechanism whereby glycosylation of CD209 by *ABO* may regulate COVID-19 disease severity.

(Continued on next page)

* Correspondence: dennis.ko@duke.edu
[1]Department of Molecular Genetics and Microbiology, School of Medicine, Duke University, 0049 CARL Building Box 3053, 213 Research Drive, Durham, NC 27710, USA
[5]Division of Infectious Diseases, Department of Medicine, Duke University Medical Center, Durham, NC 27710, USA
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusions: Thus, connecting genetically related traits across phenotypic scales links human diseases to molecular and cellular measurements that can reveal mechanisms and lead to novel biomarkers and therapeutic approaches. The iCPAGdb web portal is accessible at http://cpag.oit.duke.edu and the software code at https://github.com/tbalmat/iCPAGdb.

Keywords: Pleiotropy, Cross-phenotype association, Gout, LD-score, Colocalization, PheWAS, Hi-HOST, Idiopathic pulmonary fibrosis, Macular telangiectasia, rs2869462, rs505922, rs12610495

Background

Genome-wide association studies (GWAS) have identified hundreds of thousands of genomic regions that are associated with complex human traits and have increased our understanding of the genetic architecture of human disease [1]. While GWAS now utilize even millions of subjects through leveraging electronic medical record data [2, 3], progress towards understanding how identified genetic variants alter cellular function and physiology remains elusive. More efficient mechanisms are needed for translating knowledge of genetic disease risk and severity into insight of the underlying physiology. Integrating analysis of GWAS across different scales of biological phenotypes (molecular, cellular, and organismal) may provide novel insight into how genetic variants influence complex traits.

Comparative analyses of GWAS have revealed that numerous, seemingly unrelated traits are connected by shared underlying genetic variants [1]. This phenomenon in which genetic variants affect multiple traits or diseases is called pleiotropy. Several methods have been developed to study pleiotropic SNPs by exploring the genetic relationship of multiple phenotypes. Broadly, these approaches can be categorized into three major groups. The first method is genetic correlation, which aims to quantify the similarity of the genetic effects on pairwise traits using GWAS summary statistics such as LD-score regression [4] or from individual genotype data with GCTA GREML [5]. With large population sizes, these methods can accurately partition variance into a shared genetic component but do not reveal the genetic variants driving the genetic correlation. Genome-wide cross-trait analysis [6] has emerged as a means to follow-up such results, but these univariate meta-analyses of two traits requires genome wide summary statistics for both traits, can suffer from effect size heterogeneity in combining results from disparate traits, and cannot be easily applied to thousands of traits at once. The second approach is colocalization, which estimates how well the GWAS signals from two signals overlap in a given region while revealing plausibility of individual causal variants [7]. These two methods have successfully identified novel genetic connections across distant traits as well as pleiotropic genomic regions but have generally been used independently of each other. Finally, perhaps the most intuitive approach, is quantifying cross-phenotype SNPs that are shared across multiple phenotypes. In its simplest form, a phenome-wide association study takes a single SNP and examines the significance of association across many traits, often from electronic medical record [8]. Valuable websites, including PhenoScanner [9], GRASP [10], and GeneATLAS [11], have integrated thousands of GWAS studies with billions of SNP-traits associations and allow users to query individual SNPs across the phenome. However, such PheWAS approaches do not leverage shared genetic architecture that extends beyond individual SNPs and do not take advantage of LD information.

Motivated to simultaneously connect human phenotypes with shared genetic architecture and to identify the precise loci driving this similarity, we previously developed a method, CPAG (Cross-phenotype Analysis of GWAS), which estimated phenotype similarity of NHGRI-EBI GWAS catalog [12] traits based on shared genetic associations [13]. CPAG utilized cross-phenotype SNP associations to cluster traits into groups that were consistent with pre-defined categories and discovered novel pleiotropic SNPs connecting Crohn's disease and the fatty acid palmitoleic acid. However, CPAG could not scale sufficiently to keep up with the massive increase in the scope and scale of GWAS (facilitated through increasing use of electronic medical record (EMR)-based GWAS of huge cohorts) and the deeper phenotyping of molecular and cellular traits that can provide insight into mechanisms of pathophysiology of disease. Here, we introduce iCPAGdb (interactive Cross-Phenotype Analysis of GWAS database; https://github.com/tbalmat/iCPAGdb [14]), a new cross-phenotype analysis platform with improved identification of shared loci using pre-computed ancestry-specific LD databases and a more efficient algorithm for capturing cross-phenotype associations. These improvements facilitated integration of the NHGRI-EBI GWAS catalog with large datasets of plasma and urine metabolites and cellular host-pathogen traits. Such integration of pleiotropic analyses using GWAS datasets that include intermediate traits across biological scales are crucial for moving from lists of associated SNPs to understanding the pathophysiology of complex diseases. Finally, iCPAGdb allows users to upload their own GWAS summary statistics via web interface (http://cpag.oit.duke.edu) to identify and explore shared SNPs between their own GWAS and a deep catalog of 4418 molecular, cellular, and disease phenotypes. Using a GWAS of severe COVID-19 [15] as the querying phenotype in iCPAGdb revealed shared SNPs associated with idiopathic pulmonary fibrosis and plasma protein levels of CD209, a possible receptor for SARS-CoV-2.

Implementation

Collection of GWAS summary statistics

Publicly available GWAS summary statistics were downloaded from the following sources: 3793 traits from NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) [12] (version 1.02, downloaded on 2020/08/05), 79 traits from H2P2 cellular GWAS (http://h2p2.oit.duke.edu) [16], and 546 traits from human blood circulating metabolites and urine metabolites GWAS (http://metabolomics.helmholtz-muenchen.de/gwas/) [17, 18]. NHGRI-EBI GWAS catalog traits included annotation by Experimental Factor Ontology (EFO). All GWAS data were harmonized to genome coordinates of HG19. In total, we collected 4418 GWAS traits, and 91,323 trait-SNPs pairs. A detailed list of trait-SNP pairs at varying *p* value threshold can be found in Table 1.

Severe COVID-19 summary statistics (adjusted for top 10 principal components) generated by Ellinghaus et al. [15] were downloaded from the GRASP webpage of aggregated COVID-19 GWAS results (https://grasp.nhlbi.nih.gov/Covid19GWASResults.aspx) [10] (download date 2020/07/15). Genome coordinates were converted from GRCh38 to HG19 using UCSC liftOver. GWAS summary statistics of IPF were kindly provided by Allen et al. [19] after requesting access from https://github.com/genomicsITER/PFgenetics.

LD clumping

GWAS summary statistics were individually preprocessed by LD clumping using *PLINK v1.9* [20], based on genotypes from European populations from the 1000 Genomes project [21]. The general PLINK command was "--clump-p1 1e-5 --clump-p2 1 --clump-r2 0.4 --clump-kb 1000." For NHGRI/EBI GWAS catalog, the index SNPs were selected using the genome-wide significant *p* value threshold of 5×10^{-8} (--clump-p1 5e-8). For molecular and cellular GWAS, we used a varying *p* value cutoff from 1×10^{-3} to 1×10^{-5} for --clump-p1 parameter to choose the index SNPs.

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

LD proxy calculation

To maximize phenotypic associations due to indirect associations, pairwise LD R^2 values were computed for each leading SNP against its surrounding SNPs using the genotypes from the 1000 Genomes project (Phase 3 genotypes) [21]. Prior to calculation, all SNPs with minor allele frequency less than 0.01 and missingness > 0.1 were removed. R^2 of pairwise SNPs within 10,000 bp windows were then calculated, and only LD proxies with $R^2 > 0.4$ were retained in further analysis. The PLINK parameters for calculating LD were "--ld-window-kb 1000 --ld-window 10000 --keep-allele-order --r2 in-phase with-freqs gz." The choice of R^2 threshold of 0.4 was a practical decision made to reduce multiply counting regions with broad association signals in moderate LD, while maintaining sensitivity and specificity in detecting cross-phenotype associations. This R^2 threshold was implemented in V1.0 of the iCPAGdb software used for all analysis in this manuscript and implemented in the web portal described below. However, users can also download the iCPAGdb V1.1 source code and population-specific LD proxy databases from Github (https://github.com/tbalmat/iCPAGdb [14])

Table 1 A summary of GWAS data in iCPAGdb

	Type	Traits/diseases #	SNPs ($p <$ 5e−8)	Trait-SNP associations #	Website
NHGRI catalog	Clinical GWAS	3793	63,933	85,639	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	65,540	91,323	

GWAS summary statistics were clumped to include only a lead SNP for each trait locus and run analysis with different R^2 cutoffs to retain overlapping SNPs with higher LD ($R^2 > 0.8$) or lower LD ($R^2 > 0.2$). While performing colocalization analysis is a prudent subsequent step in pursuing any iCPAGdb results, we caution that this is especially true when using the R^2 cutoff of 0.2, because the iCPAGdb analysis may be identifying two distinct association signals in the same region or a single shared signal being harder to detect due to noisy data.

Since GWAS may be performed on diverse populations from different ancestry or continents, we calculated ancestry-specific LD proxies for European, African, and Asian populations separately. European population included 503 samples from 5 populations (CEU, TSI, FIN, GBR, IBS), African included 661 samples from 7 populations (YRI, LWK, GWD, MSL, ESN, ASW, ACB), and Asian population included 504 samples from 5 populations (CHB, JPT, CHS, CDX, and KHV). We filtered genotypes for each ancestry population by minor allele frequency more than 0.01 and retained only biallelic SNPs. SNPs which have the same genome coordinates were merged using "--merge-equal-pos." For duplicated SNPs with the same variant rsID, we kept only the first variant by using "--rm-dup force-first" using PLINK 2.0.

Cross-phenotype SNP analysis

Cross-phenotype SNPs were used to quantify the similarity of different traits. Cross-phenotype loci were identified as leading SNPs and/or their LD proxies having statistically significant associations with more than one trait/disease. If two traits shared a common leading SNP, we termed this "direct association." If a leading SNP was associated with one trait, while its LD proxy SNPs were associated with another trait, we called this "indirect association." If any shared SNP was in LD with another SNP with $R^2 > 0.4$, these SNPs were merged into a SNP block until no further LD was found across shared SNP/LD pairs. The similarity of traits pairs based on shared associated SNPs was quantified using the Chao-Sorensen and Jaccard similarity index as described [13]. In this prior work, the Chao-Sorenson index resulted in clusters with less heterogeneity based on a pre-defined disease categorization compared to other measures of similarity and thus was the primary similarity index used in this work. The significance of association for each trait pair was calculated using Fisher's exact test, based on the hypergeometric distribution:

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

where N_e is the effective number of independent SNPs in the selected population, n_1 and n_2 are the number of independent SNPs associated with trait 1 and trait 2, and k is the number of independent SNP blocks. The effective number of independent SNPs for European, African and Asian population was obtained from Table 4 from [22].

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

Comparison to LDSC

Bulik-Sullivan et al. [4] developed an innovative and unbiased method, LDSC, to estimate genetic correlation using GWAS summary statistics for all measured SNPs. Their model calculated the LD scores for a variant against all other variants in a 1 centimorgan window and hypothesized that SNPs with higher LD scores are tagged to a risk-conferring variant, and the genetic correlation among traits can be calculated by normalizing genetic covariance of SNP heritability. With this method, they estimated 276 genetic correlations for 24 diseases/traits based on full GWAS summary statistics [23]. To evaluate the power of iCPAGdb, we calculated the genetic associations on the same 24 GWAS traits. For each trait pair, only SNPs associated with each trait passing a genome-wide significant threshold (5×10^{-8}) were used by iCPAGdb. We quantified the strength of cross-phenotype similarity for each trait pair using the Chao-Sorensen similarity index. Since the p values from [23] were not corrected by multiple test correction, we calculated the p values for r_g using the R "*p.adjust*" function with a total number of 276 comparisons.

Colocalization analysis

To evaluate whether the associations of GWAS trait pairs identified by iCPAGdb were due to sharing the same causal variants, we performed colocalization analysis using the R COLOC packages [7]. COLOC uses a Bayesian framework to estimate the posterior probability that two GWAS traits share two independent causal signals (PP3) or shares a single casual variant (PP4) in the selected genome region. For each trait pair evaluated by COLOC, SNPs within 200 kb window from the lead SNP were included. Since COLOC requires minor allele frequency (MAF) for each SNP in both GWAS studies, when MAF was not available, we calculated the MAF using European populations from the 1000 Genomes Project [21]. We ran COLOC "*coloc.abf*" function using the default prior parameters, p1 = 1×10^{-4}, p2 = 1×10^{-4}, and p12 = 1×10^{-5}. We also ran the built-in "*sensitivity*" function to evaluate the robustness of predefined priors, and all tests suggested that default prior parameters were robust. Therefore, we ran all colocalization analyses with default priors.

COVID-19 transcriptomic analysis

Transcriptomics of COVID-19 peripheral blood was generated by a previous study as described [24]. For that study [24], samples were collected as part of the Molecular and Epidemiological Study of Suspected Infection (MESSI), which was conducted at Duke University Health System (DUHS) and the Durham Veterans Affairs Health Care System (DVAHCS). SARS-CoV-2 RT-PCR testing was used to confirm infection status. A total of 46 subjects were analyzed, 14 of which were assayed at more than 1 time point. In total, 77 samples were assayed. Subjects were divided into early (≤ 10 days), middle (11–21 days), and late (> 21 days) stage based on duration of symptoms. Participant self-reported symptoms were recorded at each time point for 39 symptom categories. Each symptom was scored on a scale of 0–4, with 0 indicating not present, 1 mild, 2 moderate, 3 severe, and 4 very severe symptoms. Daily symptom severity (sum of symptom scores for all symptoms) was determined for each time point. At enrollment (day 0), date of symptom onset was determined, and an initial symptom survey recorded maximum score for each symptom category between symptom onset and study enrollment. Samples used for blood transcriptomics of acute respiratory infection due to seasonal coronavirus, influenza, or bacterial pneumonia, and healthy controls were also previously described [24]. Total RNA was extracted from peripheral whole blood, and cDNA libraries prepared using NuGEN Universal Plus mRNA-seq with AnyDeplete Globin reduction were sequenced on the Illumina NovaSeq 6000, as described [24]. In brief, STAR v 2.7.1 [25] was used to align the short reads and generate the count matrix. The count matrix was further normalized using TMM method [26] and log2 transformed. Associations were performed with generalized linear models (LIMMA, [27]) and corrected for multiple testing using the Benjamini-Hochberg method [28]. RNAs-seq datasets generated in [24] are publicly available through the National Center for Biotechnology Information Gene Expression Omnibus, accession# GSE161731. For the present study, analysis of *DPP9* expression in these data was carried out in R, and p values were calculated using the Wilcoxon rank-sum test.

iCPAGdb software and website implementation

iCPAGdb is comprised of two core parts, the back-end computation and the front-end web browser. All code is available at https://github.com/tbalmat/iCPAGdb [14]. The back-end was written in python v3.6 with utilization of SQLite. SQLite tables were constructed for harmonized GWAS datasets and LD tables for different populations and are accessed using python sqlite3 package. The GWAS table stores clumped GWAS summary statistics, including trait name, trait sources, SNP rsIDs, beta values, standard error/standard deviation of beta, effective allele, and p values. The ancestry-specific LD proxy tables contain pairwise SNPs' rsID and R^2 values (R^2 > = 0.4) for different populations. All SQLite tables were indexed on unique combinations of SNP and trait or SNP pairs for LD proxy tables, which greatly reduces the searching time. To further increase calculation speed, the core cross-phenotype analysis part of iCPAGdb is parallelized by utilizing multiple threads.

Primary software components for the web portion of iCPAGdb are the R statistical programming language [29], the R package Shiny (v1.5.0) for interaction of web pages with R scripts [30], Shiny Server as a 24/7 multi-user platform to make Shiny apps publicly accessible [31], the database environment SQLite for efficient querying of GWAS and iCPAGdb results [32], and the R package RSQLite to execute SQL queries from within R scripts [33]. The results of a iCPAGdb execution are read by the R script, processed, and presented to the viewer in various tables and graphs on a web page. The iCPAGdb website is currently loaded with associations across more than 4400 public GWAS datasets that can be browsed and searched in "Review" mode. The user requests an existing iCPAGdb result set from which a corresponding table and heatmap are generated and displayed. Various filtering and graph construction controls are available for iterative sub-setting of data and selection of significance measure and number of top significant phenotype pairs to plot. The "Download" button enables the researcher to make a local copy of records appearing in the currently displayed results table. Important packages used in this mode are DT for construction of and interaction with tables and ggplot2, plotly, and heatmaply for basic plotting, interactive plotting (hover labels), and heatmap generation, respectively. The web browser also allows users to upload their own GWAS summary data, and iCPAGdb will automatically perform LD clumping based on selected population and generate an atlas of connections for the user's GWAS against > 4400 GWAS traits in the database. In this "Upload" mode, the user browses files on a local computer, selects a properly formatted GWAS result file of interest (containing, for a single phenotype, SNP rsIDs and GWAS p values), specifies format and column configuration, then uploads the file. Next, iCPAGdb computation parameter values, including iCPAGdb GWAS set to be crossed with, significance thresholds for filtering, and linkage disequilibrium (LD) population are specified. When "Compute CPAG" is pressed, the R script composes a system level command to execute the CPAG (Python) function. The future() function of the R future package [34] combined with a delaying pipe from the promises package execute iCPAGdb operations asynchronously, waiting on completion before resuming R script execution. Typical run time for a single uploaded GWAS that is already clumped to lead variants is < 30 s. For GWAS summary statistics including all SNPs in a study, run time is typically < 2 min. The results are available with downloadable tables and figures. Additional information on webapp is in Additional file 1.

Results

iCPAGdb: an atlas for discovery of cross-phenotype associations

We created iCPAGdb to facilitate exploration of cross-phenotype associations of human phenotypes and discovery of shared genetics connecting traits that were previously not known to be related. iCPAGdb utilizes 85,639 SNP-trait associations ($p < 5 \times 10^{-8}$) across 3793 traits from the NHGRI-EBI GWAS catalog, incorporates additional GWAS datasets (see below and Table 1), and allows for uploading and analysis of user GWAS summary statistics (Fig. 1a). In contrast, the original CPAG (published in 2015 [13];) used only 14198 SNP-trait associations for 887 traits from the NHGRI-EBI GWAS catalog.

Beyond this large expansion in traits and associations, we improved on the original CPAG algorithm by clumping GWAS data from each study (Additional file 2: Figure S1), creating a database of LD values based on 1000 Genomes [21], allowing selection of either European, African, or Asian LD structure, and efficiently capturing cross-phenotype associations that are driven by LD proxy (Fig. 1b). For each trait pair, iCPAGdb first selects the lead SNPs from all associated loci at a selected p value threshold ($p < 5 \times 10^{-8}$ was used for analysis of the NHGRI-EBI GWAS catalog; Additional file 3: Table S1; Additional file 4: Figure S2). These lead SNPs are compared across the trait pair to count directly shared SNPs. For SNPs that are not directly shared, iCPAGdb then checks an LD database for overlap by LD proxy. For all directly or indirectly shared SNPs, iCPAGdb further forms them into bigger SNP blocks by recursively merging them until each SNP block has no LD proxy with $R^2 >= 0.4$ against all others. iCPAGdb improves memory efficiency with built-in functions connecting to SQL GWAS and LD proxy databases and improves computational efficiency and speed by utilizing multiple CPUs. For the NHGRI-EBI GWAS Catalog, the growth of GWAS findings and improvements of iCPAGdb over the previous version of CPAG led to a 27.7-fold increase in direct cross-phenotype associations and a 47.7-fold increase in indirect cross-phenotype associations, many of which would have been missed by the original CPAG algorithm (Fig. 1c, d). Indeed, analyzing the 2013 NHGRI-EBI GWAS catalog with iCPAGdb had little effect on direct associations but increased indirect associations by 76% (Additional file 5: Figure S3).

Results of iCPAGdb are consistent with results from the orthogonal approach of genetic correlation by LD score regression [4]. Comparing the absolute values for genetic correlation of 24 phenotypes from [23] against a similarity index quantifying the degree of shared SNPs in iCPAGdb revealed that the two are significantly correlated ($p = 3.52 \times 10^{-8}$; $R^2 = 0.14$) (Fig. 1e). Nearly all phenotypes (64 of 70) that showed significant correlation by LD score regression also demonstrated a significant excess of shared SNPs in iCPAGdb. The output of iCPAGdb provides the SNPs driving the similarity between the two phenotypes, facilitating follow-up studies. Interestingly, 61% of pairwise comparisons that had significant overlap based on iCPAGdb did not have significant genetic correlation based on LD score regression. For example, LD-score regression did not detect significant genetic correlation between LDL and HDL cholesterol measurements [23], but iCPAGdb detected 92 shared SNPs, including 31 by direct overlap where the two phenotypes have the same lead SNPs ($p = 7.55 \times 10^{-195}$ by Fisher's exact test; $p = 1.49 \times 10^{-190}$ after Benjamini-Hochberg procedure. p values from iCPAGdb in the remainder of the paper are FDR-corrected for all pairwise comparisons using Benjamini-Hochberg procedure). As iCPAGdb quantifies overlapping SNPs to detect shared genetic associations rather than genetic correlation based on genome-wide summary statistics, iCPAGdb may be detecting instances where discrete loci are shared (perhaps suggesting shared biology or pleiotropy), even though the genome-wide genetic correlation is weak. Another possibility is due to the fact that iCPAGdb does not take into account directionality of effects for quantifying SNP overlap. Shared SNPs with inconsistent direction of effect between traits can lead to overall low genetic correlation by LD score regression, even when many SNPs are shared, as has been previously described for some autoimmune diseases [23].

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

In a previous study [13], we defined 4 categories of cross-phenotype associations: (1) SNP similarity between an intermediate trait/risk factor and disease, (2) SNP similarity between a disease and a consequence of disease, (3) SNP similarity between two traits affected by the same gene/pathway, and (4) SNP similarity between two traits affected by the same gene having effects in different tissues or on different pathways. Of these categories, perhaps the most clinically useful is the first category—shared SNPs that connect an intermediate

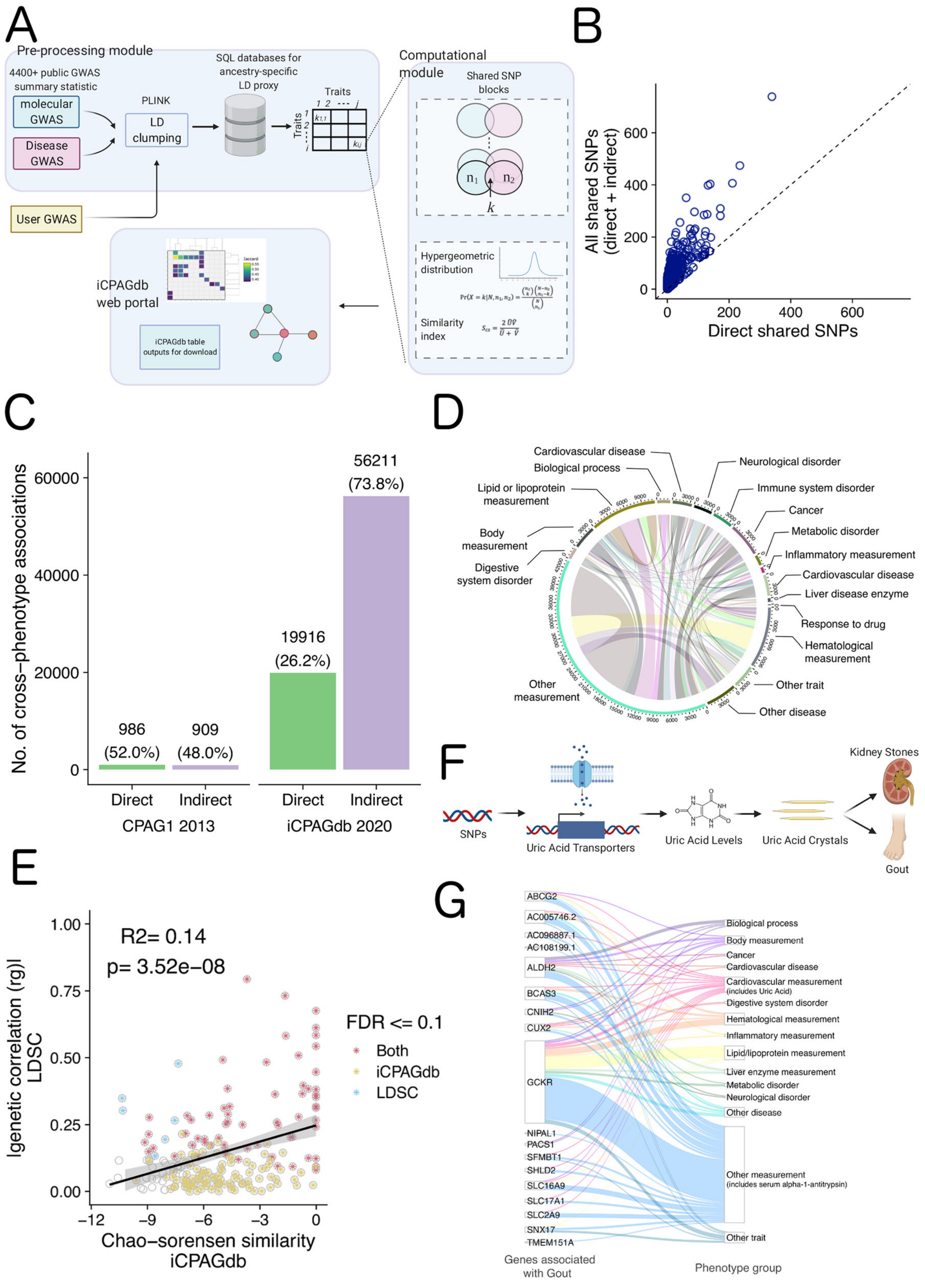


Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 An improved method for finding shared genetic architecture of human traits. **a** The overall framework of the iCPAGdb pipeline. GWAS summary statistics (from published GWAS datasets or from user-uploaded GWAS) undergo LD clumping to obtain a lead variant for each signal below a specified *p* value threshold. These SNPs are queried against an LD proxy database generated from 1000 Genomes African, Asian, or European population to identify cross-phenotype associations through direct overlap or LD proxy at $R^2 > 0.4$. Significance of overlap for each trait pair was calculated using Fisher's exact test. Outputs can be visualized/downloaded from the iCPAGdb web browser. **b** Comparison of the number of shared SNPs for each NHGRI-EBI GWAS catalog trait pair identified through direct overlap vs. both direct and indirect (LD-proxy) overlap. **c** iCPAGdb detected more significant cross-phenotypes associations than CPAG1 at FDR < 0.1. Expansion of the NHGRI-EBI GWAS catalog and improvements in capturing by LD proxy in iCPAGdb fueled a large increase in detected cross-phenotype associations across human traits. Comparisons between CPAG1 and iCPAGdb on the same 2013 dataset are in Additional file 5: Figure S3. **d** Circle plot of cross-phenotype associations detected by iCPAGdb in the NHGRI-EBI GWAS catalog. After excluding compound phenotypes (phenotypes described by NHGRI-EBI GWAS catalog as > 1 comma-separated phenotype in their ontology), a total of 1709 traits involved in a total of 53314 cross-phenotype associations were left. These were categorized into 17 EFO Parental groups. Inner ribbons link phenotypes connected by cross-phenotype associations with the width of ribbon corresponding to the number of cross-phenotype associations. The axis outside the circle represents the cumulative number of associations for each group vs all other groups. **e** Comparison of genetic correlation from LD score regression (LDSC) and the Chao-Sorensen similarity index implemented in iCPAG demonstrates significant correlation. The genetic correlation r_g of 24 diseases/trait were obtained from [23]. Since Chao-Sorensen values are bounded from 0 to 1 and r_g ranges from − 1 to 1, we used the absolute value of r_g here. Colored * indicates significant trait-pair for LDSC, iCPAGdb, or both at false discovery rate of 0.1. **f** A model demonstrating how SNPs regulate uric acid levels to impact the development of kidney stones and gout. **g** Riverplot of gout cross-phenotype associations generated from iCPAGdb output shows mapped genes associated with gout by GWAS (left) connected with NHGRI-EBI GWAS phenotypes grouped into EFO categories (right; colors are different categories). Cross-phenotype associations include causal connections (such as uric acid levels), comorbid outcomes (such as kidney stones), and regulators of disease (such as alpha-1-antitrypsin levels)

trait to a disease may reveal how molecular or cellular phenotypes mediate some aspect of the pathophysiology of disease. While the NHGRI-EBI GWAS catalog is comprised primarily of case-control GWAS of disease, we detected numerous known shared associations linking a human disease with levels of a metabolite. Metabolites are the substrates, intermediates, and products of cellular metabolism and are routinely already used as biomarkers, such as measuring glucose in diabetes management.

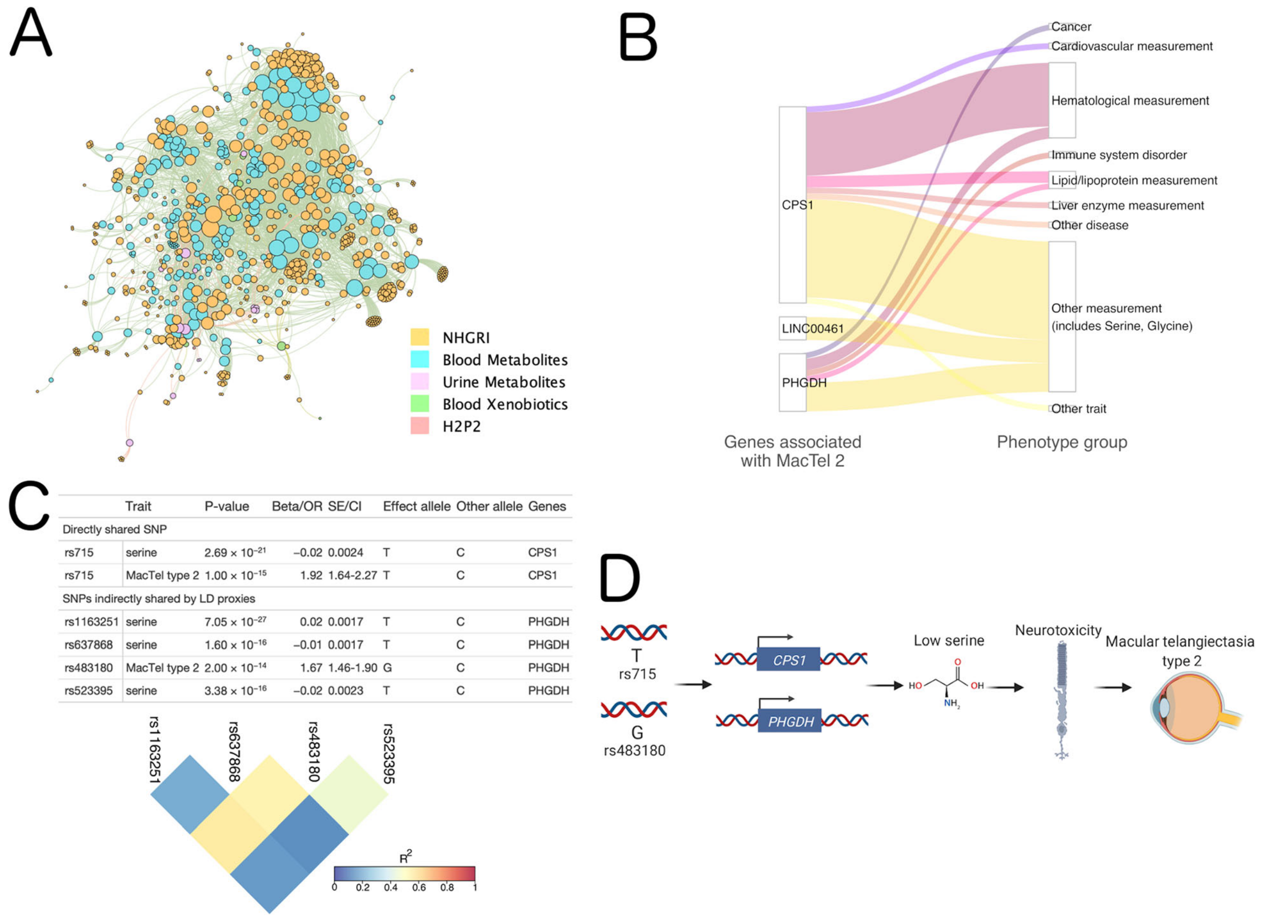
Cross-phenotype associations involving the metabolite uric acid and gout, an inflammatory arthritis driven by excess levels of uric acid [35], are illustrative of iCPAGdb's usefulness. GWAS studies have been conducted on risk of gout [36–43] as well as uric acid or urate levels [44–51]. Notably, of 32 GWAS loci for gout and 126 GWAS loci for serum uric acid levels at $p < 5 \times 10^{-8}$, 13 loci overlap, including 9 loci identified only by LD proxy (over 5000-fold enrichment; $p = 1.4 \times 10^{-42}$). These loci are spread across 7 chromosomes and include several solute carrier (SLC) and ATP-binding class (ABC) transporters that control urate absorption and secretion. Some of the loci are in close proximity but are counted separately by iCPAGdb, as could occur if different GWAS studies locate nearby peaks that fall below our $R^2 > 0.4$ threshold or if multiple causal signals are located in the same region. These data provide genetic evidence for the well-known causal role of excess uric acid in gout and further reveal multiple genes that may serve as therapeutic targets. Inhibitors of renal uric acid reabsorption through URAT1 (*SLC22A12*) are commonly used in treating gout [52], but additional transporters implicated through human genetics may also prove to be useful drug targets. Beyond uric acid levels, GWAS of kidney stones [53–55], a second manifestation of elevated uric acid levels, also share associated SNPs with gout (3 shared loci, all identified by proxy on chromosomes 2, 4, and 17; $p = 5.9 \times 10^{-9}$). Finally, gout shares 2 loci (out of 5 from [56, 57]) with levels of serum alpha-1-antitrypsin, an anti-inflammatory endogenous protease inhibitor ($p = 1.0 \times 10^{-6}$), providing a human genetic rationale for the use of alpha-1-antitrypsin-based therapeutics in acute gouty flares (as has been demonstrated to be efficacious in mice [58]). Thus, examining the gout cross-phenotype associations revealed causal connections, comorbid conditions with shared etiology, and factors that modulate inflammation in the disease (Fig. 1f, g).

Shared genetic associations reveal other well-known molecular and cellular disease relationships such as LDL cholesterol levels with cardiovascular disease (1.61×10^{-83}) and Alzheimer's disease ($p = 8.9 \times 10^{-17}$) as well as glucose with type II diabetes mellitus ($p = 5.3 \times 10^{-40}$). Other cross-phenotype associations highlight genetic variation that can extend our knowledge. For example, cross-phenotype associations were found between malaria [59–64] and red blood cell distribution width [65–69] ($p = 1.4 \times 10^{-9}$). This overlap is driven by well-known genetic variation in the beta-hemoglobin gene (*HBB)* and *ABO* blood type affecting malaria risk but also by genetic variation in *ATP2B4* which encodes a calcium transporter. To the best of our knowledge, whether size of red blood cells impacts susceptibility to malaria parasites has not been examined. These cross-phenotype associations demonstrate the promise of this approach for revealing novel relationships that can be mined through iCPAGdb.

Expansion of iCPAGdb to additional datasets of molecular and cellular traits

The above examples of clinically relevant cross-phenotype associations involving metabolite and cellular phenotypes motivated expansion of iCPAGdb to additional datasets. We used three datasets to provide molecular and cellular traits to our analysis: 491 metabolites and xenobiotics in blood [18] and 55 metabolites in urine [17], both from the Metabolomics GWAS Server (http://metabolomics.helmholtz-muenchen.de/gwas/index.php), and 79 cellular host-pathogen interaction traits from our dataset of cellular host-pathogen interaction GWAS, H2P2 (http://h2p2.oit.duke.edu) [16]. iCPAGdb revealed many connections between these molecular/cellular datasets and the NHGRI-EBI GWAS catalog (Fig. 2a; Additional file 6: Table S2).

Cross-phenotype associations with macular telangiectasia (MacTel) type 2, a disease characterized by loss of central vision due to alterations in blood vessels in the macula of the eye, confirmed the importance of the amino acid serine (Fig. 2b). A GWAS of MacTel type 2 uncovered 3 genome-wide significant loci and the authors noted that two of these loci were involved in serine/glycine metabolism, with the alleles associated with low glycine and serine conferring increased risk of



	Trait	P-value	Beta/OR	SE/CI	Effect allele	Other allele	Genes
Directly shared SNP							
rs715	serine	2.69×10^{-21}	−0.02	0.0024	T	C	CPS1
rs715	MacTel type 2	1.00×10^{-15}	1.92	1.64-2.27	T	C	CPS1
SNPs indirectly shared by LD proxies							
rs1163251	serine	7.05×10^{-27}	0.02	0.0017	T	C	PHGDH
rs637868	serine	1.60×10^{-16}	−0.01	0.0017	T	C	PHGDH
rs483180	MacTel type 2	2.00×10^{-14}	1.67	1.46-1.90	G	C	PHGDH
rs523395	serine	3.38×10^{-16}	−0.02	0.0023	T	C	PHGDH

Fig. 2 iCPAGdb integrates GWAS of different scales to reveal a biological connection between MacTel 2 and serine. **a** Multi-dataset network of cross-phenotype associations detected by iCPAGdb. Phenotypes that demonstrated significant overlap (FDR ≤0.1) are color-coded in the indicated colors. **b** Riverplot of macular telangiectasia type 2 (MacTel type 2) cross-phenotype associations generated from iCPAGdb shows mapped genes associated with MacTel type 2 (left) connected with NHGRI-EBI GWAS phenotypes grouped into EFO categories (right; colors are different categories). SNPs in *CPS1* and *PHGDH* are associated with MacTel type 2 and are also associated with serine levels, which are believed to play a causal role in the disease. Other connections may represent causal connections, comorbid outcomes, and regulators of disease. **c** Cross-phenotype associations connecting MacTel type 2 and serine. One locus demonstrated direct SNP overlap (rs715). A second locus demonstrated indirect overlap based on 4 SNPs in LD as visualized in the heatmap color-coded by LD. **d** A model for how SNPs regulate serine levels to impact pathogenesis of MacTel type 2 based on iCPAGdb and prior work described in the text

MacTel type 2 [70]. The authors speculated that the low serine levels could lead to high levels of ammonia and glutamate causing neurotoxicity and stress-induced angiogenesis [70]. Gantner et al. have since provided evidence that low serine levels result in elevated levels of deoxysphingolipids to trigger cell death in photoreceptors [71]. iCPAGdb rediscovered the connection of two loci being shared between serine in serum (measured by [18]) and risk of MacTel (Fig. 2c, d; p = 3.9 × 10^{-7}; 99, 084-fold enrichment). iCPAGdb also revealed 7 other serum metabolites including glycine that shared an association with rs715 but not with the second MacTel locus. While serine was not part of the urine metabolomics dataset, iCPAGdb did detect overlap of glycine in urine and MacTel type 2 (p = 0.01).

We also included host-pathogen traits from H2P2, a cellular GWAS we previously carried out using 528 lymphoblastoid cell lines (LCLs) exposed to 7 different pathogens [16]. Notably, unlike the metabolomics datasets, H2P2 identified SNPs associated with traits at baseline and in response to stimuli. Further, as pathogens have likely been drivers of human evolution [72, 73], comparing H2P2 to human disease GWAS may reveal unintended consequences of past pandemics on the human genome. Previously, we reported colocalization of a locus regulating CXCL10 levels following *Chlamydia trachomatis* infection (rs2869462) and risk of inflammatory bowel disease [16]. iCPAGdb revealed shared genetic variants for this H2P2 phenotype and blood levels of CXCL9 (MIG) [74] (Fig. 3a; p = 0.04). p values for the two associations are strongly correlated (Fig. 3b), and the effect size for SNPs associated with both chemokines are significantly positively correlated (Fig. 3c). We utilized COLOC, which uses a Bayesian framework to determine whether GWAS signals in the same region are likely due to the same causal variant [7]. The posterior probability that both CXCL10 protein levels from cells and CXCL9 levels in blood share the same causal variant is 0.90 (Table 2), with rs2869462 identified as the most likely causal SNP (Additional file 7: Table S3). The genes encoding these two chemokines are adjacent to each other on chromosome 4, and this result points to variants regulating expression of both genes that will make it challenging to disentangle their effects in disease.

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

We applied iCPAGdb to a recently published GWAS of severe COVID-19 with respiratory failure [15]. While this study focused on two genome-wide significant associations at the *ABO* locus and in a cluster of chemokine receptors and other genes on Chromosome 3, we relaxed the p value threshold for iCPAGdb to 1×10^{-5}, resulting in 24 suggestive loci after LD clumping. Not surprisingly, iCPAGdb revealed that the genome-wide significant association near the blood type locus *ABO* is in LD with multiple other SNPs in this region associated with other human diseases and traits (Fig. 4a; Additional file 8: Table S4). This included the classic association with malaria resistance [64], but also less well known associations with duodenal ulcer [75], pancreatic cancer [76], and heart failure [77]. Multiple studies have now reported the association of the *ABO* locus with risk of COVID-19 [15, 78]. The causal effect on COVID-19 may involve A and B antigens on blood cells, antibodies against A and B antigens, the enzymatic activity of the ABO glycosyltransferase on possibly other glycoproteins, or even other genes in the region. Insight into these possible mechanisms was revealed by iCPAGdb, which identified association of this locus with levels of 8 individual proteins in the NHGRI-EBI GWAS catalog. These proteins, all encoded on different chromosomes than ABO, include IL-6, TNF-α, CD209 (DC-SIGN), Tie-1, mannose-binding protein C, FGF23, and clotting factors (factor VIII and vWF). In each of these cases, the association of the locus to both molecular trait and disease provides a plausible causal chain from SNP to *cis*-effect on *ABO* to trans-effect on a protein to severe COVID-19 disease. For example, association with VWF and Factor VIII may indicate ABO affects COVID-19 through regulation of thrombosis, as patients with severe COVID-19 can have thromboembolic complications as part of a hyper-inflammatory state [79]. In fact, both VWF and factor VIII are targets of glycosylation by ABO [80–82] and levels of these proteins are reported to be regulated by ABO [83–87]. Further, regulation of levels of IL-6 and TNF-α suggest possible regulation of inflammation, as "cytokine storm" plays an important role during severe COVID-19 [88]. Most interestingly, the *ABO* locus is associated with both COVID-19 and CD209 (p = 0.008). A preprint recently confirmed this association across populations, and these authors speculated that ABO may affect CD209 levels to regulate SARS-CoV-2 entry [89]. Indeed, there has since been evidence from two preprints that CD209 can bind to SARS-CoV-2 and can act as a receptor for entry into immune cells [90, 91].

The "A" allele of rs657152 associated with increased risk of COVID-19 with respiratory failure is also associated with increased levels of CD209 (Fig. 4b). We performed colocalization analysis of the GWAS signals for COVID-19 [15] and CD209 protein levels [57]. This analysis indicated the two are likely driven by the same causal variants (Fig. 4c; COLOC posterior probability PP4 = 0.98 with the lead causal SNP of rs505922; Additional file 9: Table S5). Thus, iCPAGdb and subsequent

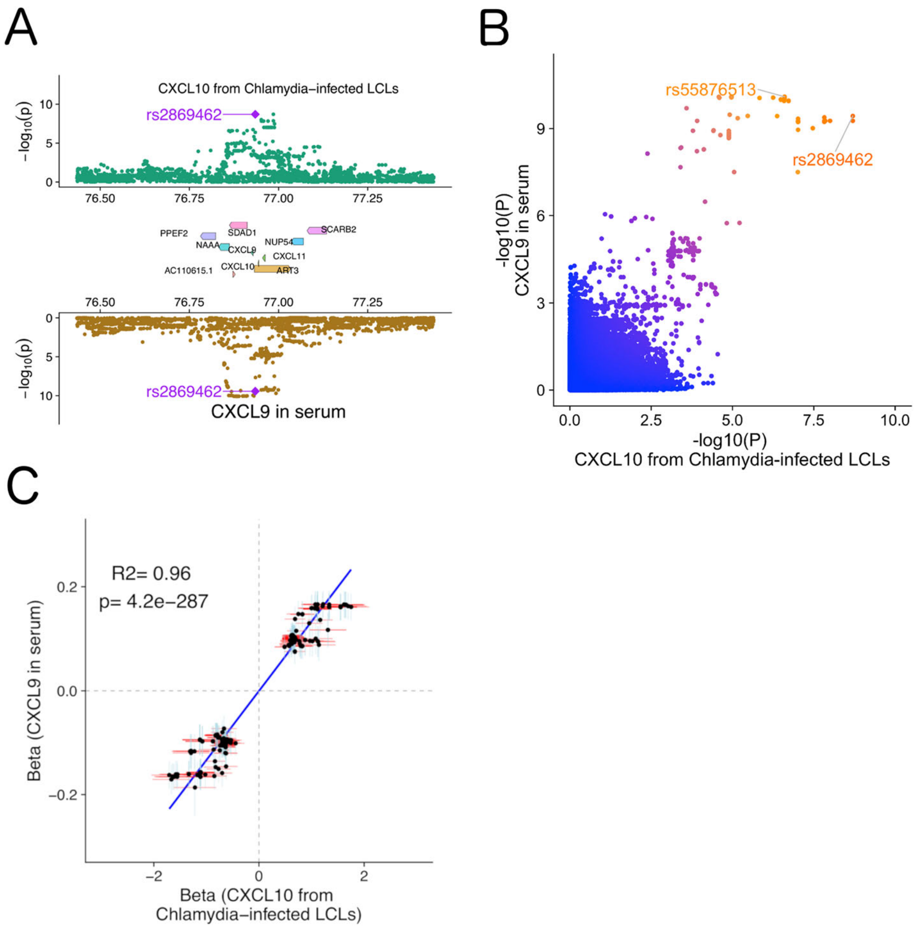


Fig. 3 Cross-phenotype association analysis reveals the same genetic locus impacts both *Chlamydia*-induced CXCL10 levels and MIG level in serum. **a** Regional Miami colocalization plot demonstrates a genetic locus that impacts both CXCL10 level in lymphoblastoid cell lines following *Chlamydia trachomatis* infection and CXCL9 (MIG) levels in serum. **b** Comparison of -log10(*p* value) for GWAS of CXCL10 following *C. trachomatis* infection and levels of CXCL9 (MIG) in serum. The lead SNP in the region for each phenotype is marked. **c** Scatter plot demonstrates a highly positive correlation of the effect coefficients of cellular CXCL10 after *C. trachomatis* infection and of SNPs associated with blood CXCL9 levels. Each dot represents a SNP which has *p* value < 0.01 for both phenotypes. A total of 413 SNPs from a 4-mb window surrounding the leading SNP rs2869462 was selected. The blue vertical or red horizontal bar shows the standard error of the beta value for each SNP

colocalization analysis support a model where *ABO* regulates CD209 protein levels to impact COVID-19 risk, though much future experimental and clinical studies will be required to fully test this hypothesis (Fig. 4d). The pleiotropic effects of *ABO* on levels of multiple proteins will make defining the mechanism challenging.

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

Beyond *ABO,* a locus in the dipeptidyl peptidase 9 (*DPP9*) gene associated at $p < 1 \times 10^{-5}$ with severe COVID-19 was identified as being shared with a GWAS of fibrotic idiopathic interstitial pneumonia [92] and a

Table 2 COLOC analysis output

Trait1	Trait2	Locus	SNP #	PP3	PP4	PP3 + PP4	PP4/PP3	Lead causal SNP
CXCL10 level after *Chlamydia* 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

PP3 is the posterior probability for the model where the two traits have independent causal variants. PP4 is the posterior probability for the model where the two traits share a single causal variant

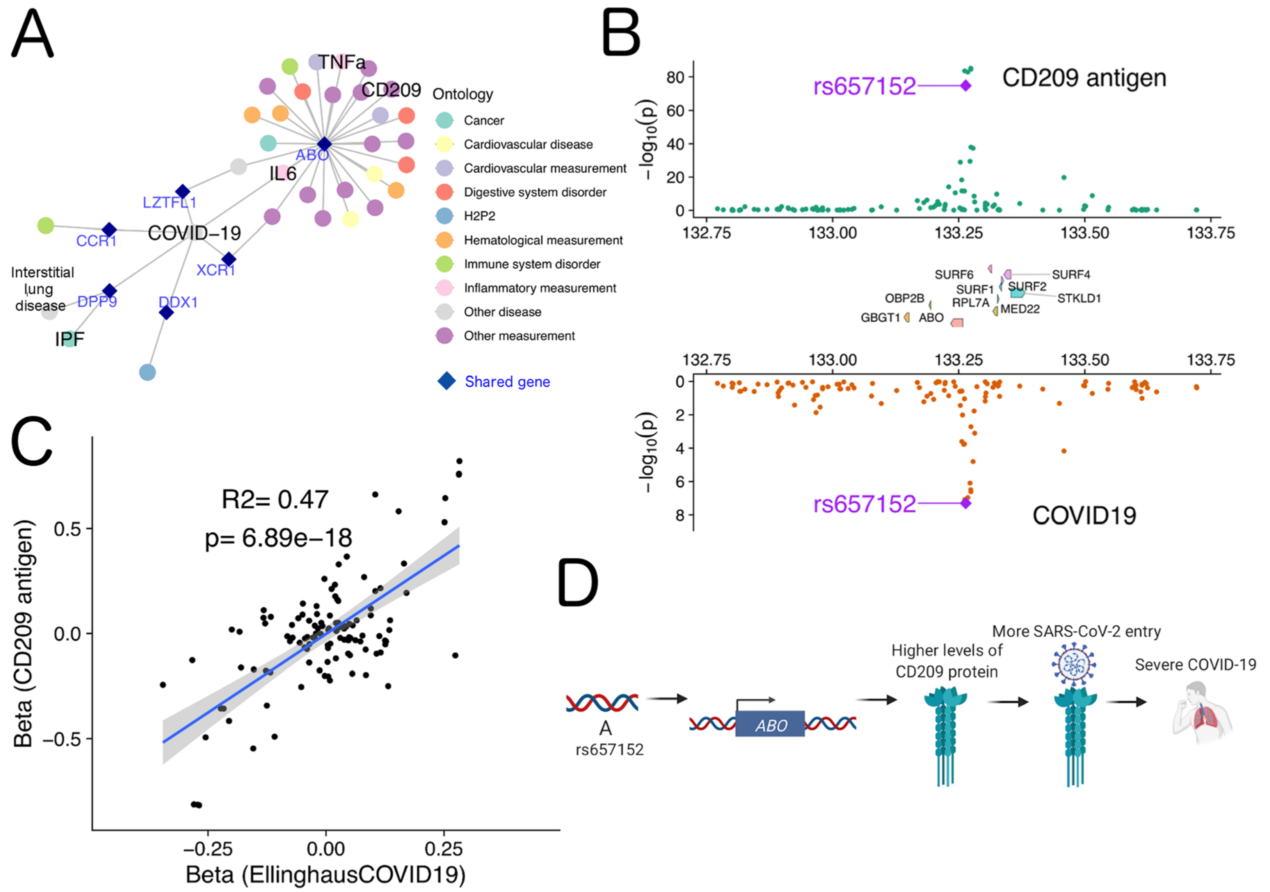


Fig. 4 Cross-phenotype association of *ABO* reveals a possible role for CD209 in severe COVID-19. **a** A network of genetic associations involving severe COVID-19. Each node represents either a disease/trait (filled circles) or a gene (dark blue diamond). The *ABO* locus was associated with multiple other diseases and levels of specific proteins, while DPP9 connects COVID-19 only with IPF and interstitial lung disease (idiopathic interstitial pneumonia). **b** Regional Miami colocalization plot demonstrates the *ABO* locus impacts both CD209 protein levels and risk of severe COVID-19. **c** A significant positive correlation for effect size of SNPs in the *ABO* locus on CD209 protein levels and risk of severe COVID-19. **d** Model of how *ABO* may affect CD209 and severe COVID-19

recent GWAS of the most severe form of that group of diseases, idiopathic pulmonary fibrosis (IPF) [19]. rs12610495 was the lead variant for each of these GWAS studies as well as the suggestive peak for severe COVID-19 (p = 5.2 × 10^{-6} [15];). Much evidence has already accumulated that pulmonary fibrosis is a hallmark of severe COVID-19 [93, 94]. While the association of rs12610495 with COVID-19 did not reach genome-wide significance in Ellinghaus et al. 2020 [15], this SNP is in LD with the lead variant from a recent GWAS of critically ill COVID-19 patients that does surpass genome-wide significance (p = 3.98 × 10^{-12} [95];; R^2 = 0.95 in 1000 Genomes European populations). Thus, iCPAGdb alerted us to the importance of a suggestive COVID-19 susceptibility locus that has since been validated in an independent cohort.

We determined that rs12610495 is an eQTL in lung tissue for the gene for *DPP9* (and no other genes in the region) in GTEx (p = 4.5 × 10^{-9} [96];), with the "G" allele being associated with lower expression (Fig. 5a). Interestingly, DPP9 is a protease in the same family as DPP4, the receptor for MERS-coronavirus [97]. Additionally, DPP9 is an inhibitor of inflammasome activation by NLRP3 [98–100]. Colocalization analysis confirmed the signals from severe COVID-19 and IPF are likely driven by the same causal variant (Fig. 5b; COLOC posterior probability PP4 = 0.994, lead SNP rs12610495; Additional file 10: Table S6). Based on these data and the known biology, we developed alternative hypotheses for how this SNP might be regulating risk of severe COVID-19: *DPP9* may be acting as a previously unrecognized receptor for SARS-CoV-2 or it may be inhibiting inflammation during COVID-19 infection. Based on the directionality of effect of rs12610495 on *DPP9* gene expression, the "G" allele should lead to lower *DPP9* expression and less entry if the receptor model is correct. However, the "G" allele is instead associated with increased risk of severe COVID-19 (Fig. 5c).

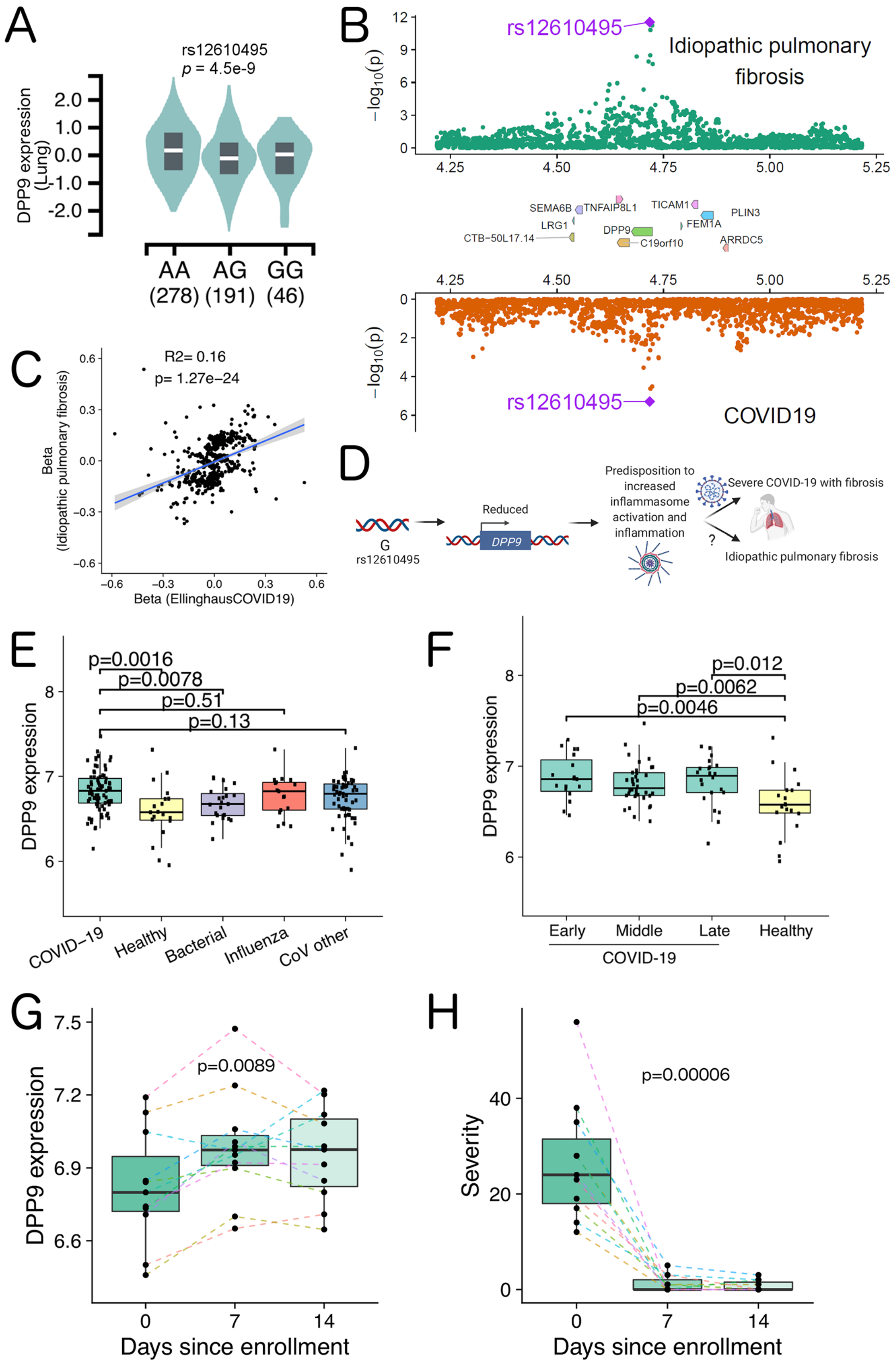


Fig. 5 (See legend on next page.)

(See figure on previous page.)
Fig. 5 Cross-phenotype analysis and COVID-19 patient transcriptomics reveals a role for *DPP9* in severe COVID-19. **a** Lung eQTL data from GTEx shows rs12610495 "G" allele is associated with reduced expression of *DPP9*. **b** Regional Miami colocalization plot demonstrates the *DPP9* locus impacts both idiopathic pulmonary fibrosis and risk of severe COVID-19. **c** A significant positive correlation for effect size of SNPs in the *DPP9* locus on idiopathic pulmonary fibrosis and risk of severe COVID-19. **d** Model of how *DPP9* may affect idiopathic pulmonary fibrosis and risk of severe COVID-19. **e** *DPP9* expression in peripheral blood is significantly higher in COVID-19 patients (n = 77 samples) compared to healthy (n = 19) and bacteria-infected patients (n = 23). The p values were calculated using the Wilcoxon rank-sum test. **f** COVID-19 patients demonstrate significantly higher *DPP9* expression compared to healthy controls during early (days 1–10; n = 19 samples), middle (days 11–20; n = 36), and late (21+ days; n=22) stages of SARS-CoV-2 infection. The p values were calculated using the Wilcoxon rank-sum test. **g** *DPP9* demonstrates increased expression during recovery from COVID-19. A total of 11 patients were measured sequentially at enrollment (day 0), day 7, and day 14. The colored dash line connects measurements from the same patient across time points. p value was calculated using Friedman test. **h** Decreased symptom severity scores of COVID-19 patients over time. The eleven subjects in G were assessed for symptom severity at days 0, 7, and 14. The colored dash line connects measurements from the same patient across time points. p value was calculated using Friedman test

Alternatively, the "G" allele could lead to lower *DPP9* to increase inflammasome activation in lung tissue, a model consistent with "G" increasing risk of severe COVID-19 and this allele also increasing risk of idiopathic pulmonary fibrosis (Fig. 5d).

To further examine the role of *DPP9* in COVID-19, we analyzed previously generated transcriptomics of peripheral blood from COVID-19 patients [24]. Levels of *DPP9* expression across 46 COVID-19 patients were compared to individuals with seasonal coronavirus, influenza, bacterial pneumonia, and healthy controls. *DPP9* levels were significantly increased in COVID-19 patients compared to the other groups (fold-change = 1.15, p = 0.003 adjusted by Benjamini-Hochberg method). Comparing COVID-19 data vs. each comparator individually revealed that *DPP9* levels were elevated vs. healthy controls (p = 0.0016) and bacterial infection (p = 0.0078) but not influenza or other coronavirus infection (Fig. 5e). This data supports a role for *DPP9* in the host response to viral infections. In examining all samples in the cohort, increased *DPP9* was observed both early and late in COVID-19 infection (Fig. 5f). However, eleven subjects that did not require hospitalization had repeated measurements at day 0 (initial enrollment into the study), day 7, and day 14 that revealed changes in *DPP9* expression as infection resolved. While *DPP9* expression increased from day 0 compared to 7 days and 14 days (Fig. 5g; p = 0.0089), symptom severity dramatically improved over this period (Fig. 5h; p = 0.00006). We speculate that *DPP9* may be induced to effectively turn off the inflammatory response to SARS-CoV-2 to minimize tissue damage and fibrosis. Combined with our human genetic data, these findings suggest that insufficient induction of *DPP9* expression could predispose to severe COVID-19.

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

As the above examples demonstrate, iCPAGdb analysis can rapidly generate hypotheses connecting molecular and cellular traits to human disease. The website allows quick access to the pre-calculated cross-phenotype associations results described in this manuscript. Users can also upload their own GWAS summary statistics for comparing against all 4418 GWAS traits in the iCPAGdb website, facilitating the discovery of new cross-phenotype relationships. Total time for uploading, clumping of summary statistics, and calculation of cross-phenotype associations is typically < 2 min.

Discussion

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

In examining cross-phenotype connections, it is important to carefully examine the overlapping SNPs provided as part of the iCPAGdb output to determine (1) the genome location where the variants are located, as some may be adjacent/overlapping loci in weak LD and not truly distinct and (2) how well identified GWAS signals from two traits overlap. Indeed, we view iCPAGdb as the first step in a pipeline for gaining greater understanding of any GWAS that then moves to colocalization analysis (see Figs. 3a, 4b, 5b; Additional file 7: Table S3; Additional file 9: Table S5; Additional file 10: Table S6), to further dissect GWAS signals in the same region. Making summary statistics more readily available for all GWAS, especially earlier studies in NHGRI-EBI GWAS, would facilitate these validation studies. Finally, functional studies in model systems and clinical studies are needed to test the proposed hypothesis and deeply understand the underlying mechanisms.

While the current web implementation of iCPAGdb uses NHGRI-EBI GWAS catalog [12], H2P2 [16], and metabolomics GWAS datasets [17, 18], additional datasets of molecular, cellular, and disease GWAS can be easily added. Analysis of user-uploaded GWAS may be the most useful application of iCPAGdb and will lead to discovery of new connections among human phenotypes to encourage experimental and clinical follow-up studies. Our studies of COVID-19 provide a test case for this and revealed possible mechanisms underlying the associations of severe COVID-19 with *ABO* and *DPP9.*

While our work highlights shared genetic architecture regulating *ABO*, protein abundance, and COVID-19, much work remains to be done to understand the mechanisms underlying these connections. The *ABO* locus controls abundance of many proteins. Some of these proteins, such as VWF and Factor VIII, have already been shown to be regulated by glycosylation of ABO [80–82]. For CD209, *ABO* is a protein quantitative trait locus, but it is unknown whether CD209 protein abundance is regulated by ABO glycosylation. CD209 has a predicted N-linked glycosylation site (N80) and glycosylation has been observed by mass spectrometry (http://glycositeatlas.biomarkercenter.org/glycosites/33001/) [101]. Whether human genetic variation also impacts CD209 glycosylation is also an unanswered question. Previous studies have examined protein glycosylation as a GWAS trait, resulting in 16 genome-wide significant loci [102–104], 15 of which have been recently replicated [105]. However, these studies quantified total plasma N-glycans released from proteins and did not specifically quantify glycosylation and glycoforms for individual proteins. Future GWAS quantifying individual glycosylated protein isoforms, as well as other post-translational modifications, may therefore be valuable.

The shared underlying genetic risk factors for IPF and COVID-19 suggest that *DPP9* may have a common role in pathogenesis in these diseases. iCPAGdb was able to identify this connection in the first published COVID19 GWAS despite the *DPP9* allele being below genome-wide significance in that cohort, demonstrating the utility of iCPAGdb in expanding the power of GWAS studies on emerging and understudied diseases. We speculate that characteristics of inflammasome-mediated responses, normally suppressed by high expression of *DPP9*, may predispose to fibrosis. The shared genetic architecture also suggests that therapeutic approaches targeting fibrosis may be beneficial in both conditions. Pirfenidone and Nintedanib are anti-fibrotic FDA-approved drugs used to treat IPF, and our findings support the idea that these drugs may prove beneficial in COVID-19 [106–108].

Conclusions

As our examination of COVID-19 demonstrates, iCPAGdb is a powerful hypothesis engine that will lead to a deeper understanding of the genetic underpinnings of human disease risk, severity, and drug response.

Availability and requirements

Project name: iCPAGdb

Project home page: http://cpag.oit.duke.edu and https://github.com/tbalmat/iCPAGdb [14]

Operating system(s): Red Hat Enterprise Linux OS, v7.9

Programming language: Python v3.6 and R Shiny v1.5

Other requirements: PLINK v1.9; R packages: shiny, shinyjs, DT, RSQLite, ggplot2, heatmaply, plotly, future; Python packages: panda, scipy, joblib, tqdm, sqlite

License: GPL v3

Any restrictions to use by non-academics: No restriction for non-academic use

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00904-z.

Additional file 1. iCPAGdb Web App Supplement.

Additional file 2: Figure S1. Clumping of GWAS results from NHGRI-EBI GWAS catalog.

Additional file 3: Table S1. iCPAGdb output for cross-phenotype associations from NHGRI-EBI GWAS catalog (downloaded on August 5, 2020).

Additional file 4: Figure S2. Histogram of shared SNPs for each trait pair in NHGRI-EBI GWAS catalog from iCPAGdb at false discovery rate of 0.1.

Additional file 5: Figure S3. Comparison of CPAG1 and iCPAGdb using the NHGRI-EBI GWAS catalog summary statistics downloaded on September 4, 2013.

Additional file 6: Table S2. iCPAGdb output for cross-phenotype associations between molecular and cellular datasets and NHGRI-EBI GWAS catalog.

Additional file 7: Table S3. COLOC analysis of *CXCL10* following *C. trachomatis* infection and levels of *CXCL9* (MIG) in whole blood.

Additional file 8: Table S4. Cross-phenotype associations for COVID-19 (p value < 1×10^{-5}) against 4418 GWAS traits in iCPAGdb (p value < 5×10^{-8}).

Additional file 9: Table S5. COLOC analysis output for *ABO* region between COVID-19 and CD209 antigen levels.

Additional file 10: Table S6. COLOC analysis output for *DPP9* region between COVID-19 and idiopathic pulmonary fibrosis.

Acknowledgements

We thank Benjamin Schott, Jeffrey Bourgeois, and Kyle Gibbs for thoughtful discussions. We thank Dr. Richard J. Allen and colleagues for sharing idiopathic pulmonary fibrosis GWAS summary statistics.

Authors' contributions

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

Funding

LW, TJB, AI, MRD, ERH, and DCK were supported by NIH R21AI133305. LW, ALA, and DCK were supported by NIH R01AI118903. FJC, RH, TWB, MTM, XS, ELT, ERK, and CWW were supported by DARPA/USMRAA W911NF1920111 and NIH UM1AI104681. The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Institutes of Health.

Availability of data and materials

Web resources:
iCPAGdb website: http://cpag.oit.duke.edu
iCPAGdb GitHub: https://github.com/tbalmat/iCPAGdb [14]
NHGRI GWAS Catalog: https://www.ebi.ac.uk/gwas/ [12]
H2P2 cellular GWAS: http://h2p2.oit.duke.edu [16]
Human metabolite GWAS summary statistics: http://metabolomics.helmholtz-muenchen.de/gwas/index.php?task=download [17, 18]
COVID-19 GWAS summary statistics from Ellinghaus et al. (2020) [15] were downloaded from the GRASP webpage of aggregated COVID-19 GWAS results [10]: https://grasp.nhlbi.nih.gov/Covid19GWASResults.aspx
IPF GWAS: download link was obtained by applying for access following the collaborative protocol from https://github.com/genomicsITER/PFgenetics [19]
RNA-seq datasets generated in [24] are publicly available through the National Center for Biotechnology Information Gene Expression Omnibus, accession# GSE161731 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161731
Tools for visualization:
R packages:
ggplot2: https://cran.r-project.org/web/packages/ggplot2/ [109]
gggenes: https://cran.r-project.org/web/packages/gggenes/index.html [110]
tidygraph: https://cran.r-project.org/web/packages/tidygraph/ [111]
ggnetwork: https://cran.r-project.org/web/packages/ggnetwork/ [112]
circlize: https://cran.r-project.org/web/packages/circlize/ [113]
ggpubr: https://cran.r-project.org/web/packages/ggpubr/ [114]
DT: https://cran.r-project.org/web/packages/DT [115]
plotly: https://cran.r-project.org/web/packages/plotly/ [116]
heatmaply: https://cran.r-project.org/web/packages/heatmaply/ [117]
promises: https://CRAN.R-project.org/package=promises [118]
All iCPAGdb output described in this manuscript are available for browsing from http://cpag.oit.duke.edu. Supplemental files also contain iCPAGdb output and COLOC analysis results. Code is available at GitHub https://github.com/tbalmat/iCPAGdb [14].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Molecular Genetics and Microbiology, School of Medicine, Duke University, 0049 CARL Building Box 3053, 213 Research Drive, Durham, NC 27710, USA. [2]Duke Research Computing, Duke University, Durham, NC 27710, USA. [3]Center for Applied Genomics and Precision Medicine, Department of Medicine, Duke University, Durham, NC 27710, USA. [4]Durham Veterans Affairs Health Care System, Durham, NC 27705, USA. [5]Division of Infectious Diseases, Department of Medicine, Duke University Medical Center, Durham, NC 27710, USA. [6]Department of Hospital Medicine, Duke Regional Hospital, Durham, NC 27705, USA. [7]Department of Biomedical Engineering, Woo Center for Big Data and Precision Health, Duke University, Durham, NC 27710, USA. [8]Duke Molecular Physiology Institute and Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710, USA. [9]Cooperative Studies Program Epidemiology Center-Durham, Durham VA Health Care System, Durham, NC 27705, USA.

Received: 11 January 2021 Accepted: 5 May 2021
Published online: 17 May 2021

References

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101(1):5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.
2. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018; 562(7726):203–9. https://doi.org/10.1038/s41586-018-0579-z.
3. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4(1):13. https://doi.org/10.1186/1755-8794-4-13.
4. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291–5.
5. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28(19):2540–2. https://doi.org/10.1093/bioinformatics/bts474.
6. Zhu Z, Anttila V, Smoller JW, Lee PH. Statistical power and utility of meta-analysis methods for cross-phenotype genome-wide association studies. PLoS One. 2018;13(3):e0193256. https://doi.org/10.1371/journal.pone.0193256.
7. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383. https://doi.org/10.1371/journal.pgen.1004383.
8. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205–10.
9. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics. 2016;32(20):3207–9. https://doi.org/10.1093/bioinformatics/btw373.
10. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014;30(12):i185–94. https://doi.org/10.1093/bioinformatics/btu273.
11. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018;50(11):1593–9. https://doi.org/10.1038/s41588-018-0248-z.
12. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–D12. https://doi.org/10.1093/nar/gky1120.
13. Wang L, Oehlers SH, Espenschied ST, Rawls JF, Tobin DM, Ko DC. CPAG: software for leveraging pleiotropy in GWAS to reveal similarity between human traits links plasma fatty acids and intestinal inflammation. Genome Biol. 2015;16(1):190. https://doi.org/10.1186/s13059-015-0722-1.
14. Wang L, Balmat T, Ko DC. iCPAGdb: GitHub; 2021. Available from: https://github.com/tbalmat/iCPAGdb.
15. Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, et al. Genomewide association study of severe COVID-19 with respiratory failure. N Engl J Med. 2020;383:1522–34.
16. Wang L, Pittman KJ, Barker JR, Salinas RE, Stanaway IB, Williams GD, et al. An atlas of genetic variation linking pathogen-induced cellular traits to human disease. Cell Host Microbe. 2018;24(2):308–23 e6.
17. Raffler J, Friedrich N, Arnold M, Kacprowski T, Rueedi R, Altmaier E, et al. Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality. PLoS Genet. 2015;11(9):e1005487. https://doi.org/10.1371/journal.pgen.1005487.
18. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014; 46(6):543–50. https://doi.org/10.1038/ng.2982.
19. Allen RJ, Guillen-Guio B, Oldham JM, Ma SF, Dressen A, Paynton ML, et al. Genome-wide association study of susceptibility to idiopathic pulmonary

fibrosis. Am J Respir Crit Care Med. 2020;201(5):564–74. https://doi.org/10.1164/rccm.201905-1017OC.

20. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.

21. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74.

22. Li MX, Yeung JM, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Human genetics. 2012;131(5):747–56. https://doi.org/10.1007/s00439-011-1118-2.

23. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47(11):1236–41. https://doi.org/10.1038/ng.3406.

24. McClain MT, Constantine FJ, Henao R, Liu Y, Tsalik EL, Burke TW, et al. Dysregulated transcriptional responses to SARS-CoV-2 in the periphery. Nat Commun. 2021;12(1):1079. https://doi.org/10.1038/s41467-021-21289-y.

25. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.

26. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25.

27. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

28. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

29. Team RC. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2020. http://www.R-project.org/

30. Cheng W, Cheng J, Allaire JJ, Xie Y, McPherson J. Shiny: web application framework for R. 1.5.0 ed; 2020.

31. RStudio. Shiny Server: Put Shiny Web Apps Online. 1.5.0 ed. 2020.

32. Hipp RD. SQLite; 2020.

33. Muller K, Wickham H, James DA, Falcon S. RSQLite: 'SQLite' Interface for R. 2.2.1 ed; 2020.

34. Bengtsson H. A unifying framework for parallel and distributed processing in r using futures; 2020.

35. Bodofsky S, Merriman TR, Thomas TJ, Schlesinger N. Advances in our understanding of gout as an auto-inflammatory disease. Semin Arthritis Rheum. 2020;50(5):1089–100. https://doi.org/10.1016/j.semarthrit.2020.06.015.

36. Chen CJ, Tseng CC, Yen JH, Chang JG, Chou WC, Chu HW, et al. ABCG2 contributes to the development of gout and hyperuricemia in a genome-wide association study. Sci Rep. 2018;8(1):3137. https://doi.org/10.1038/s41598-018-21425-7.

37. Lai HM, Chen CJ, Su BY, Chen YC, Yu SF, Yen JH, et al. Gout and type 2 diabetes have a mutual inter-dependent effect on genetic risk factors and higher incidences. Rheumatology (Oxford). 2012;51(4):715–20. https://doi.org/10.1093/rheumatology/ker373.

38. Lee MG, Hsu TC, Chen SC, Lee YC, Kuo PH, Yang JH, et al. Integrative genome-wide association studies of eQTL and GWAS data for gout disease susceptibility. Sci Rep. 2019;9(1):4981. https://doi.org/10.1038/s41598-019-41434-4.

39. Li C, Li Z, Liu S, Wang C, Han L, Cui L, et al. Genome-wide association analysis identifies three new risk loci for gout arthritis in Han Chinese. Nat Commun. 2015;6(1):7041. https://doi.org/10.1038/ncomms8041.

40. Nakayama A, Nakaoka H, Yamamoto K, Sakiyama M, Shaukat A, Toyoda Y, et al. GWAS of clinically defined gout and subtypes identifies multiple susceptibility loci that include urate transporter genes. Ann Rheum Dis. 2017;76(5):869–77. https://doi.org/10.1136/annrheumdis-2016-209632.

41. Nakayama A, Nakatochi M, Kawamura Y, Yamamoto K, Nakaoka H, Shimizu S, et al. Subtype-specific gout susceptibility loci and enrichment of selection pressure on ABCG2 and ALDH2 identified by subtype genome-wide meta-analyses of clinically defined gout patients. Ann Rheum Dis. 2020;79(5):657–65. https://doi.org/10.1136/annrheumdis-2019-216644.

42. Sulem P, Gudbjartsson DF, Walters GB, Helgadottir HT, Helgason A, Gudjonsson SA, et al. Identification of low-frequency variants associated with gout and serum uric acid levels. Nat Genet. 2011;43(11):1127–30. https://doi.org/10.1038/ng.972.

43. Matsuo H, Yamamoto K, Nakaoka H, Nakayama A, Sakiyama M, Chiba T, et al. Genome-wide association study of clinically defined gout identifies multiple risk loci and its association with clinical subtypes. Ann Rheum Dis. 2016;75(4):652–9.

44. Dehghan A, Kottgen A, Yang Q, Hwang SJ, Kao WL, Rivadeneira F, et al. Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. Lancet. 2008;372(9654):1953–61.

45. Kottgen A, Albrecht E, Teumer A, Vitart V, Krumsiek J, Hundertmark C, et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. Nat Genet. 2013;45(2):145–54. https://doi.org/10.1038/ng.2500.

46. Tin A, Woodward OM, Kao WH, Liu CT, Lu X, Nalls MA, et al. Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele. Hum Mol Genet. 2011;20(20):4056–68.

47. Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat Genet. 2010;42(3):210–5.

48. Boocock J, Leask M, Okada Y, Asian Genetic Epidemiology Network C, Matsuo H, Kawamura Y, et al. Genomic dissection of 43 serum urate-associated loci provides multiple insights into molecular mechanisms of urate control. Hum Mol Genet. 2020;29(6):923–43. https://doi.org/10.1093/hmg/ddaa013.

49. Li S, Sanna S, Maschio A, Busonero F, Usala G, Mulas A, et al. The GLUT9 gene is associated with serum uric acid levels in Sardinia and Chianti cohorts. PLoS Genet. 2007;3(11):e194. https://doi.org/10.1371/journal.pgen.0030194.

50. Doring A, Gieger C, Mehta D, Gohlke H, Prokisch H, Coassin S, et al. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. Nat Genet. 2008;40(4):430–6.

51. Tin A, Marten J, Halperin Kuhns VL, Li Y, Wuttke M, Kirsten H, et al. Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. Nat Genet. 2019;51(10):1459–74. https://doi.org/10.1038/s41588-019-0504-x.

52. Dong Y, Zhao T, Ai W, Zalloum WA, Kang D, Wu T, et al. Novel urate transporter 1 (URAT1) inhibitors: a review of recent patent literature (2016-2019). Expert Opin Ther Pat. 2019;29(11):871–9. https://doi.org/10.1080/13543776.2019.1676727.

53. Thorleifsson G, Holm H, Edvardsson V, Walters GB, Styrkarsdottir U, Gudbjartsson DF, et al. Sequence variants in the CLDN14 gene associate with kidney stones and bone mineral density. Nat Genet. 2009;41(8):926–30. https://doi.org/10.1038/ng.404.

54. Oddsson A, Sulem P, Helgason H, Edvardsson VO, Thorleifsson G, Sveinbjornsson G, et al. Common and rare variants associated with kidney stones and biochemical traits. Nat Commun. 2015;6(1):7975. https://doi.org/10.1038/ncomms8975.

55. Howles SA, Wiberg A, Goldsworthy M, Bayliss AL, Gluck AK, Ng M, et al. Genetic variants of calcium and vitamin D metabolism in kidney stone disease. Nat Commun. 2019;10(1):5175.

56. Setoh K, Terao C, Muro S, Kawaguchi T, Tabara Y, Takahashi M, et al. Three missense variants of metabolic syndrome-related genes are associated with alpha-1 antitrypsin levels. Nat Commun. 2015;6(1):7754. https://doi.org/10.1038/ncomms8754.

57. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357.

58. Joosten LA, Crisan TO, Azam T, Cleophas MC, Koenders MI, van de Veerdonk FL, et al. Alpha-1-anti-trypsin-Fc fusion protein ameliorates gouty arthritis by reducing release and extracellular processing of IL-1beta and by the induction of endogenous IL-1Ra. Ann Rheum Dis. 2016;75(6):1219–27. https://doi.org/10.1136/annrheumdis-2014-206966.

59. Band G, Le QS, Jostins L, Pirinen M, Kivinen K, Jallow M, et al. Imputation-based meta-analysis of severe malaria in three African populations. PLoS Genet. 2013;9(5):e1003509. https://doi.org/10.1371/journal.pgen.1003509.

60. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet. 2009;41(6):657–65. https://doi.org/10.1038/ng.388.

61. Malaria Genomic Epidemiology N. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nat Commun. 2019;10(1):5732.

62. Malaria Genomic Epidemiology N, Band G, Rockett KA, Spencer CC, Kwiatkowski DP. A novel locus of resistance to severe malaria in a region of

ancient balancing selection. Nature. 2015;526(7572):253–7. https://doi.org/10.1038/nature15390.

63. Ravenhall M, Campino S, Sepulveda N, Manjurano A, Nadjm B, Mtove G, et al. Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. PLoS Genet. 2018;14(1): e1007172. https://doi.org/10.1371/journal.pgen.1007172.
64. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature. 2012;489(7416):443–6. https://doi.org/10.1038/nature11334.
65. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell. 2016;167(5):1415–29 e19.
66. Chen MH, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell. 2020;182(5):1198–213 e14.
67. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging polygenic functional enrichment to improve GWAS power. Am J Hum Genet. 2019;104(1):65–75. https://doi.org/10.1016/j.ajhg.2018.11.008.
68. Fatumo S, Carstensen T, Nashiru O, Gurdasani D, Sandhu M, Kaleebu P. Complimentary methods for multivariate genome-wide association study identify new susceptibility genes for blood cell traits. Front Genet. 2019;10: 334.
69. Chen Z, Tang H, Qayyum R, Schick UM, Nalls MA, Handsaker R, et al. Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Hum Mol Genet. 2013;22(12):2529–38. https://doi.org/10.1093/hmg/ddt087.
70. Scerri TS, Quaglieri A, Cai C, Zernant J, Matsunami N, Baird L, et al. Genome-wide analyses identify common variants associated with macular telangiectasia type 2. Nat Genet. 2017;49(4):559–67. https://doi.org/10.1038/ng.3799.
71. Gantner ML, Eade K, Wallace M, Handzlik MK, Fallon R, Trombley J, et al. Serine and lipid metabolism in macular disease and peripheral neuropathy. N Engl J Med. 2019;381(15):1422–33. https://doi.org/10.1056/NEJMoa1815111.
72. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 2011;7(11): e1002355. https://doi.org/10.1371/journal.pgen.1002355.
73. Pittman KJ, Glover LC, Wang L, Ko DC. The legacy of past pandemics: common human mutations that protect against infectious disease. PLoS Pathog. 2016;12(7):e1005680. https://doi.org/10.1371/journal.ppat.1005680.
74. Ahola-Olli AV, Wurtz P, Havulinna AS, Aalto K, Pitkanen N, Lehtimaki T, et al. Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. Am J Hum Genet. 2017;100(1): 40–50. https://doi.org/10.1016/j.ajhg.2016.11.007.
75. Tanikawa C, Urabe Y, Matsuo K, Kubo M, Takahashi A, Ito H, et al. A genome-wide association study identifies two susceptibility loci for duodenal ulcer in the Japanese population. Nat Genet. 2012;44(4):430–4, S1-2. https://doi.org/10.1038/ng.1109.
76. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet. 2009;41(9):986–90. https://doi.org/10.1038/ng.429.
77. Shah S, Henry A, Roselli C, Lin H, Sveinbjornsson G, Fatemifar G, et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. Nat Commun. 2020;11(1):163. https://doi.org/10.1038/s41467-019-13690-5.
78. Zhao J, Yang Y, Huang H, Li D, Gu D, Lu X, et al. Relationship between the ABO blood group and the COVID-19 susceptibility. Clin Infect Dis. 2020. https://doi.org/10.1093/cid/ciaa1150.
79. Wool GD, Miller JL. The impact of COVID-19 disease on platelets and coagulation. Pathobiology. 2021;88(1):15–27.
80. Canis K, Anzengruber J, Garenaux E, Feichtinger M, Benamara K, Scheiflinger F, et al. In-depth comparison of N-glycosylation of human plasma-derived factor VIII and different recombinant products: from structure to clinical implications. J Thromb Haemost. 2018;16(8):1592–603. https://doi.org/10.1111/jth.14204.
81. Matsui T, Titani K, Mizuochi T. Structures of the asparagine-linked oligosaccharide chains of human von Willebrand factor. Occurrence of blood group A, B, and H(O) structures. J Biol Chem. 1992;267(13):8723–31. https://doi.org/10.1016/S0021-9258(19)50338-6.
82. Sodetz JM, Paulson JC, McKee PA. Carbohydrate composition and identification of blood group A, B, and H oligosaccharide structures on human factor VIII/von Willebrand factor. J Biol Chem. 1979;254(21):10754–60. https://doi.org/10.1016/S0021-9258(19)86586-9.
83. Gallinaro L, Cattini MG, Sztukowska M, Padrini R, Sartorello F, Pontara E, et al. A shorter von Willebrand factor survival in O blood group subjects explains how ABO determinants influence plasma von Willebrand factor. Blood. 2008;111(7):3540–5.
84. Shima M, Fujimura Y, Nishiyama T, Tsujiuchi T, Narita N, Matsui T, et al. ABO blood group genotype and plasma von Willebrand factor in normal individuals. Vox Sang. 1995;68(4):236–40. https://doi.org/10.1111/j.1423-0410.1995.tb02579.x.
85. Albanez S, Ogiwara K, Michels A, Hopman W, Grabell J, James P, et al. Aging and ABO blood type influence von Willebrand factor and factor VIII levels through interrelated mechanisms. J Thromb Haemost. 2016;14(5):953–63. https://doi.org/10.1111/jth.13294.
86. Song J, Chen F, Campos M, Bolgiano D, Houck K, Chambless LE, et al. Quantitative influence of ABO blood groups on factor VIII and its ratio to von Willebrand factor, novel observations from an ARIC study of 11,673 subjects. PLoS One. 2015;10(8):e0132626. https://doi.org/10.1371/journal.pone.0132626.
87. Murray GP, Post SR, Post GR. ABO blood group is a determinant of von Willebrand factor protein levels in human pulmonary endothelial cells. J Clin Pathol. 2020;73(6):347–9. https://doi.org/10.1136/jclinpath-2019-206182.
88. Mangalmurti N, Hunter CA. Cytokine storms: understanding COVID-19. Immunity. 2020;53(1):19–25. https://doi.org/10.1016/j.immuni.2020.06.017.
89. Katz DH, Tahir UA, Ngo D, Benson MD, Bick AG, Pampana A, et al. Proteomic profiling in biracial cohorts implicates DC-SIGN as a mediator of genetic risk in COVID-19. medRxiv. 2020. https://doi.org/10.1101/2020.06.09.20125690.
90. Amraie R, Napoleon MA, Yin W, Berrigan J, Suder E, Zhao G, et al. CD209L/L-SIGN and CD209/DC-SIGN act as receptors for SARS-CoV-2 and are differentially expressed in lung and kidney epithelial and endothelial cells. bioRxiv. 2020. https://doi.org/10.1101/2020.06.22.165803.
91. Gao C, Zeng J, Jia N, Stavenhagen K, Matsumoto Y, Zhang H, et al. SARS-CoV-2 spike protein interacts with multiple innate immune receptors. bioRxiv. 2020. https://doi.org/10.1101/2020.07.29.227462.
92. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet. 2013;45(6):613–20. https://doi.org/10.1038/ng.2609.
93. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis. 2020;20(4):425–34.
94. Ojo AS, Balogun SA, Williams OT, Ojo OS. Pulmonary fibrosis in COVID-19 survivors: predictive factors and risk reduction strategies. Pulm Med. 2020; 2020:6175964.
95. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021;591(7848): 92–8.
96. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348(6235):648–60. https://doi.org/10.1126/science.1262110.
97. Raj VS, Mou H, Smits SL, Dekkers DH, Muller MA, Dijkman R, et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. Nature. 2013;495(7440):251–4. https://doi.org/10.1038/nature12005.
98. Okondo MC, Johnson DC, Sridharan R, Go EB, Chui AJ, Wang MS, et al. DPP8 and DPP9 inhibition induces pro-caspase-1-dependent monocyte and macrophage pyroptosis. Nat Chem Biol. 2017;13(1):46–53.
99. Okondo MC, Rao SD, Taabazuing CY, Chui AJ, Poplawski SE, Johnson DC, et al. Inhibition of Dpp8/9 Activates the Nlrp1b Inflammasome. Cell Chem Biol. 2018; 25(3):262–7 e5. https://doi.org/10.1016/j.chembiol.2017.12.013.
100. Zhong FL, Robinson K, Teo DET, Tan KY, Lim C, Harapas CR, et al. Human DPP9 represses NLRP1 inflammasome and protects against autoinflammatory diseases via both peptidase activity and FIIND domain binding. J Biol Chem. 2018;293(49):18864–78. https://doi.org/10.1074/jbc.RA118.004350.
101. Sun S, Hu Y, Ao M, Shah P, Chen J, Yang W, et al. N-GlycositeAtlas: a database resource for mass spectrometry-based human N-linked glycoprotein and glycosylation site mapping. Clin Proteomic. 2019;16(1):35. https://doi.org/10.1186/s12014-019-9254-0.

102. Lauc G, Essafi A, Huffman JE, Hayward C, Knezevic A, Kattla JJ, et al. Genomics meets glycomics-the first GWAS study of human N-Glycome identifies HNF1alpha as a master regulator of plasma protein fucosylation. PLoS Genet. 2010;6(12):e1001256. https://doi.org/10.1371/journal.pgen.1001256.
103. Sharapov SZ, Tsepilov YA, Klaric L, Mangino M, Thareja G, Shadrina AS, et al. Defining the genetic control of human blood plasma N-glycome using genome-wide association study. Hum Mol Genet. 2019;28(12):2062–77. https://doi.org/10.1093/hmg/ddz054.
104. Huffman JE, Knezevic A, Vitart V, Kattla J, Adamczyk B, Novokmet M, et al. Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. Hum Mol Genet. 2011;20(24):5000–11. https://doi.org/10.1093/hmg/ddr414.
105. Sharapov SZ, Shadrina AS, Tsepilov YA, Elgaeva EE, Tiys ES, Feoktistova SG, et al. Replication of fifteen loci involved in human plasma protein N-glycosylation in 4,802 samples from four cohorts. Glycobiology. 2021;31(2):82–8.
106. George PM, Wells AU, Jenkins RG. Pulmonary fibrosis and COVID-19: the potential role for antifibrotic therapy. Lancet Respir Med. 2020;8(8):807–15.
107. Seifirad S. Pirfenidone: a novel hypothetical treatment for COVID-19. Med Hypotheses. 2020;144:110005.
108. Ferrara F, Granata G, Pelliccia C, La Porta R, Vitiello A. The added value of pirfenidone to fight inflammation and fibrotic state induced by SARS-CoV-2: anti-inflammatory and anti-fibrotic therapy could solve the lung complications of the infection? Eur J Clin Pharmacol. 2020;76(11):1615–8.
109. Wickham H. Ggplot2: elegant graphics for data analysis, vol. viii. New York: Springer; 2009. p. 212.
110. Wilkins D. gggenes: Draw Gene Arrow Maps in 'ggplot2'. R package version 0.4.1. 2020. Available from: https://CRAN.R-project.org/package=gggenes.
111. Pedersen TL. tidygraph: A tidy API for graph manipulation. R package version 1.2.0. 2020. Available from: https://CRAN.R-project.org/package=tidygraph.
112. Briatte F. ggnetwork: geometries to plot networks with 'ggplot2'. R package version 0.5.8. 2020. Available from: https://github.com/briatte/ggnetwork.
113. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30(19):2811–2. https://doi.org/10.1093/bioinformatics/btu393.
114. Kassambara A. ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0. 2020. Available from: https://CRAN.R-project.org/package=ggpubr.
115. Xie Y, Cheng J, Tan X. DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.17. 2021. Available from: https://CRAN.R-project.org/package=DT.
116. Sievert C. Interactive web-based data visualization with R, plotly, and shiny. Boca Raton: CRC Press, Taylor and Francis Group; 2020. pages cm p
117. Galili T, O'Callaghan A, Sidi J, Sievert C. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics. 2018;34(9):1600–2.
118. Cheng J. promises: abstractions for promise-based asynchronous programming. R package v1.1.1. 2020. Available from: https://CRAN.R-project.org/package=promises.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.