**Supplement 2:** Genome-wide association studies for the susceptibility and the severity of SARS-CoV-2 infections in UK Biobank

The UK Biobank (UKB) analysis was based on a standardized phenotype definitions and analysis plan described in COVID-19 HGI1.

***Phenotype definitions***

UKB is a population-based cohort which recruited people aged between 40 – 69 years from across the UK. Covid-19 tests results were provided to UKB by Public Health England for tests dated from March 16th, 2020 to July 26th, 2020. We included all white-British individuals (N = 440,346, more detailed methods were described in the **Determining the white-British subset** section below) and subsequently excluded those with no sex chromosome imputation results (N = 291), those who withdrew consent recently (N = 86), and those who died before March 16th, 2020, derived by Data-field 40000. This process retained 413,809 individuals. We used “covid19\_result” table for the PCR test results. The UKB describes Covid-19 testing procedures as follows: “The vast majority of samples tested for Covid-19 disease are from combined nose/throat swabs, that are transported in a medium suitable for viruses (a balanced salt solution), for PCR to be performed. In intensive care settings, lower respiratory samples may also be analysed”2. UK Biobank released death register data (“death” and “death\_cause” tables) up to June 28th, but full follow-up is only guaranteed until May 31st. Inpatient hospital data (“hesin” tables; admitted patient care in England) is available up to May 31st.

For the susceptibility phenotype, case status was defined as individuals with at least one positive test results or using primary care electronic health record(N = 1,333) and controls were defined as those with only negative test results (N = 10,174). Of note, among individuals with hospital encounter data, there were none with ICD-10 codes of “U071” or “U072” and who did not have a PCR test performed to confirm their Covid-19 status. To extract Covid-19 cases from TPP primary care data (“covid19\_tpp\_gp\_clinical” table; only available for England until June 3rd, 2020), we used the following codes2: “Y22b7” (Patient reported having already had Covid-19 in the recent past [not tested]) and “Y22b8” (Patient reported having already had Covid-19 in the recent past [tested].

For the hospitalized phenotype, cases were defined as hospitalized Covid-19 whereas controls were non-hospitalized Covid-19 patients. Case status was defined as SARS-CoV-2 PCR positive individuals with either a “hospital inpatient” flag (reqorg = 1 in UKB) when they were tested, patients who died after the positive test (based on death registry), or patient with at least one inpatient hospital episode ending after the positive test (N = 723). Controls were defined as PCR positive individuals who did not meet the criteria of cases (N = 607).

For the severe disease phenotype, both cases and controls had a positive SARS-CoV-2 PCR result. Cases were further defined as individuals who died after their test (based on death registry), or with a hospital admission (ending after the positive test) associated with ICD-10 codes J80, J9600, J9609, Z991 or OPCS4 codes E851 or E852 (N = 293). Controls were further defined as being alive after their positive test, and without a hospital admission ending after the positive test (N = 607).

For all three extended phenotypes, case definitions were the same as the corresponding analyses above. However, controls were defined as all individuals with only negative tests or those without test results in England. As the COVID-19 test results were only available from Public Health England, we excluded individuals living in Wales and Scotland on UKB enrolment. For this, we used the assessment centre (Data-Field: 54) of their initial visit and removed those assessed at Cardiff (Data-Coding [DC]: 11003), Edinburg (DC: 11004), Glasgow (DC: 11004), Swansea (DC: 11022), and Wrexham (DC: 11023). The final control count for the extended phenotypes was 364,359. For extended susceptibility phenotypes, there were 364,356 controls, as three individuals were defined to be cases using primary care electronic health record.

***Ethical compliance***

Ethical approval was granted from the Northwest Multi-Centre Research Ethics Committee for the UKB and informed consent was obtained from all participants prior to participation.

***Determining the white-British subset***

To minimize bias due to population stratification, analyses were restricted to individuals of white-British ancestry, comprising the largest ancestral sub-population within the UK Biobank. We previously defined3 the white-British subset by first selecting high-quality LD-pruned SNPs from HapMap3 (minor allele frequency [MAF] > 0.01, minor allele count [MAC] > 5, Hardy-Weinberg equilibrium p > 1 x 10-6). Using these SNPs we then computed principal components within 1000 Genomes4 using FlashPCA5, which we then used to project UK Biobank individuals. Clustering analysis of these projected principal components was used to identify the largest cluster that overlapped the GBR sub-population of 1000 Genomes, which we henceforth refer to as the white-British subset.

***Genome-wide association analysis in UKB***

We tested autosomal and X-chromosomal genetic variants for association with COVID-19 severity, as defined above, assuming an additive allelic effect, using an mixed linear model(MLM)-based association analysis implemented in the SAIGE6 to account for population structure and cryptic relatedness. We used all variants with imputation score (INFO) > 0.6, MAF > 0.0001 and MAC > 20 as was done in the recent GWAS for COVID-19 severity7. The following covariates were included as fixed effects in the models: age, age2, sex, age\*sex, and ancestry informative principal components (PCs) 1 to 10. Age was calculated using the age at the recruitment (Data-field: 21003), the date of attending assessment centre (Data-field: 53) and the testing date for SARS-Cov2 infection. For those without test results, the date was set to the last date of the available tests (July 29th, 2020). To calculate PCs, variants with MAF < 0.05, call-rate < 95%, Hardy-Weinberg Equilibrium p value < 1 x 10-6, and variants within high LD region were first removed, followed by pruning with plink --indep-pairwise 50 5 0.05. PCs were calculated using plink with all white-British individuals defined above.

***GWAS meta-analysis***

We restricted to the cohorts of European ancestry included in COVID-19 HGI to reduce the risk of bias from population stratification. As the GWAS summary released in COVID-19 HGI was built on GRCh38, we first converted the chromosomal positions in UKB GWAS from hg19 to GRCh38 using liftOver8 and then meta-analyzed using METAL9. All positions which failed to be converted were discarded from the analysis. We included the variants with INFO > 0.6 and MAF > 0.0001 in each cohort and performed meta-analysis using METAL assuming fixed effects as was done in COVID-19 HGI.

**References**

1. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European journal of human genetics : EJHG* 2020; **28**(6): 715-8.

2. UK Biobank Covid-19 Test Results Data; <http://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/tppgp4covid19.pdf>.

3. Morris JA, Kemp JP, Youlten SE, et al. An atlas of genetic influences on osteoporosis in humans and mice. *Nat Genet* 2019; **51**(2): 258-66.

4. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature* 2015; **526**(7571): 68-74.

5. Abraham G, Inouye M. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLOS ONE* 2014; **9**(4): e93766.

6. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018; **50**(9): 1335-41.

7. Ellinghaus D, Degenhardt F, Bujanda L, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *New England Journal of Medicine* 2020.

8. Hinrichs AS, Karolchik D, Baertsch R, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 2006; **34**(Database issue): D590-D8.

9. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* 2010; **26**(17): 2190-1.