Whole genome sequencing identifies multiple loci for critical illness caused by COVID-19

Athanasios Kousathanas[‡,1], Erola Pairo-Castineira[‡,2,3], Konrad Rawlik[2], Alex Stuckey[1], Christopher A Odhams[1], Susan Walker[1], Clark D Russell[2,4], Tomas Malinauskas[5], Jonathan Millar[2], Katherine S Elliott[5], Fiona Griffiths[2], Wilna Oosthuyzen[2], Kirstie Morrice[6], Sean Keating[7], Bo Wang[2], Daniel Rhodes[1], Lucija Klaric[3], Marie Zechner[2], Nick Parkinson[2], Andrew D. Bretherick[3], Afshan Siddiq[1], Peter Goddard[1], Sally Donovan[1], David Maslove[8], Alistair Nichol[9], Malcolm G Semple[10,11], Tala Zainy[1], Fiona Maleady-Crowe[1], Linda Todd[1], Shahla Salehi[1], Julian Knight[5], Greg Elgar[1], Georgia Chan[1], Prabhu Arumugam[1], Tom A Fowler[12,13], Augusto Rendon[1], Manu Shankar-Hari[14], Charlotte Summers[15], Paul Elliott[16], Jian Yang[17], Yang Wu, GenOMICC Investigators , 23andMe , Covid-19 Human Genetics Initiative , Angie Fawkes[6], Lee Murphy[6], Kathy Rowan[18], Chris P Ponting[3], Veronique Vitart[3], James F Wilson[3,19], Richard H Scott[1,20], Sara Clohisey[*,2], Loukas Moutsianas[*,1], Andy Law[*,2], Mark J Caulfield[*,12,21], J. Kenneth Baillie[*,2,3,4,7].

‡ - joint first authors

* - joint last authors

[1]Genomics England, London UK
[2]Roslin Institute, University of Edinburgh, Easter Bush, Edinburgh, EH25 9RG, UK
[3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK
[4]Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, 47 Little France Crescent, Edinburgh, UK
[5]Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK
[6]Edinburgh Clinical Research Facility, Western General Hospital, University of Edinburgh, EH4 2XU, UK
[7]Intensive Care Unit, Royal Infirmary of Edinburgh, 54 Little France Drive, Edinburgh, EH16 5SA, UK
[8]Department of Critical Care Medicine, Queen's University and Kingston Health Sciences Centre, Kingston, ON, Canada
[9]Clinical Research Centre at St Vincent's University Hospital, University College Dublin, Dublin, Ireland
[10]NIHR Health Protection Research Unit for Emerging and Zoonotic Infections, Institute of Infection, Veterinary and Ecological Sciences University of Liverpool, Liverpool, L69 7BE, UK
[11]Respiratory Medicine, Alder Hey Children's Hospital, Institute in The Park, University of Liverpool, Alder Hey Children's Hospital, Liverpool, UK

[12]Genomics England, Queen Mary University of London
[13]Test and Trace, the Health Security Agency, Department of Health and Social Care, Victoria St, London, UK
[14]Department of Intensive Care Medicine, Guy's and St. Thomas NHS Foundation Trust, London, UK
[15]Department of Medicine, University of Cambridge, Cambridge, UK
[16]Imperial College, London
[17]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang 310024, China
[18]Intensive Care National Audit & Research Centre, London, UK
[19]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, Teviot Place, Edinburgh EH8 9AG, UK
[20]Great Ormond Street Hospital, London UK
[21]William Harvey Research Institute, Queen Mary University of London, Charterhouse Square, London EC1 6BQ

Abstract

Critical illness in COVID-19 is caused by inflammatory lung injury, mediated by the host immune system. We and others have shown that host genetic variation influences the development of illness requiring critical care[1] or hospitalisation[2;3;4] following SARS-Co-V2 infection. The GenOMICC (Genetics of Mortality in Critical Care) study is designed to compare genetic variants in critically-ill cases with population controls in order to find underlying disease mechanisms.

Here, we use whole genome sequencing and statistical fine mapping in 7,491 critically-ill cases compared with 48,400 population controls to discover and replicate 22 independent variants that significantly predispose to life-threatening COVID-19. We identify 15 new independent associations with critical COVID-19, including variants within genes involved in interferon signalling (*IL10RB*, *PLSCR1*), leucocyte differentiation (*BCL11A*), and blood type antigen secretor status (*FUT2*). Using transcriptome-wide association and colocalisation to infer the effect of gene expression on disease severity, we find evidence implicating expression of multiple genes, including reduced expression of a membrane flippase (*ATP11A*), and increased mucin expression (*MUC1*), in critical disease.

We show that comparison between critically-ill cases and population controls is highly efficient for genetic association analysis and enables detection of therapeutically-relevant mechanisms of disease. Therapeutic predictions arising from these findings require testing in clinical trials.

Introduction

Critical illness in COVID-19 is both an extreme disease phenotype, and a relatively homogeneous clinical definition including patients with hypoxaemic respiratory failure [5] with acute lung injury, [6] and excluding many patients with non-pulmonary clinical presentations [7] who are known to have divergent responses to therapy. [8] In the UK, the critically-ill patient group is younger, less likely to have significant comorbidity, and more severely affected than a general hospitalised cohort, [5] characteristics which may amplify observed genetic effects. In addition, since development of critical illness is in itself a key clinical endpoint for therapeutic trials, [8] using critical illness as a phenotype in genetic studies enables detection of directly therapeutically-relevant genetic effects. [1]

Using microarray genotyping in 2,244 cases, we previously reported that critical COVID-19 is associated with genetic variation in the host immune response to viral infection (*OAS1*, *IFNAR2*, *TYK2*) and the inflammasome regulator *DPP9*. [1] In collaboration with international groups, we recently extended these findings to include a variant near *TAC4* (rs77534576). [4] Several variants have been associated with milder phenotypes, such as the need for hospitalisation or management in the community, including the ABO blood type locus, [3] a pleiotropic inversion in chr17q21.31, [9] and associations in 5 additional loci including the T lymphocyte-associated transcription factor, *FOXP4*. [4] An enrichment of rare loss-of-function variants in candidate interferon signalling genes has been reported, [2] but this has yet to be replicated at genome-wide significance thresholds. [10;11]

We established a partnership between the GenOMICC Study and Genomics England to perform whole genome sequencing (WGS) to improve resolution and deepen fine-mapping of significant signals to enhance the biological insights into critical COVID-19. Here, we present results from a cohort of 7,491 critically-ill patients from 224 intensive care units, compared with 48,400 population controls, describing discovery and validation of 22 gene loci for susceptibility to life-threatening COVID-19.

Results

Study design

Cases were defined by the presence of COVID-19 critical illness in the view of the treating clinician - specifically, the need for continuous cardio-respiratory monitoring. Patients were recruited from 224 intensive care units across the UK in the GenOMICC (Genetics Of Mortality In Critical Care) study. As a control population, unrelated participants recruited to the 100,000 Genomes Project were selected, excluding those with a known positive COVID-19 test, as severity information was not available. The 100,000 Genomes Project cohort (100k cohort) is comprised of UK individuals with a broad range of rare diseases or cancer and their family members. We included an additional prospectively-recruited cohort of volunteers (mild cohort) who self-reported testing positive for SARS-CoV-2 infection, and experienced mild or asymptomatic disease.

GWAS analysis

Whole genome sequencing and subsequent alignment and variant calling was performed for all subjects as described below (Methods). Following quality control procedures, we used a logistic mixed model regression, implemented in SAIGE, [12] to perform association analyses with unrelated

chr:pos (hg38)	rsid	REF	ALT	RAF	pop	OR	OR_{CI}	Pval	HetPVal	Consequence	Gene	Expression
1:155066988	rs114301457	C	T*	0.0058	EUR	2.40	1.82-3.16	6.8×10^{-10}	1	synonymous	*EFNA4*	-
1:155175305	rs7528026	G	A*	0.032	META	1.39	1.24-1.55	7.16×10^{-9}	0.96	intron	*TRIM46*	-
1:155197995	rs41264915	A*	G	0.89	EUR	1.28	1.19-1.37	1.02×10^{-12}	0.29	intron	*THBS3*	*MUC1*
2:60480453	rs1123573	A*	G	0.61	META	1.13	1.09-1.18	9.85×10^{-10}	0.29	intron	*BCL11A*	-
3:45796521	rs2271616	G	T*	0.14	EUR	1.29	1.21-1.37	9.9×10^{-17}	0.0011	5' UTR	*SLC6A20*	*SLC6A20, CCR5*
3:45859597	rs73064425	C	T*	0.077	EUR	2.71	2.51-2.94	1.97×10^{-133}	0.010	intron	*LZTFL1*	*LZTFL1, CCR9*
3:146517122	rs343320	G	A*	0.081	EUR	1.25	1.16-1.35	4.94×10^{-9}	0.53	missense	*PLSCR1*	-
5:131995059	rs56162149	C	T*	0.17	EUR	1.20	1.13-1.26	7.65×10^{-11}	0.17	intron	*ACSL6*	*ACSL6, FNIP1*
6:32623820	rs9271609	T*	C	0.65	EUR	1.14	1.09-1.19	3.26×10^{-9}	0.24	upstream	*HLA-DQA1*	*HLA-DQA1, HLA-DQA2*
6:41515007	rs2496644	A*	C	0.015	META	1.45	1.32-1.60	7.59×10^{-15}	0.49	intron	*LINC01276*	-
9:21206606	rs28368148	C	G*	0.013	EUR	1.74	1.45-2.09	1.93×10^{-9}	1	missense	*IFNA10*	-
11:34482745	rs61882275	G*	A	0.62	EUR	1.15	1.10-1.20	1.61×10^{-10}	0.29	intron	*ELF5*	-
12:132489230	rs56106917	GC	G*	0.49	EUR	1.13	1.09-1.18	2.08×10^{-9}	0.90	upstream	*FBRSL1*	-
13:112889041	rs9577175	C	T*	0.23	EUR	1.18	1.12-1.24	3.71×10^{-11}	0.10	downstream	*ATP11A*	*ATP11A*
15:93046840	rs4424872	T*	A	0.0079	EUR	2.37	1.87-3.01	8.61×10^{-13}	1.82×10^{-7}	intron	*RGMA*	-
16:89196249	rs117169628	G	A*	0.15	EUR	1.19	1.12-1.26	4.4×10^{-9}	0.80	missense	*SLC22A31*	*SLC22A31, CDH15*
17:46152620	rs2532300	T*	C	0.77	EUR	1.16	1.10-1.22	4.19×10^{-9}	0.32	intron	*KANSL1*	*ARHGAP27*
17:49863260	rs3848456	C	A*	0.029	EUR	1.50	1.33-1.70	4.19×10^{-11}	0.14	regulatory	.	-
19:4717660	rs12610495	A	G*	0.31	EUR	1.32	1.27-1.38	3.91×10^{-36}	0.069	missense	*DPP9*	-
19:10305768	rs73510898	G	A*	0.093	EUR	1.28	1.19-1.37	1.57×10^{-11}	0.011	intron	*ZGLP1*	-
19:10352442	rs34536443	G	C*	0.050	EUR	1.50	1.36-1.65	6.98×10^{-17}	0.63	missense	*TYK2*	*TYK2, PDE4A*
19:48697960	rs368565	C	T*	0.44	EUR	1.15	1.1-1.2	3.55×10^{-11}	0.22	intron	*FUT2*	*FUT2, NTN5, RASIP1*
21:33230000	rs17860115	C	A*	0.32	EUR	1.24	1.19-1.3	9.69×10^{-22}	0.63	5' UTR	*IFNAR2*	-
21:33287378	rs8178521	C	T*	0.27	EUR	1.18	1.12-1.23	3.53×10^{-12}	0.67	intron	*IL10RB*	-
21:33959662	rs35370143	T	TAC*	0.083	EUR	1.26	1.17-1.36	1.24×10^{-9}	1	intron	*LINC00649*	-

Table 1: Lead variants from independent regions in the per-population GWAS and trans-ancestry meta-analysis. Variants and the reference and alternate allele are reported with hg38 build coordinates. Asterisk (*) indicates the risk allele. For each variant, we report the risk allele frequency in Europeans (RAF), the odds ratio and 95% confidence interval, and the association P-value. Consequence indicates the worst consequence predicted by VEP99, and Gene indicates the VEP99-predicted gene, but not necessarily the causal mediator. Expression indicates genes where is evidence of gene expression affecting COVID-19 severity, found by TWAS and colocalisation analysis.

individuals (critically-ill cases $n = 7,491$, controls (100k) $n = 46,770$, controls (mild COVID-19) $n = 1,630$) (Methods, Supplementary Table 2). 1,339 of these cases were included in the primary analysis for our previous report.[1] Genome wide association studies (GWAS) were performed separately for genetically predicted ancestry groups (European - EUR, South Asian - SAS, African - AFR, East Asian - EAS, see Methods). Subsequently, we conducted inverse-variance weighted fixed effects meta-analysis across the four predicted ancestry cohorts using METAL[13] (Methods). In order to reduce the risk of spurious associations arising from genotyping or pipeline errors, we required supporting evidence from variants in linkage disequilibrium for all genome-wide significant variants: observed z-scores for each variant were compared to imputed z-scores for the same variant, with discrepant values being excluded (see Methods, Supplementary Figure 12).

In population-specific analyses, we discovered 22 independent genome-wide significant associations in the EUR ancestry group (Figure 1, Supplementary Figure 11 and Table 1) at a P-value threshold adjusted for multiple testing for 2,264,479 independent linkage disequilibrium-pruned genetic variants: 2.2×10^{-08} (Supplementary Table 3). The strong association at 3p21.31 also reached genome-wide significance in the SAS ancestry group (Supplementary Figure 11).

In trans-ancestry meta-analysis, we identified an additional three loci with genome-wide significant associations (Figure 1, Table 1). We tested the meta-analysed set of 25 loci for heterogeneity of effect size between predicted ancestries and detected significant (at $P < 1.83 \times 10^{-3}$) evidence for heterogeneity for two variants (Table 1, Supplementary Figure 13).

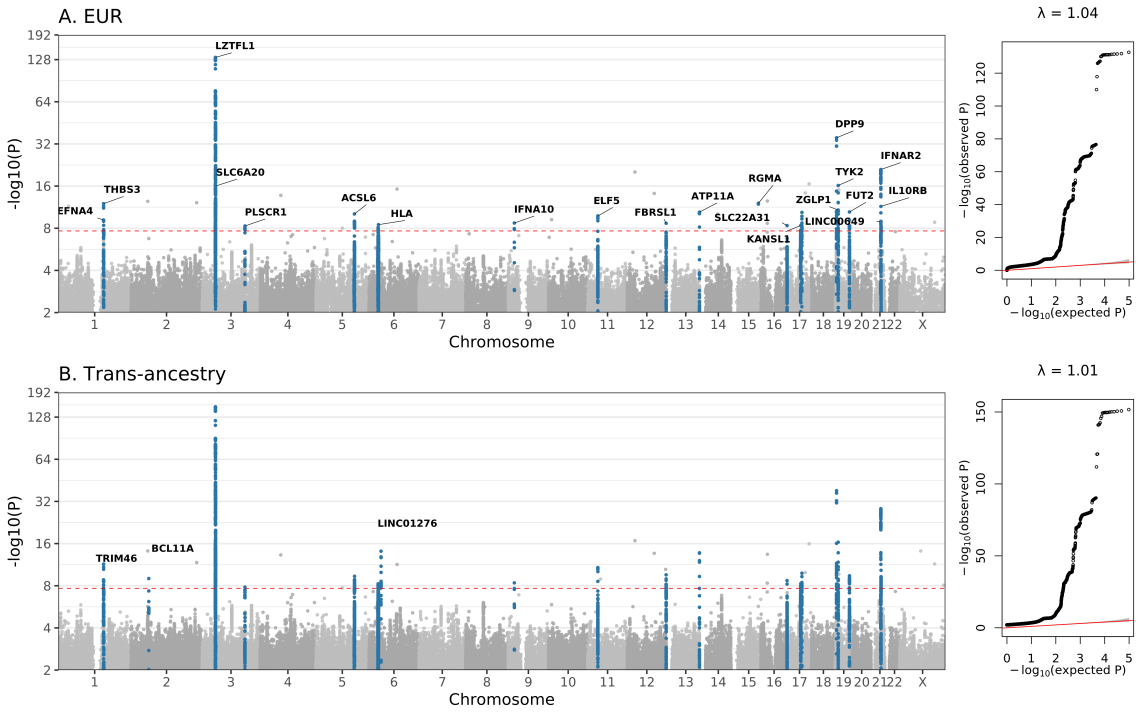

Figure 1: GWAS results for EUR ancestry group, and trans-ancestry meta-analysis. Manhattan plots are shown on the left and quantile–quantile (QQ) plots of observed versus expected P values are shown on the right, with genomic inflation (λ) displayed for each analysis. Highlighted results in blue in the Manhattan plots indicate variants that are LD-clumped (r^2=0.1, P_2=0.01, EUR LD) with the lead variants at each locus. Gene name annotation by Variant Effect Predictor (VEP) indicates genes impacted by the predicted consequence type of each lead variant. The red dashed line shows the Bonferroni-corrected P-value=2.2×10^{-8}.

chr:pos (hg38)	rsid	REF	ALT	OR	OR_{CI}	Pval	$OR_{hgi.23m}$	$OR_{CIhgi.23m}$	$Pval_{hgi.23m}$	Gene	Citation
1:155066988	rs114301457	C	T	2.40	1.81-3.18	1.51×10^{-9}	1.46	1.21-1.77	0.00011 *	*EFNA4*	-
1:155175305	rs7528026	G	A	1.39	1.24-1.55	7.16×10^{-9}	1.14	1.07-1.22	0.00012 *	*TRIM46*	-
1:155197995	rs41264915	A	G	0.80	0.76-0.86	3.79×10^{-12}	0.9	0.87-0.933	1.51×10^{-9} *	*THBS3*	-
2:60480453	rs1123573	A	G	0.88	0.85-0.92	9.85×10^{-10}	0.95	0.93-0.97	0.000018 *	*BCL11A*	-
3:45796521	rs2271616	G	T	1.26	1.19-1.34	2.45×10^{-15}	1.11	1.07-1.15	4.95×10^{-9} *	*SLC6A20*	([4])
3:45859597	rs73064425	C	T	2.52	2.35-2.70	2.18×10^{-152}	1.46	1.4-1.51	1.02×10^{-77} *	*LZTFL1*	[3]
3:146517122	rs343320	G	A	1.24	1.15-1.33	1.52×10^{-8}	1.08	1.04-1.13	0.00028 *	*PLSCR1*	-
5:132441275	rs10066378	T	C	1.20	1.13-1.27	4.48×10^{-10}	1.05	1.02-1.08	0.00074 *	*IRF1-AS1*	-
6:32623820	rs9271609	T	C	0.88	0.84-0.92	1.27×10^{-8}	1	0.98-1.03	0.89	*HLA-DQA1*	-
6:41515007	rs2496644	A	C	0.69	0.63-0.76	7.59×10^{-15}	0.87	0.83-0.92	3.17×10^{-7} *	*LINC01276*	-
9:21206606	rs28368148	C	G	1.74	1.45-2.1	4.09×10^{-9}	1.21	1.07-1.37	0.0024	*IFNA10*	-
11:34482745	rs61882275	G	A	0.87	0.84-0.91	1.62×10^{-11}	0.93	0.91-0.95	1.9×10^{-10} *	*ELF5*	-
12:132479205	rs4883585	G	A	1.13	1.09-1.18	1.12×10^{-9}	1.04	1.02-1.06	0.00047 *	*FBRSL1*	-
13:112889041	rs9577175	C	T	1.18	1.13-1.23	1.61×10^{-12}	1.07	1.04-1.09	1.29×10^{-6} *	*ATP11A*	-
15:93046840	rs4424872	T	A	0.64	0.53-0.76	1.99×10^{-6}	-	-	-	*RGMA*	-
16:89196249	rs117169628	G	A	1.18	1.12-1.25	6.04×10^{-9}	1.1	1.07-1.14	6.57×10^{-9} *	*SLC22A31*	-
17:46152620	rs2532300	T	C	0.87	0.82-0.91	1.4×10^{-8}	0.92	0.89-0.94	2.49×10^{-9} *	*KANSL1*	[9]
17:49863260	rs3848456	C	A	1.42	1.27-1.58	1.47×10^{-10}	1.15	1.09-1.21	1.34×10^{-7} *	.	[4]
19:4717660	rs12610495	A	G	1.32	1.27-1.38	6.44×10^{-39}	1.11	1.09-1.14	5.74×10^{-19} *	*DPP9*	[1]
19:10305768	rs73510898	G	A	1.24	1.16-1.33	1.47×10^{-9}	1.08	1.04-1.12	0.00016 *	*ZGLP1*	-
19:10352442	rs34536443	G	C	1.50	1.37-1.66	4.22×10^{-17}	1.22	1.15-1.29	4.06×10^{-11} *	*TYK2*	[1]
19:48697960	rs368565	C	T	1.13	1.09-1.18	3.74×10^{-10}	1.04	1.02-1.06	0.00087 *	*FUT2*	-
21:33230000	rs17860115	C	A	1.26	1.21-1.31	6.28×10^{-28}	1.11	1.08-1.13	1.77×10^{-18} *	*IFNAR2*	[1]
21:33287378	rs8178521	C	T	1.17	1.12-1.22	4.23×10^{-12}	1.06	1.03-1.09	8.02×10^{-6} *	*IL10RB*	-
21:33914436	rs12626438	A	G	1.22	1.14-1.31	1.78×10^{-8}	1.1	1.06-1.14	2.33×10^{-7} *	*LINC00649*	-

Table 2: Replication in a combined data from external studies - combined meta-analysis of HGI freeze 6 B2 and 23andMe. Odds ratios and *P*-values are shown for variants in LD with the lead variant that were genotyped/imputed in both sources. Chromosome, reference and alternate allele correspond to the build hg38. An asterisk (*) next to the HGI and 23andme meta-analysis *P*-value indicates that the lead signal is replicated with *P*-value<0.002 with a concordant direction of effect. Citation lists the first publication of confirmed genome-wide associations with critical illness or (in brackets) any COVID-19 phenotype.

Replication

Replication was performed using summary statistics generously shared by collaborators: data from the COVID-19 Host Genetics Initiative (HGI) data freeze 6 were combined using meta-analysis with data shared by 23andMe (Methods). Although the HGI programme included an analysis intended to mirror the GenOMICC study (analysis "A2"), there are currently insufficient cases from other sources available to attempt replication, so we used the broader hospitalised phenotype (analysis "B2") for replication. We removed signals in the HGI data derived from GenOMICC cases using mathematical subtraction (see Methods) to ensure independence. Using LD clumping to find variants genotyped in both the discovery and replication studies, we required $P < 0.002$ (0.05/25) and concordant direction of effect (Table 2) for replication.

We replicated 22 of the 25 significant associations identified in the population specific and/or trans-ancestry GWAS. Two of the three loci not replicated correspond to rare alleles that may not be well represented in the replication datasets which are dominated by SNP genotyping data. Although not replicated, for rs28368148 (9:21206606:C:G, *IFNA10*) we observed both a consistent direction of effect and odds ratio. The third locus is within the human leukocyte antigen (HLA) locus (see below).

We inferred credible sets of variants using Bayesian fine-mapping with susieR [14], by analysing the

GWAS summaries of 17 3Mbp regions that were flanking groups of lead signals. We obtained 22 independent credible sets of variants for EUR and one for SAS that each had posterior inclusion probability > 0.95.

Fine mapping of the association signals revealed putative causal variants for several genes (See Supplementary Information). For example, we detected variants at 3q24 and 9p21.3 predicted to be missense mutations by Variant Effect Predictor (VEP). These impact *PLSCR1* and *IFNA10* respectively, and both are predicted to be deleterious by the Combined Annotation Dependent Depletion (CADD) tool[15] (*PLSCR1* (chr3:146517122:G:A, rs343320,p.His262Tyr, OR:1.24, 95%CIs [1.15-1.33], CADD:22.6; *IFNA10* (chr9:21206606:C:G, rs28368148,p.Trp164Cys, OR:1.74, 95% CIs [1.45-2.09], CADD:23.9). Structural predictions for these loci suggest functional effects (Figure 3 and Supplementary Figure 15.

Gene burden testing

To assess the contribution of rare variants to critical illness, we performed gene-based analysis using SKAT-O as implemented in SAIGE-GENE[16], using a subset of 12,982 individuals from our cohort (7,491 individuals with critical COVID-19 and 5,391 controls) for which the genome sequencing data were processed with the same alignment and variant calling pipeline. We tested the burden of rare (MAF<0.5%) variants considering the predicted variant consequence type. We assessed burden using a strict definition for damaging variants (high-confidence loss-of-function (pLoF) variants as identified by LOFTEE[17]) and a lenient definition (pLoF plus missense variants with CADD ≥ 10)[15] , but found no significant associations at a gene-wide significance level. All individual rare variants included in the tests had P-values $>10^{-5}$.

We then further examined the association with 13 genes involved in the regulation of type I and III interferon immunity that were implicated in critical COVID-19 pneumonia[2] but, as with other recent studies[10], we did not find any significant gene burden test associations (tests for all genes had P-value>0.05, Supplementary File AVTsuppinfo.xlsx). We also did not replicate the reported association[10] for the toll-like receptor 7 (*TLR7*) gene.

Transcriptome-wide association study

In order to infer the effect of genetically-determined variation in gene expression on disease susceptibility, we performed a transcriptome-wide association study (TWAS) using gene expression data (GTExv8) for two disease-relevant tissues, lung and whole blood. We found 14 genes with significant association between predicted expression and critical COVID-19 in the lung and 6 in whole blood analyses (Supplementary File: TWAS.xlsx). To increase statistical power using eQTLs from multiple tissues, we performed a TWAS meta-analysis using all available tissues in GTExv8, revealing 51 transcriptome-wide significant genes. Since TWAS uses a composite signal derived from multiple eQTLs, we used colocalisation to find specific eQTLs in whole blood (eqtlGen and GTExv8) and lung (GTExv8[18]) which share the same signal with GWAS (EUR) associations. We found 16 genes which significantly colocalise in at least one of the studied tissues, shown in Figure 2.

We repeated the TWAS analysis using models of intron excision rate from GTExv8 to obtain splicing TWAS. We found 40 signals in lung, affecting 16 genes and 20 signals in whole blood which affect 9 genes. In a meta-analysis of splicing TWAS using all GTExv8 tissues, we found 91 significant introns in a total of 33 genes. Using GTExv8 lung and whole blood sqtls to find colocalising

signals with splicing TWAS significant results, we found 11 genes with colocalising splicing signals (Supplementary File: TWAS.xlsx).

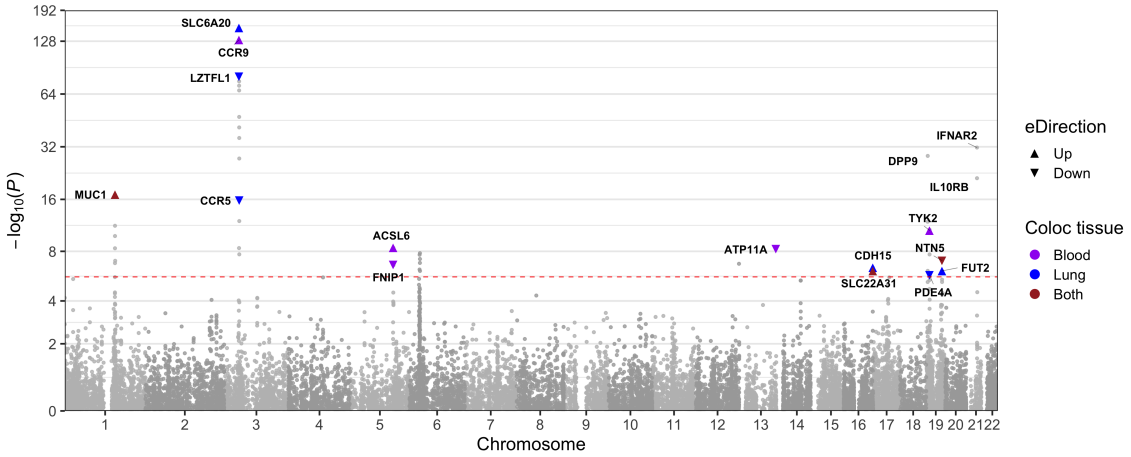

Figure 2: Gene-level Manhattan plot showing results from TWAS meta-analysis and highlighting genes that colocalise with GWAS signals or have strong metaTWAS associations. Highlighting color is different for lung and blood tissue data that were used for colocalisation. Arrows show direction of change in gene expression associated with an increased disease risk. Red dashed line shows significance threshold at $P < 2.3 \times 10^{-6}$.

HLA region

To investigate the contribution of specific HLA alleles to the observed association in the HLA region, we imputed HLA alleles at a four digit (two-field) level using HIBAG [19]. The only allele that reached genome-wide significance was HLA-DRB1*04:01 ($OR = 0.80, 95\%CI = 0.75 - 0.86, P = 1.6 \times 10^{-10}$ in EUR), which has a stronger P-value than the lead SNP in the region ($OR : 0.88, 95\%CIs : 0.84 - 0.92, P = 3.3 \times 10^{-9}$ in EUR) and is a better fit to the data ($AIC_{DRB1*04:01} = 30241.34$, $AIC_{leadSNP} = 30252.93$). Results are shown in supplementary figure 25.

Discussion

We report 22 replicated genetic associations with life-threatening COVID-19, and 3 additional loci, discovered in only 7,491 cases. This demonstrates the efficiency of the design of the GenOMICC study, which is an open-source international research programme [20] focusing on critically-ill patients with infectious disease and other critical illness phenotypes (https://genomicc.org). By using whole genome sequencing we were able to detect multiple distinct signals with high confidence for several of the associated loci, in some cases implicating different biological mechanisms.

Several variants associated with life-threatening disease are linked to interferon signalling. A coding variant in a ligand, *IFNA10A*, and reduced expression of its receptor *IFNAR2* (Figure 2), were associated with critical COVID-19. The narrow failure of replication for the *IFNA10* variant

(rs28368148, replication $P = 0.00243$, significance threshold $P < 0.002$) may be due to limited power in the replication cohort. The lead variant in *TYK2* in whole genome sequencing is a well-studied protein-coding variant with reduced phosphorylation activity, consistent with that reported recently,[4] but associated with significantly increased *TYK2* expression (Figure 2, Methods). Fine mapping reveals a significant critical illness association with an independent missense variant in *IL10RB*, a receptor for Type III (lambda) interferons (rs28368148,p.Trp164Cys, Table 1). Overall, variants predicted to be associated with reduction in interferon signalling are associated with critical disease. Importantly, systemic administration of interferon in a large clinical trial, albeit late in disease, did not reduce mortality.[21]

Phospholipid scramblase 1 (*PLSCR1*; chr3:146517122:G:A) functions as a nuclear signal for the antiviral effect of interferon,[22] and has been shown to control replication of other RNA viruses including vesicular stomatitis virus, encephalomyocarditis virus and Influenza A virus.[23;22] The risk allele at the lead variant (chr3:146517122:G:A, rs343320) encodes a substitution, H262Y, which is predicted to disrupt the non-canonical nuclear localisation signal[24] by eliminating a hydrogen bond with importin (Figure 3). Deletion of this nuclear localisation signal has been shown to prevent neutrophil maturation.[25] Although *PLSCR1* is strongly up-regulated when membrane lipid asymmetry is lost (see below), it may not act directly on this process.[26]

We report significant associations in several genes implicated in B-cell lymphopoesis and differentiation of myeloid cells. *BCL11A* is essential in B- and T-lymphopoiesis[27] and promotes plasmacytoid dendritic cell differentiation.[28] *TAC4*, reported previously,[4] encodes a regulator of B-cell lymphopoesis[29] and antibody production,[30] and promotes survival of dendritic cells.[31] Finally, although the strongest fine mapping signal at 5q31.1 (chr5:131995059:C:T, rs56162149) is in an intron of *ACSL6* (locus, p), the credible set includes a missense variant in *CSF2* of uncertain significance (chr5:132075767:T:C). *CSF2* encodes granulocyte-macrophage colony stimulating factor, a key differentiation factor in the mononuclear phagocyte system which is strongly up-regulated in critical COVID-19,[32] and is already under investigation as a target for therapy.[33]

Several new genetic associations implicate genes known to be involved in lung disease. The second variant in the credible set at 13q14 (chr13:112882313:A:G, rs1278769, in *ATP11A*), has been reported as a lead variant for idiopathic pulmonary fibrosis.[34] *ATP11A* encodes a flippase which maintains the asymmetric distribution of phospholipids in cell membranes;[35] disruption of this asymmetry is a phagocytic signal on apoptotic cells, and is required for platelet activation.[36;37] TWAS and colocalisation demonstrate that genetic variants predicted to decrease expression of *ATP11A* in lung are associated with critical illness. A combination of fine mapping, colocalisation with eQTL signals (GTEx and eQTLgen) and TWAS results provide evidence in support of *MUC1* as the mediator of the association with rs41264915 (Table 1). This may indicate an important role for mucins in the development of critical illness in COVID-19. The direction of effect (Figure 2) suggests that agents that reduce *MUC1* expression, and by extension its abundance, may be a therapeutic option. Finally, the association on 11p13 (rs61882275) includes GTEx eQTL for the lung fibroblast transcription factor *ELF5* in lung tissue, and the gene encoding the antioxidant enzyme catalase (*CAT*) in whole blood with evidence of colocalisation in both signals (supplementary material: TWAS.xlsx).[18] The protective allele at this locus is weakly associated with reduced lung function in a previous GWAS.[38]

FUT2 encodes alpha-(1,2)fucosyltransferase, which controls the secretion of ABO blood type glycans into body fluids and expression on epithelial surfaces. An association with critical COVID-19 was reported previously in a candidate gene association study by Mankelow et al.[39] The credible set for the

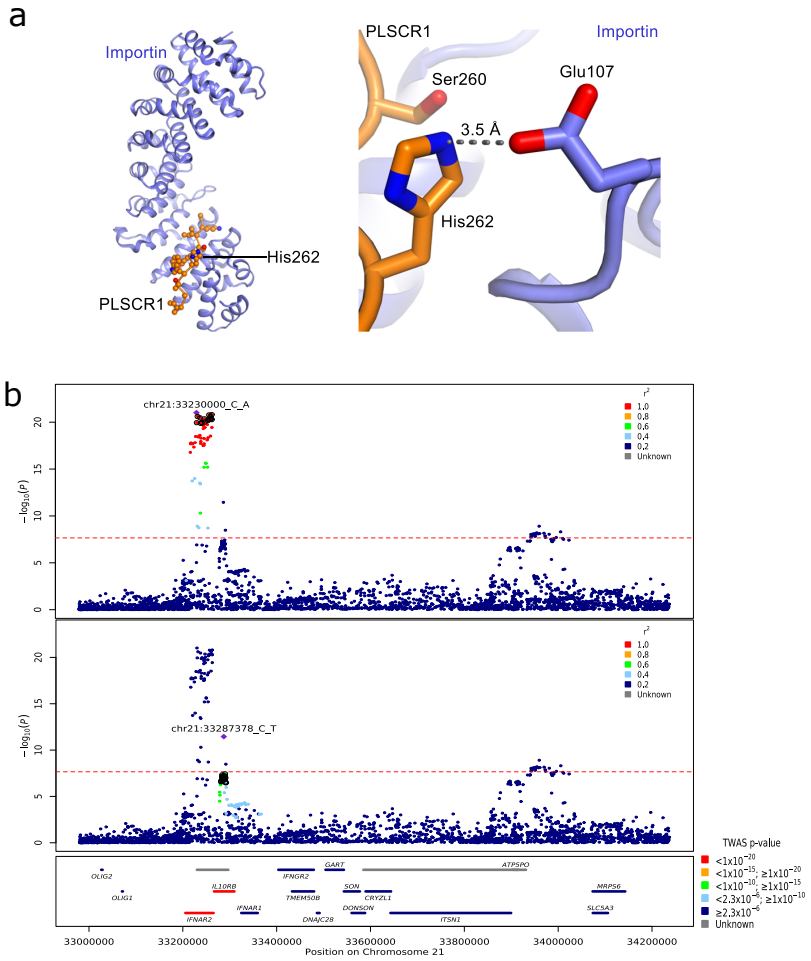

Figure 3: (a) Predicted structural consequences of lead variant at *PLSCR1*. Left panel shows the crystal structure of PLSCR1 nuclear localization signal (orange, Gly257–Ile266, numbering correspond to UniProt entry O15162) in complex with Importin α (blue), Protein Data Bank (PDB) ID 1Y2A. Side chains of PLSCR1 are shown as connected spheres with carbon atoms coloured in orange, nitrogens in blue and oxygens in red. Hydrogen atoms were not determined at this resolution (2.20 Å) and are not shown. Right panel: a closeup view showing side chains of PLSCR1 Ser260, His262 and Importin Glu107 as sticks. Distance (in Å) between selected atoms (PLSCR1 His262 $N\epsilon 2$ and Importin Glu107 carboxyl O) is indicated. A hydrogen bond between PLSCR1 His262 and Importin Glu107 is indicated with a dashed line. The risk variant is predicted to eliminate this bond, disrupting nuclear import, an essential step for effect on antiviral signalling [22] and neutrophil maturation. [25] (b) Regional detail showing fine-mapping to separate two adjacent independent signals. Top two panels: variants in linkage disequilibrium with the lead variants shown. The loci that are included in two independent credible sets are displayed with black outline circles. Bottom panel: locations of protein-coding genes, coloured by TWAS *P*-value.

FUT2 locus includes rs492602 (chr19:48703160:A:G) which is linked to a stop codon gain mutation (chr19:48703417:G:A), leading to the well-described non-secretor phenotype in homozygotes.[40;41] We show that the stop-gain, non-secretor allele is protective against life-threatening COVID-19. The protective variant in our study has been previously reported to protect against other viruses (rotavirus,[42] mumps and common colds[43]), to enhance antibody responses to polyomavirus BK[44] and to increase susceptibility to infection with some encapsulated bacteria.[45]

Limitations

In contrast to microarray genotyping, whole genome sequencing is rapidly evolving and a relatively new technology for genome-wide association studies, with relatively few sources of population controls. We used selected controls from the 100,000 genomes project, sequenced on a different platform (illumina HiSeqX) from the cases (illumina NovaSeq6000)(Supplementary Table 1). To minimise the risk of false positive associations arising due to sequencing or genotyping errors, we required all significant associations to be supported by local variants in linkage disequilibrium, which may be excessively stringent (see Methods). Although this approach may remove some true associations, our priority is to maximise confidence in the reported signals. Of 25 variants meeting this requirement, 22 are replicated in an independent study, and the remaining 3 may well be true associations that have failed due to a lack of coverage or power in the replication dataset.

The design of our study incorporates genetic signals for every stage in the disease progression into a single phenotype. This includes exposure, viral replication, inflammatory lung injury and hypoxaemic respiratory failure. Although we can have considerable confidence that the replicated associations with critical COVID-19 we report are robust, we cannot determine at which stage in the disease process, or in which tissue, the relevant biological mechanisms are active, which can have therapeutic implications.

Conclusions

The genetic associations here implicate new biological mechanisms underlying the development of life-threatening COVID-19, several of which may be amenable to therapeutic targeting. In the context of the ongoing global pandemic, translation to clinical practice is an urgent priority. As with our previous work, large-scale randomised trials are essential before translating our findings into clinical practice.

Acknowledgements

We thank the patients and their loved ones who volunteered to contribute to this study at one of the most difficult times in their lives, and the research staff in every intensive care unit who recruited patients at personal risk during the most extreme conditions we have ever witnessed in UK hospitals.

GenOMICC was funded by the Department of Health and Social Care (DHSC), LifeArc, the Medical Research Council, UKRI, Sepsis Research (the Fiona Elizabeth Agnew Trust), the Intensive Care Society, a Wellcome-Beit Prize award to J. K. Baillie (Wellcome Trust 103258/Z/13/A) and a BBSRC Institute Program Support Grant to the Roslin Institute (BBS/E/D/20002172, BBS/E/D/10002070 and BBS/E/D/30002275). Whole-genome sequencing was performed by Illumina at the NHS Genomic Sequencing Centre in partnership and was overseen by Genomics England. We would like to thank all at Genomics England who have contributed to the supporting the processing of the sequencing and clinical data. We thank DHSC, the Medical Research Council, UKRI, LifeArc, Genomics England Ltd and Illumina Inc for funding sequencing. Genomics England and the 100,000 Genomes Project was funded by the National Institute for Health Research, the Wellcome Trust, the Medical Research Council, Cancer Research UK, the Department of Health and Social Care and NHS England. We are grateful for the support from Professor Dame Sue Hill and the team in NHS England and the 13 Genomic Medicine Centres that successfully delivered the 100,000 Genomes Project which provide the control sequences for this study. We thank the participants of the 100,000 Genomes Project who made this study possible and the Genomics England Participant Panel for their strategic advice, involvement and engagement. We acknowledge NHS Digital, Public Health England and the Intensive Care National Audit and Research Centre who provided life course longitudinal clinical data on the participants. This work forms part of the portfolio of research of the NIHR Biomedical Research Centre at Barts. Mark Caulfield is an NIHR Senior Investigator. This study owes a great deal to the National Institute of Healthcare Research Clinical Research Network (NIHR CRN) and the Chief Scientist Office (Scotland), who facilitate recruitment into research studies in NHS hospitals, and to the global ISARIC and InFACT consortia.

The views expressed are those of the authors and not necessarily those of the DHSC, DID, NIHR, MRC, Wellcome Trust or PHE.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on August 22nd, 2021 (GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2).

Data availability

Summary statistics will be shared openly with international collaborators to accelerate discovery. Data can be obtained from genomicc.org/data

Individual-level data will be available in the UK Outbreak Analysis Platform at the University of Edinburgh and through the Genomics England research environment.

Contributions

AK, EP-C, KR, AS, CAO, SW, TM, KSE, BW, DR, LK, MZ, NP, ADB, JY, YW, SC, LMo, AL and JKB contributed to data analysis. AK, EP-C, KR, AS, CAO, SW, CDR, JM, AR, SC, LMo and AL contributed to bioinformatics. AK, EP-C, KR, CDR, JM, DM, AN, MGS, SC, LMo, MJC and JKB contributed to writing and reviewing the manuscript. EP-C, KR, KM, SK, AF, LM, KRo, CPP, VV, JFW, SC, AL, MJC and JKB contributed to design. SW, FG, WO, PG and SD contributed to project management. FG, WO, KM, SK, PG, SD, DM, AN, MGS, SS, JK, TAF, MS-H, CS, PE, AF, LM, KRo, CPP, RHS, SC and AL contributed to oversight. FG, WO, FM-C and JKB contributed to ethics and governance. KM, ASi, AF and LM contributed to sample handling and sequencing. and ASi contributed to data collection. and TZ contributed to sample handing. TZ and GE contributed to sequencing. and LT contributed to recruitment of controls. GC, PA and KRo contributed to clinical data management. KRo, CPP, SC and JKB contributed to conception. KRo, CPP, VV and JFW contributed to reviewing the manuscript. MJC and JKB contributed to scientific leadership.

Conflict of interest

All authors declare that they have no conflicts of interest relating to this work.

Genomics England Ltd is a wholly owned Department of Health and Social Care company created in 2013 to work with the NHS to introduce advanced genomic technologies and analytics into healthcare. All Genomics England affiliated authors are, or were, salaried by Genomics England during this programme.

Materials and Methods

Ethics

GenOMICC was both approved by the following research ethics committees: Scotland "A" Research Ethics Committee, 15/SS/0110; Coventry and Warwickshire Research Ethics Committtee (England, Wales and Northern Ireland), 19/WM/0247). Current and previous versions of the study protocol are available at genomicc.org/protocol. All participants gave informed consent.

Recruitment of cases

Patients recruited to the GenOMICC study (genomicc.org) had confirmed COVID-19 according to local clinical testing and were deemed, in the view of the treating clinician, to require continuous cardiorespiratory monitoring. In UK practice this kind of monitoring is undertaken in high-dependency or intensive care units. This study was approved by research ethics committees in the recruiting countries (Scotland 15/SS/0110, England, Wales and Northern Ireland: 19/WM/0247). Current and previous versions of the study protocol are available at genomicc.org/protocol. All participants gave informed consent.

Recruitment of controls

Mild/asymptomatic controls

Participants were recruited to the mild COVID-19 cohort on the basis of having experienced mild (non-hospitalised) or asymptomatic COVID-19. Participants volunteered to take part in the study via a microsite and were required to self-report the details of a positive COVID-19 test. Volunteers were prioritised for genome sequencing based on demographic matching with the critical COVID-19 cohort considering self-reported ancestry, sex, age and location within the UK. We refer to this cohort as the covid-mild cohort.

100,000 Genomes project controls

Participants were enrolled in the 100,000 Genomes Project from families with a broad range of rare diseases, cancers and infection by 13 regional NHS Genomic Medicine Centres across England and in Northern Ireland, Scotland and Wales. For this analysis, participants for whom a positive SARS-CoV-2 test had been recorded as of March, 2021 were not included due to uncertainty in the severity of COVID-19 symptoms. Only participants for whom genome sequencing was performed from blood derived DNA were included and participants with haematological malignancies were excluded to avoid potential tumour contamination.

DNA extraction

For critical COVID-19 cases and mild cohort controls, DNA was extracted from whole blood using Nucleon Kit (Cytiva) with the BACC3 protocol. DNA samples were re-suspended in 1 ml TE buffer pH 7.5 (10mM Tris-Cl pH 7.5, 1mM EDTA pH 8.0). The yield of the DNA was measured using Qubit and normalised to 50ng/μl before WGS or genotyping.

WGS sequencing

For all three cohorts, DNA was extracted from whole-blood using standard protocols. Sequencing libraries were generating using the Illumina TruSeq DNA PCR-Free High Throughput Sample Preparation kit and sequenced with 150bp paired-end reads in a single lane of an Illumina Hiseq X instrument (for 100,000 Genomes Project samples) or NovaSeq instrument (for the COVID-19 critical and mild cohorts).

Sequencing data QC

All genome sequencing data were required to meet minimum quality metrics and quality control measures were applied for all genomes as part of the bioinformatics pipeline. The minimum data requirements for all genomes were $> 85 \times 10^{-9}$ bases with $Q \geq 30$ and $\geq 95\%$ of the autosomal genome covered at $\geq 15x$ calculated from reads with mapping quality > 10 after removing duplicate reads and overlapping bases, after adaptor and quality trimming. Assessment of germline cross-sample contamination was performed using VerifyBamID and samples with $> 3\%$ contamination were excluded. Sex checks were performed to confirm that the sex reported for a participant was concordant with the sex inferred from the genomic data.

WGS Alignment and variant calling

COVID-19 cohorts

For the critical and mild COVID-19 cohorts, sequencing data alignment and variant calling was performed with Genomics England pipeline 2.0 which uses the DRAGEN software (v3.2.22). Alignment was performed to genome reference GRCh38 including decoy contigs and alternate haplotypes (ALT contigs), with ALT-aware mapping and variant calling to improve specificity.

100,000 Genome Project cohort (100K-genomes)

All genomes from the 100,000 Genomes Project cohort were analysed with the Illumina North Star Version 4 Whole Genome Sequencing Workflow (NSV4, version 2.6.53.23); which is comprised of the iSAAC Aligner (version 03.16.02.19) and Starling Small Variant Caller (version 2.4.7). Samples were aligned to the Homo Sapiens NCBI GRCh38 assembly with decoys.

A subset of the genomes from the Cancer program of the 100,000 Genomes Project were reprocessed (alignment and variants calling) using the same pipeline used for the COVID-19 cohorts (DRAGEN v3.2.22) for equity of alignment and variant calling.

Aggregation

Aggregation was conducted separately for the samples analysed with Genomics England pipeline 2.0 (severe-cohort, mild-cohort, cancer-realigned-100K), and those analysed with the Illumina North Star Version 4 pipeline (100K-Genomes).

For the first three, the WGS data were aggregated from single sample gVCF files to multi-sample VCF files using GVCFGenotyper (GG) v3.8.1, which accepts gVCF files generated via the DRAGEN pipeline as input. GG outputs multi-allelic variants (several ALT variants per position on the same row), and for downstream analyses the output was decomposed to bi-allelic variants per row using software vt v0.57721. We refer to the aggregate as aggCOVID_vX, where X is the specific freeze. The analysis in this manuscript uses data from freeze v4.2 and the respective aggregate is referred to as aggCOVID_v4.2.

Aggregation for the 100K-Genomes cohort was performed using Illumina's gvcfgenotyper v2019.02.26, merged with bcftools v1.10.2 and normalised with vt v0.57721.

Sample Quality Control (QC)

Samples that failed any of the following four BAM-level QC filters: freemix contamination (>3%), mean autosomal coverage (<25X), percent mapped reads (<90%), and percent chimeric reads (>5%) were excluded from the analysis.

Additionally, a set of VCF-level QC filters were applied post-aggregation on all autosomal bi-allelic SNVs (akin to gnomAD v3.1[17]). Samples were filtered out based on the residuals of eleven QC metrics (calculated using bcftools) after regressing out the effects of sequencing platform and the first three ancestry assignment principal components (including all linear, quadratic, and interaction terms) taken from the sample projections onto the SNP loadings from the individuals of 1000 Genomes Project phase 3 (1KGP3). Samples were removed that were four median absolute deviations (MADs)

above or below the median for the following metrics: ratio heterozygous-homozygous, ratio insertions-deletions, ratio transitions-transversions, total deletions, total insertions, total heterozygous snps, total homozygous snps, total transitions, total transversions. For the number of total singletons (snps), samples were removed that were more than 8 MADs above the median. For the ratio of heterozygous to homozygous alternate snps, samples were removed that were more than 4 MADs above the median.

After quality control, 79,803 individuals were included in the analysis with the breakdown according to cohort shown in Supplementary Table 2.

Selection of high-quality (HQ) independent SNPs

We selected high-quality independent variants for inferring kinship coefficients, performing PCA, assigning ancestry and for the conditioning on the Genetic Relatedness matrix by the logistic mixed model of SAIGE and SAIGE-GENE. To avoid capturing platform and/or analysis pipeline effects for these analyses, we performed very stringent variant QC as described below.

HQ common SNPs

We started with autosomal, bi-allelic SNPs which had frequency > 5% in aggV2 (100K participant aggregate) and in the 1KGP3. We then restricted to variants that had missingness <1%, median genotype quality QC>30, median depth (DP) >=30 and >= 90% of heterozygote genotypes passing an ABratio binomial test with P-value > 10^{-2} for aggV2 participants. We also excluded variants in complex regions from the list available in , and variants where the ref/alt combination was CG or AT (C/G, G/C, A/T, T/A). We also removed all SNPs which were out of Hardy Weinberg Equilibrium (HWE) in any of the AFR, EAS, EUR or SAS super-populations of aggV2, with a P-value cutoff of pHWE < 10^{-5}. We then LD-pruned using plink v1.9 with an $r^2 = 0.1$ and in 500kb windows. This resulted in a total of 63,523 high-quality sites from aggV2.

We then extracted these high-quality sites from the aggCOVID_v4.2 aggregate and further applied variant quality filters (missingness <1%, median QC>30, median depth >=30 and >= 90% of heterozygote genotypes passing an ABratio binomial test with P-value > 10^{-2}), per batch of sequencing platform (i.e, HiseqX, NovaSeq6000).

After applying variant filters in aggV2 and aggCOVID_v4.2, we merged the genomic data from the two aggregates for the intersection of the variants which resulted in a final total of 58,925 sites.

HQ rare SNPs

We selected high-quality rare (MAF< 0.005) bi-allelic SNPs to be used with SAIGE for aggregate variant testing analysis. To create this set, we applied the same variant QC procedure as with the common variants: We selected variants that had missingness <1%, median QC>30, median depth >=30 and >= 90% of heterozygote genotypes passing an ABratio binomial test with P-value > 10^{-2} per batch of sequencing and genotyping platform (i.e, HiSeq+NSV4, HiSeq+Pipeline 2.0, NovaSeq+Pipeline 2.0). We then subsetted those to the following groups of MAC/MAF categories: MAC 1, 2, 3, 4, 5, 6-10, 11-20, MAC 20 - MAF 0.001, MAF 0.001 - 0.005.

Relatedness, ancestry and principal components

Kinship

We calculated kinship coefficients among all pairs of samples using software plink2 and its implementation of the KING robust algorithm. We used a kinship cutoff < 0.0442 to select unrelated individuals with argument "–king-cutoff".

Genetic Ancestry Prediction

To infer the ancestry of each individual we performed principal components analysis (PCA) on unrelated 1KGP3 individuals with GCTA v1.93.1_beta software using HQ common SNPs and inferred the first 20 PCs. We calculated loadings for each SNP which we used to project aggV2 and aggCOVID_v4.2 individuals onto the 1KGP3 PCs. We then trained a random forest algorithm from R-package randomForest with the first 10 1KGP3 PCs as features and the super-population ancestry of each individual as labels. These were 'AFR' for individuals of African ancestry, 'AMR' for individuals of American ancestry, 'EAS' for individuals of East Asian ancestry, 'EUR' for individuals of European ancestry, and 'SAS' for individuals of South Asian ancestry. We used 500 trees for the training. We then used the trained model to assign probability of belonging to a certain super-population class for each individual in our cohorts. We assigned individuals to a super-population when class probability >=0.8. Individuals for which no class had probability >=0.8 were labelled as "unassigned" and were not included in the analyses.

Principal component analysis

After labelling each individual with predicted genetic ancestry, we calculated ancestry-specific PCs using GCTA v1.93.1_beta, *i.e.*. We computed 20 PCs for each of the ancestries that were used in the association analyses (AFR, EAS, EUR, and SAS).

Variant Quality Control

Variant QC was performed to ensure high quality of variants and to minimise batch effects due to using samples from different sequencing platforms (NovaSeq6000 and HiseqX) and different variant callers (Strelka2 and DRAGEN). We first masked low-quality genotypes setting them to missing, merged aggregate files and then performed additional variant quality control separately for the two major types of association analyses, GWAS and AVT, which concerned common and rare variants, respectively.

Masking

Prior to any analysis we masked low quality genotypes using bcftools setGT module. Genotypes with DP<10, GQ<20, and heterozygote genotypes failing an AB-ratio binomial test with P-value $< 10^{-3}$ were set to missing.

We then converted the masked VCF files to plink and bgen format using plink v.2.0.

Merging of aggregate samples

Merging of aggV2 and aggCOVID_v4.2 samples was done using plink files with masked genotypes and the merge function of plink v.1.9.[46] for variants that were found in both aggregates.

GWAS analyses

Variant QC

We restricted all GWAS analyses to common variants applying the following filters using plink v1.9: MAF > 0 in both cases and controls, MAF$> 0.5\%$ and MAC >20, missingness $< 2\%$, Differential missingness between cases and controls, mid-P-value $< 10^{-5}$, HWE deviations on unrelated controls, mid-P-value $< 10^{-6}$, Multi-allelic variants were additionally required to have MAF $> 0.1\%$ in both aggV2 and aggCOVID_v4.2.

Control-control QC filter

100K aggV2 samples that were aligned and genotype called with the Illumina North Star Version 4 pipeline represented the majority of control samples in our GWAS analyses, whereas all of the cases were aligned and called with Genomics England pipeline 2.0 (Supplementary Table 1). Therefore, the alignment and genotyping pipelines partially match the case/control status which necessitates additional filtering for adjusting for between-pipeline differences in alignment and variant calling. To control for potential batch effects, we used the overlap of 3,954 samples from the Genomics England 100K participants that were aligned and called with both pipelines. For each variant, we computed and compared between platforms the inferred allele frequency for the population samples. We then filtered out all variants that had $> 1\%$ relative difference in allele frequency between platforms. The relative difference was computed on a per-population basis for EUR (n=3,157), SAS (n=373), AFR (n=354) and EAS (n=81).

Model

We used a 2-step logistic mixed model regression approach as implemented in SAIGE v0.44.5 for single variant association analyses. In step 1, SAIGE fits the null mixed model and covariates. In step 2, single variant association tests are performed with the saddlepoint approximation (SPA) correction to calibrate unbalanced case-control ratios. We used the HQ common variant sites for fitting the null model and sex, age, age^2, $age * sex$ and 20 principal components as covariates in step 1. The principal components were computed separately by predicted genetic ancestry (i.e, EUR-specific, AFR-specific, etc.), to capture subtle structure effects.

Analyses

All analyses were done on unrelated individuals with pairwise kinship coefficient < 0.0442. We conducted GWAS analyses per genetic ancestry, for all populations for which we had >100 cases and >100 controls (AFR, EAS, EUR, and SAS).

Multiple testing correction

As our study is testing variants that were directly sequenced by WGS and not imputed, we calculated the P-value significance threshold by estimating the effective number of tests. After selecting the

final filtered set of tested variants for each population, we LD-pruned in a window of 250Kb and $r^2 = 0.8$ with plink 1.9. We then computed the Bonferroni-corrected P-value threshold as 0.05 divided by the number of LD-pruned variants. The P-value thresholds that were used for declaring statistical significance are given in Supplementary Table 3.

LD-clumping

We used plink1.9 to do clumping of variants that were genome-wide significant for each analysis with $P1$ set to per-population P-value from table X, $P2 = 0.01$, clump distance 1500Mb and $r^2 = 0.1$.

Conditional analysis

To find the set of independent variants in the per-population analyses, we performed a step-wise conditional analysis with the GWAS summary statistics for each population using GTCA 1.9.3 –cojo-slct function. The parameters for the function were $pval = 2.2 \times 10^{-8}$, a distance of 10,000 kb and a colinear threshold of 0.9[47].

Fine-mapping

We performed fine-mapping for genome-wide significant signals using Rpackage SusieR v0.11.42[48]. For each genome-wide significant variant locus, we selected the variants 1.5 Mbp on each side and computed the correlation matrix among them with plink v1.9. We then run the susieR summary-statistics based function susie_rss and provided the summary z-scores from SAIGE (i.e, effect size divided by its standard error) and the correlation matrix computed with the same samples that were used for the corresponding GWAS. We required coverage >0.95 for each identified credible set and minimum and median correlation coefficients (purity) of r=0.1 and 0.5, respectively.

Functional annotation of credible sets

We annotated all variants included in each credible set identified by SusieR using VEP v99. We also selected the worst consequence across transcripts using bcftools +split-vep -s worst. We also ranked each variant within each credible set according to the predicted consequence and the ranking was based on the table provided by Ensembl: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html.

Trans-ancestry meta-analysis

We performed a meta-analysis across all ancestries using a inverse-variance weighted method and control for population stratification for each separate analysis in the METAL software[13]. The meta-analysed variants were filtered for variants with heterogeneity P-value $p < 2.22 \times 10^{-8}$ and variants that are not present in at least half of the individuals. We used the meta R package to plot forest plots of the clumped trans-ancestry meta-analysis variants[49].

LD-based validation of lead GWAS signals

In order to quantify the support for genome-wide significant signals from nearby variants in LD, we assessed the internal consistency of GWAS results of the lead variants and their surroundings. To this end, we compared observed z-scores at lead variants with the expected z-scores based on those

observed at neighbouring variants. Specifically, we computed the observed z-score for a variant i as $s_i = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$ and, following the approach of[50], the imputed z-score at a target variant t as

$$\hat{s}_t = \mathbf{\Sigma}_{t,P}(\mathbf{\Sigma}_{P,P} + \lambda \mathbf{I})^{-1}\mathbf{s}_P$$

where $\mathbf{s}_P$ are the observed z-scores at a set P of predictor variants, $\mathbf{\Sigma}_{x,y}$ is the empirical correlation matrix of dosage coded genotypes computed on the GWAS sample between the variants in x and y, and λ is a regularization parameter set to 10^{-5}. The set P of predictor variants consisted of all variants within 100 kb of the target variant with a genotype correlation with the target variant greater than 0.25.

Replication

We used the Host Genetic Initiative (HGI) GWAS meta-analysis round 6 hospitalised COVID vs population (B2 analysis), including all genetic ancestries. In order to remove overlapping signals we performed a mathematical subtraction of the GenOMICC GWAS of European genetic ancestry. The HGI data was downloaded from https://www.covid19hg.org/results/r6/. The subtraction was performed using MetaSubtract package (version 1.60) for R (version 4.0.2) after removing variants with the same genomic position and using the lambda.cohortswith genomic inflation calculated on the GenOMICC summary statistics. Then, we calculated a trans-ancestry meta-analysis for the three ancestries with summary statistics in 23andMe: African, Latino and European using variants that passed the 23andMe ancestry QC, with imputation score > 0.6 and with maf > 0.005. And finally we performed a final meta-analysis of 23andMe and HGI B2 without GenOMICC to create the final replication set. Meta-analysis were performed using METAL[13], with the inverse-variance weighting method (STDERR mode) and genomic control ON. We considered that a hit was replicating if the direction of effect in the GenOMICC-subtracted HGI summary statistics was the same as in our GWAS, and the P-value was significant after Bonferroni correction for the number of attempted replications ($pval < 0.05/25$). If the main hit was not present in the HGI-23andMe meta-analysis or if the hit was not replicating we looked for replication in variants in high LD with the top variant ($r^2 > 0.9$), which helped replicate two regions.

Stratified analysis

We also performed sex-specific analysis (male and females separately) as well as analysis stratified by age (*i.e.*, participants <60 and >=60 years old) for each super-population set. To compare effect of variants within groups for the age and sex stratified analysis we first adjusted the effect and error of each variant for the standard deviation of the trait in each stratified group and then used the following t-statistic, as in previous studies[51;52]

$$t = \frac{b_1 - b_2}{\sqrt{se_1^2 + se_2^2 - 2 \cdot r se_1 \cdot r se_2}}$$

where b_1 is the adjusted effect for group 1, b_2 is the adjusted effect for group 2, se_1 and se_2 are the adjusted standard errors for group 1 and 2 respectively and r is the Spearman rank correlation between groups across all genetic variants.

HLA Imputation and Association Analysis

HLA types were imputed at two field (4-digit) resolution for all samples within aggV2 and aggCOVID_v4.2 for the following seven loci: HLA-A, HLA-C, HLA-B, HLA-DRB1, HLA-DQA1,

HLA-DQB1, and HLA-DPB1 using the HIBAG package in R [19]. At time of writing, HLA types were also imputed for 82% of samples using HLA*LA [53]. Inferred HLA alleles between HIBAG and HLA*LA were >96% identical at 4-digit resolution. HLA association analysis was run under an additive model using SAIGE; in an identical fashion to the SNV GWAS. The multi-sample VCF of aggregated HLA type calls from HIBAG were used as input where any allele call with posterior probability (T) < 0.5 were set to missing.

Aggregate variant testing (AVT)

Aggregate variant testing on aggCOVID_v4.2 was performed using SKAT-O as implemented in SAIGE-GENE v0.44.5 [16] on all protein-coding genes. Variant and sample QC for the preparation and masking of the aggregate files has been described elsewhere. We further excluded SNPs with differential missingness between cases and controls (mid-P value $< 10^{-5}$) or a site-wide missingness above 5%. Only bi-allelic SNPs with a MAF<0.5% were included.

We filtered the variants to include in the aggregate variant testing by applying two functional annotation filters: A putative loss of function (*pLoF*) filter, where only variants that are annotated by LOFTEE [17] as high confidence loss of function were included, and a more lenient (*missense*) filter where variants that have a consequence of missense or worse as annotated by VEP, with a CADD_PHRED score of ≥ 10, were also included. All variants were annotated using VEP v99. SAIGE-GENE was run with the same covariates used in the single variant analysis: sex, age, age^2, $age * sex$ and 20 (population-specific) principal components generated from common variants (MAF $\geq 5\%$).

We ran the tests separately by genetically predicted ancestry, as well as across all four ancestries as a mega-analysis. We considered a gene-wide significant threshold on the basis of the genes tested per ancestry, correcting for the two masks (*pLoF* and *missense*, Supplementary Table 4).

Post-GWAS analysis

Transcriptome-wide Association Studies (TWAS)

We performed TWAS in the MetaXcan framework and the GTExv8 eQTL and sQTL MASHR-M models available for download in (http://predictdb.org/). We first calculated, using the European summary statistics, individual TWAS for whole blood and lung with the S-PrediXcan function [54;55]. Then we performed a metaTWAS including data from all tissues to increase statistical power using s-MultiXcan [56]. We applied Bonferroni correction to the results in order to choose significant genes and introns for each analysis.

Colocalisation analysis

Significant genes from TWAS, splicing TWAS, metaTWAS and splicing metaTWAS, as well as genes where one of the top variants was a significant eQTL or sQTL were selected for a colocalisation analysis using the coloc R package [57]. We chose the lead SNPS from the European ancestry GWAS summary statistics and a region of ±200 kb around each SNP to do the colocalisation with the identified genes in the region. GTExv8 whole blood and lung tissue summary statistics and eqtlGen (which has blood eQTL summary statistics for $> 30,000$ individuals) were used for the analysis [18;58]. We first performed a sensitivity analysis of the posterior probability of colocalisation (PPH4) on the

prior probability of colocalisation (p12), going from $p12 = 10^{-}8$ to $p12 = 10^{-}4$ with the default threshold being $p12 = 10^{-}5$. eQTL signal and GWAS signals were deemed to colocalise if these two criteria were met: (1) At $P12 = 5 \times 10^{-}5$ the probability of colocalisation $PPH4 > 0.5$ and (2) At $p12 = 10^{-}5$ the probability of independent signal (PPH3) was not the main hypothesis ($PPH3 < 0.5$). These criteria were chosen to allow eQTLs with weaker P-values due to lack of power in GTExv8, to be colocalised with the signal when the main hypothesis using small priors was that there wasn't any signal in the eQTL data.

As the chromosome 3 associated interval is larger than 200 kb, we performed additional colocalisation including a region up to 500 kb, but no further colocalisations were found.

References

[1] Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in Covid-19. *Nature* 1–1 (2020).

[2] Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science (New York, N.y.)* **370**, eabd4570 (2020). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7857407/.

[3] Ellinghaus, D. *et al.* Genomewide association study of severe covid-19 with respiratory failure. *The New England journal of medicine* **383**, 1522–1534 (2020).

[4] COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* (2021). URL https://doi.org/10.1038/s41586-021-03767-x.

[5] Docherty, A. B. *et al.* Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Prospective observational cohort study. *BMJ* **369** (2020).

[6] Dorward, D. A. *et al.* Tissue-Specific Immunopathology in Fatal COVID-19. *American Journal of Respiratory and Critical Care Medicine* **203**, 192–201 (2021).

[7] Millar, J. E. *et al.* Robust, reproducible clinical patterns in hospitalised patients with COVID-19. *medRxiv* 2020.08.14.20168088 (2020).

[8] Horby, P. *et al.* Dexamethasone in Hospitalized Patients with Covid-19 — Preliminary Report. *New England Journal of Medicine* (2020).

[9] Degenhardt, F. *et al.* New susceptibility loci for severe COVID-19 by detailed GWAS analysis in European populations (2021).

[10] Kosmicki, J. A. *et al.* Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. *American Journal of Human Genetics* **108**, 1350–1355 (2021). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8173480/.

[11] Povysil, G. *et al.* Rare loss-of-function variants in type i ifn immunity genes are not associated with severe covid-19. *The Journal of clinical investigation* **131** (2021).

[12] Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335–1341 (2018). URL http://www.nature.com/articles/s41588-018-0184-y.

[13] Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* **26**, 2190–2191 (2010).

[14] Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273–1300 (2020). URL https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssb.12388https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12388https://rss.onlinelibrary.wiley.com/doi/10.1111/rssb.12388.

[15] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**, D886–D894 (2018). URL https://doi.org/10.1093/nar/gky1016.

[16] Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature Genetics* **52**, 634–639 (2020). URL https://www.nature.com/articles/s41588-020-0621-6.

[17] Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). URL https://www.nature.com/articles/s41586-020-2308-7.

[18] Consortium, T. G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). URL https://science.sciencemag.org/content/369/6509/1318. Publisher: American Association for the Advancement of Science _eprint: https://science.sciencemag.org/content/369/6509/1318.full.pdf.

[19] Zheng, X. *et al.* HIBAG - HLA genotype imputation with attribute bagging. *Pharmacogenomics Journal* **14**, 192–200 (2014).

[20] Dunning, J. W. *et al.* Open source clinical science for emerging infections. *The Lancet Infectious Diseases* **14**, 8–9 (2014).

[21] Repurposed Antiviral Drugs for Covid-19 — Interim WHO Solidarity Trial Results. *New England Journal of Medicine* **0**, null (2020).

[22] Dong, B. *et al.* Phospholipid scramblase 1 potentiates the antiviral activity of interferon. *Journal of virology* **78**, 8983–93 (2004).

[23] Luo, W. *et al.* Phospholipid scramblase 1 interacts with influenza a virus np, impairing its nuclear import and thereby suppressing virus replication. *PLoS pathogens* **14**, e1006851 (2018).

[24] Chen, M.-H. *et al.* Phospholipid Scramblase 1 Contains a Nonclassical Nuclear Localization Signal with Unique Binding Site in Importin A^*. *Journal of Biological Chemistry* **280**, 10599–10606 (2005).

[25] Chen, C.-W., Sowden, M., Zhao, Q., Wiedmer, T. & Sims, P. J. Nuclear phospholipid scramblase 1 prolongs the mitotic expansion of granulocyte precursors during G-CSF-induced granulopoiesis. *Journal of Leukocyte Biology* **90**, 221–233 (2011).

[26] Bevers, E. M. & Williamson, P. L. Phospholipid scramblase: An update. *FEBS Letters* **584**, 2724–2730 (2010).

[27] Yu, Y. *et al.* Bcl11a is essential for lymphoid development and negatively regulates p53. *The Journal of experimental medicine* **209**, 2467–83 (2012).

[28] Reizis, B. Plasmacytoid Dendritic Cells: Development, Regulation, and Function. *Immunity* **50**, 37–50 (2019).

[29] Zhang, Y., Lu, L., Furlonger, C., Wu, G. E. & Paige, C. J. Hemokinin is a hematopoietic-specific tachykinin that regulates b lymphopoiesis. *Nature immunology* **1**, 392–7 (2000).

[30] Wang, W. *et al.* Hemokinin-1 activates the mapk pathway and enhances b cell proliferation and antibody production. *Journal of immunology (Baltimore, Md. : 1950)* **184**, 3590–7 (2010).

[31] Janelsins, B. M. *et al.* Proinflammatory tachykinins that signal through the neurokinin 1 receptor promote survival of dendritic cells and potent cellular immunity. *Blood* **113**, 3017–26 (2009).

[32] Thwaites, R. S. *et al.* Inflammatory profiles across the spectrum of disease reveal a distinct role for GM-CSF in severe COVID-19. *Science Immunology* **6** (2021).

[33] Lang, F. M., Lee, K. M.-C., Teijaro, J. R., Becher, B. & Hamilton, J. A. Gm-csf-based treatments in covid-19: reconciling opposing therapeutic approaches. *Nature reviews. Immunology* **20**, 507–514 (2020).

[34] Moore, C. *et al.* Resequencing Study Confirms That Host Defense and Cell Senescence Gene Variants Contribute to the Risk of Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **200**, 199–208 (2019). URL https://www.atsjournals.org/doi/10.1164/rccm.201810-1891OC. Publisher: American Thoracic Society - AJRCCM.

[35] Takatsu, H. *et al.* Phospholipid flippase activities and substrate specificities of human type iv p-type atpases localized to the plasma membrane. *The Journal of biological chemistry* **289**, 33543–56 (2014).

[36] Bevers, E. M., Comfurius, P. & Zwaal, R. F. Changes in membrane phospholipid distribution during platelet activation. *Biochimica et biophysica acta* **736**, 57–66 (1983).

[37] Zwaal, R. F., Comfurius, P. & van Deenen, L. L. Membrane asymmetry and blood coagulation. *Nature* **268**, 358–60 (1977).

[38] Shrine, N. *et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nature genetics* **51**, 481–493 (2019).

[39] Mankelow, T. J. *et al.* Blood group type A secretors are associated with a higher risk of COVID-19 cardiovascular disease complications. *eJHaem* **2**, 175–187 (2021).

[40] Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G. & Lowe, J. B. Sequence and expression of a candidate for the human secretor blood group alpha(1,2)fucosyltransferase gene (fut2). homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *The Journal of biological chemistry* **270**, 4640–9 (1995).

[41] Ferrer-Admetlla, A. *et al.* A natural history of fut2 polymorphism in humans. *Molecular biology and evolution* **26**, 1993–2003 (2009).

[42] Imbert-Marcille, B.-M. *et al.* A fut2 gene common polymorphism determines resistance to rotavirus a of the p[8] genotype. *The Journal of infectious diseases* **209**, 1227–30 (2014).

[43] Tian, C. *et al.* Genome-wide association and hla region fine-mapping studies identify susceptibility loci for multiple common infections. *Nature communications* **8**, 599 (2017).

[44] Kachuri, L. *et al.* The landscape of host genetic factors involved in immune response to common viral infections. *medRxiv : the preprint server for health sciences* (2020).

[45] Blackwell, C. C. *et al.* Non-secretion of abo antigens predisposing to infection by neisseria meningitidis and streptococcus pneumoniae. *Lancet (London, England)* **2**, 284–5 (1986).

[46] Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007). URL https://www.sciencedirect.com/science/article/pii/S0002929707613524.

[47] Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369–375 (2012). URL https://doi.org/10.1038/ng.2213.

[48] Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **82**, 1273–1300 (2020). URL https://rss.onlinelibrary.wiley.com/doi/10.1111/rssb.12388.

[49] Balduzzi, S., Rücker, G. & Schwarzer, G. How to perform a meta-analysis with R: a practical tutorial. *Evidence-Based Mental Health* **22**, 153–160 (2019). URL https://ebmh.bmj.com/content/22/4/153. Publisher: Royal College of Psychiatrists Section: Statistics in practice.

[50] Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014). URL https://doi.org/10.1093/bioinformatics/btu416. https://academic.oup.com/bioinformatics/article-pdf/30/20/2906/17147061/btu416.pdf.

[51] Winkler, T. W. *et al.* The influence of age and sex on genetic associations with adult body size and shape: A large-scale genome-wide interaction study. *PLOS Genetics* **11**, 1–42 (2015). URL https://doi.org/10.1371/journal.pgen.1005378.

[52] Bernabeu, E. *et al.* Sexual differences in genetic architecture in uk biobank. *bioRxiv* (2020). URL https://www.biorxiv.org/content/early/2020/07/21/2020.07.20.211813. https://www.biorxiv.org/content/early/2020/07/21/2020.07.20.211813.full.pdf.

[53] Dilthey, A. T. *et al.* HLA*LA—HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–4396 (2019). URL https://doi.org/10.1093/bioinformatics/btz235. https://academic.oup.com/bioinformatics/article-pdf/35/21/4394/30330845/btz235.pdf.

[54] Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**, 1825 (2018). URL https://doi.org/10.1038/s41467-018-03621-1.

[55] Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**, 1091–1098 (2015). URL https://doi.org/10.1038/ng.3367.

[56] Barbeira, A. N. *et al.* Integrating predicted transcriptome from multiple tissues improves association detection. *PLOS Genetics* **15**, 1–20 (2019). URL https://doi.org/10.1371/journal.pgen.1007889. Publisher: Public Library of Science.

[57] Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004383. Publisher: Public Library of Science.

[58] Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018). URL http://biorxiv.org/content/early/2018/10/19/447367.abstract.