Supplementary information

Initial Study of Human Genetic Contribution to COVID-19 Severity and Susceptibility

Fang Wang[1]*, Shujia Huang[2,3]*, Rongsui Gao[1]*, Yuwen Zhou[2,4]*, Changxiang Lai[1]*, Zhichao Li[2,4]*, Wenjie Xian[1], Xiaobo Qian[2,4], Zhiyu Li[1], Yushan Huang[2,4]，Qiyuan Tang[1], Panhong Liu[2,4], Ruikun Chen[1], Rong Liu[2], Xuan Li[1], Xin Tong[2], Xuan Zhou[1], Yong Bai[2], Gang Duan[1], Tao Zhang[2], Xun Xu[2,5], Jian Wang[2,6], Huanming Yang[2,6], Siyang Liu[2#], Qing He[1#], Xin Jin[2,3#], Lei Liu[1#]

1. The Third People's Hospital of Shenzhen, National Clinical Research Center for Infectious Disease, The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen 518112, Guangdong, China
2. BGI-Shenzhen, Shenzhen 518083, Guangdong, China
3. School of Medicine, South China University of Technology, Guangzhou 510006, Guangdong, China
4. BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, Guangdong, China
5. Guangdong Provincial Key Laboratory of Genome Read and Write，BGI-Shenzhen, Shenzhen, 518120，China
6. James D. Watson Institute of Genome Science, 310008 Hangzhou, China

*Those authors contribute equally

Corresponding to any of the followings:

Lei Liu liuleiszsdsrmyy@163.com

Xin Jin jinxin@genomics.cn

Qing He heqingjoe@163.com

Siyang Liu liusiyang@genomics.cn

Supplementary Figures

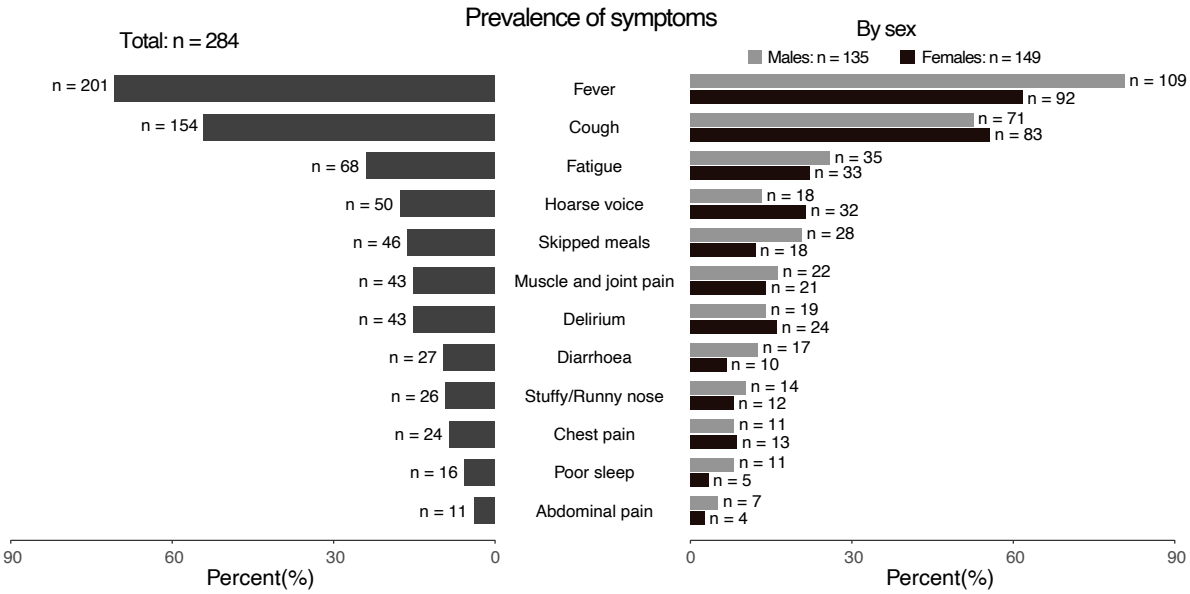


Figure S1. Clinical manifestation of the 284 unrelated individuals. Statistics were calculated from the patients' complaints in the electronic health records.

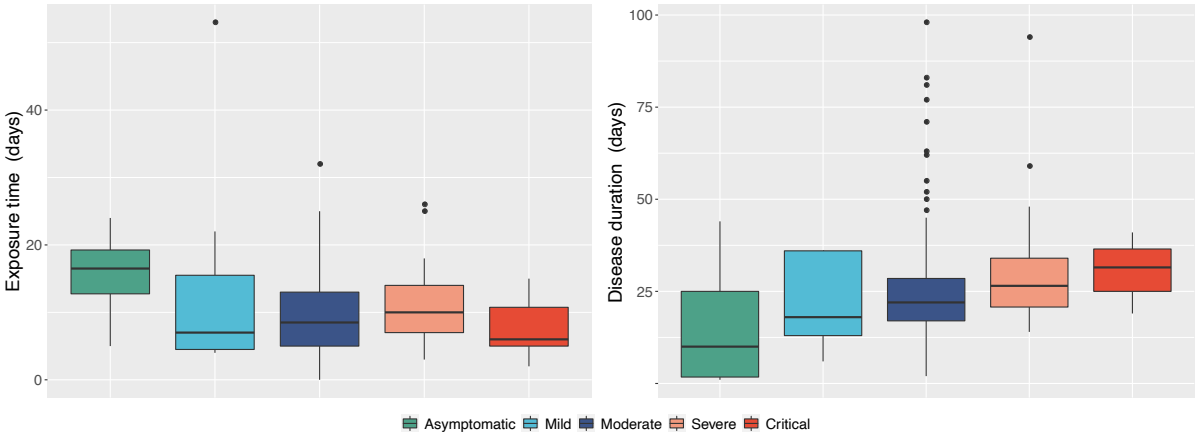


Figure S2. Exposure time and disease duration of the 332 patients by five categories.
A) Exposure time is defined as the time duration between the oral report of the first infected contact and the disease onset. B) Disease duration is defined as the time duration between disease onset and the first negative PCR-test.

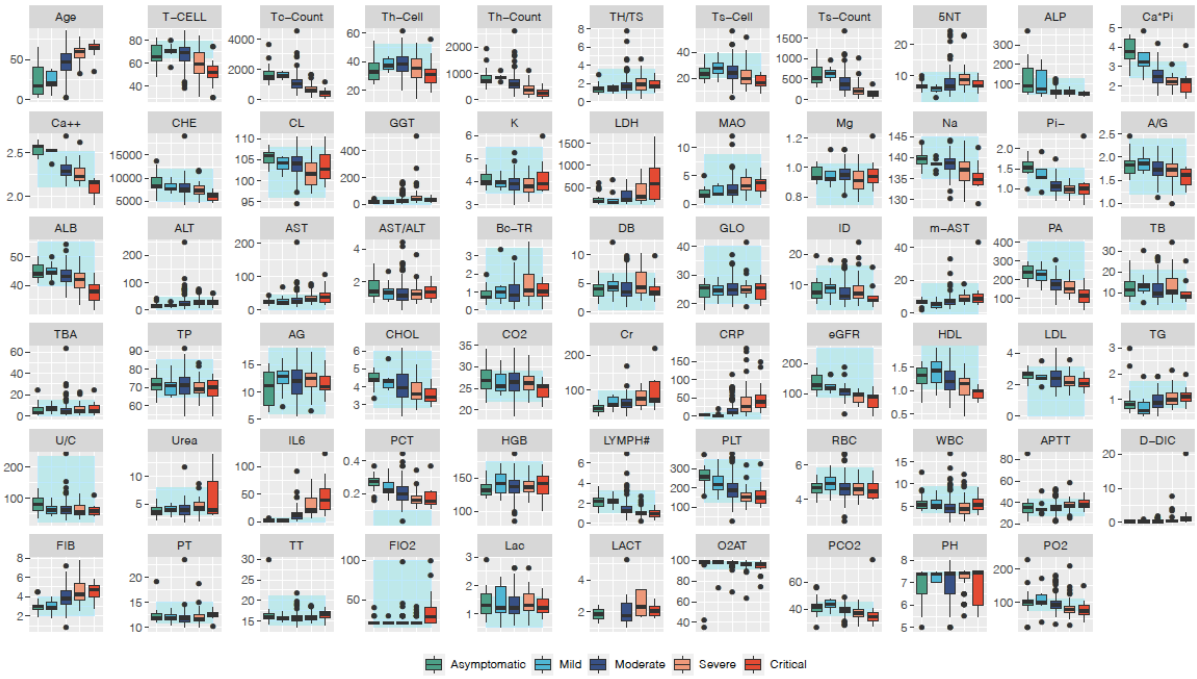


Figure S3. Distribution of age and sixty-four clinical laboratory assessments for the five groups of patients. Light blue boxes indicate the baseline of the healthy population in the electronic health records. The sixty-four laboratory belongs to the seven categories 1) blood count and blood chemical analysis (PLT, HGB, LYMPH#,RBC,WBC) 2) assessments of liver function (A/G, ALB, ALT, AST, AST/ALT, Bc-TR, DB, GLO, ID, m-AST, PA, TB, TBA, T), 3) assessments of renal function (AG, CHOL, CO2, Cr, CRP, eGFR, HDL, LDL, TG, U/C, Urea), 4) tests of humoral immunity (PCT, IL6), 5) tests of coagulation (APTT, D-DIC, FIB, PT, TT) and 6) measures of electrolyte (5NT, ALP, Ca*Pi, Ca++, CHE, CL, GGT, K, LDH, MAO, Mg, Na, Pi-) and 7) blood gas electrolyte (FIO2, Lac, O2AT, PCO2, PH, PO2, LACT).

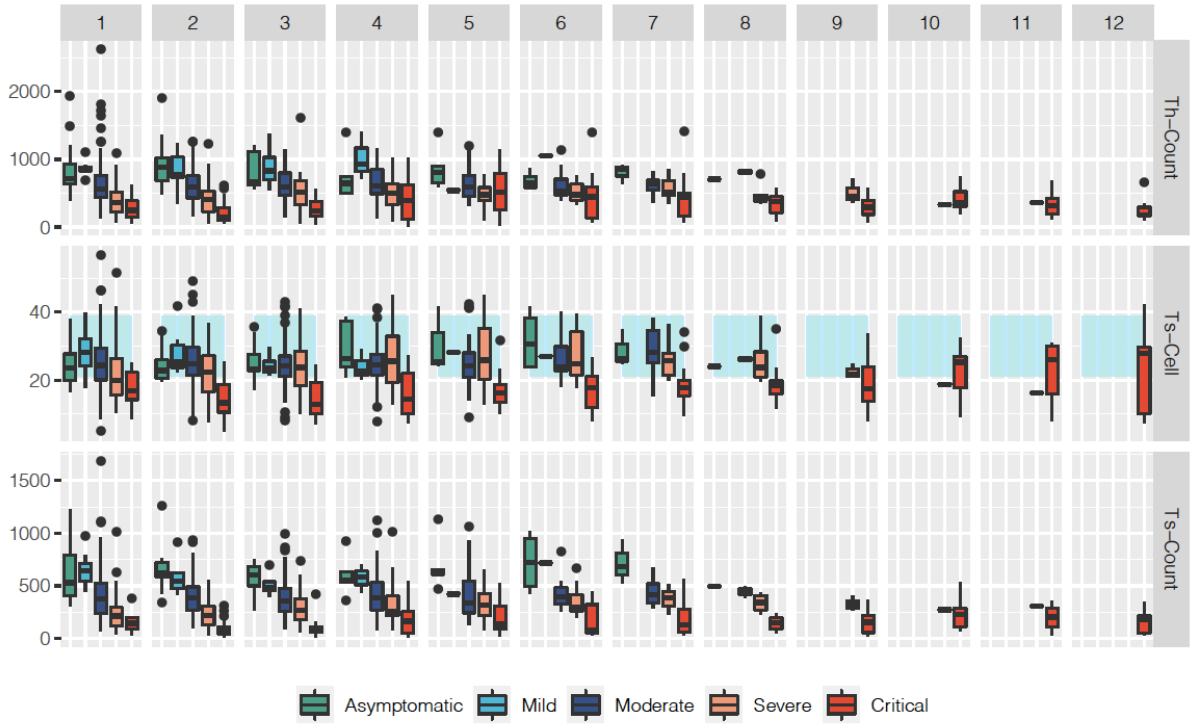
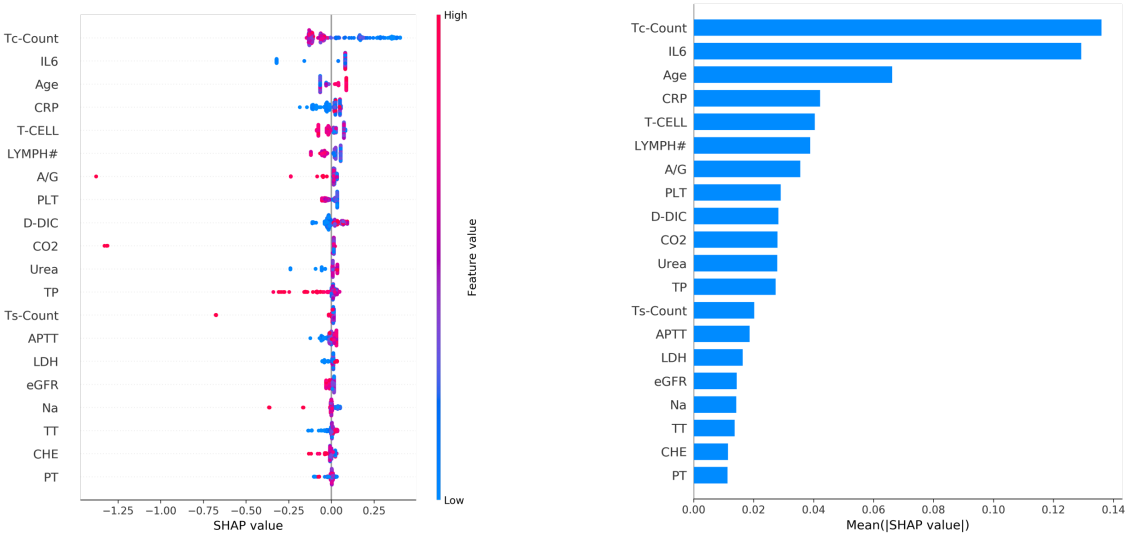


Figure S4. Distribution of the T lymphocyte subgroups for the five disease categories as a function of time ranging from the first to the twelfth assessment.

284 unrelated patients



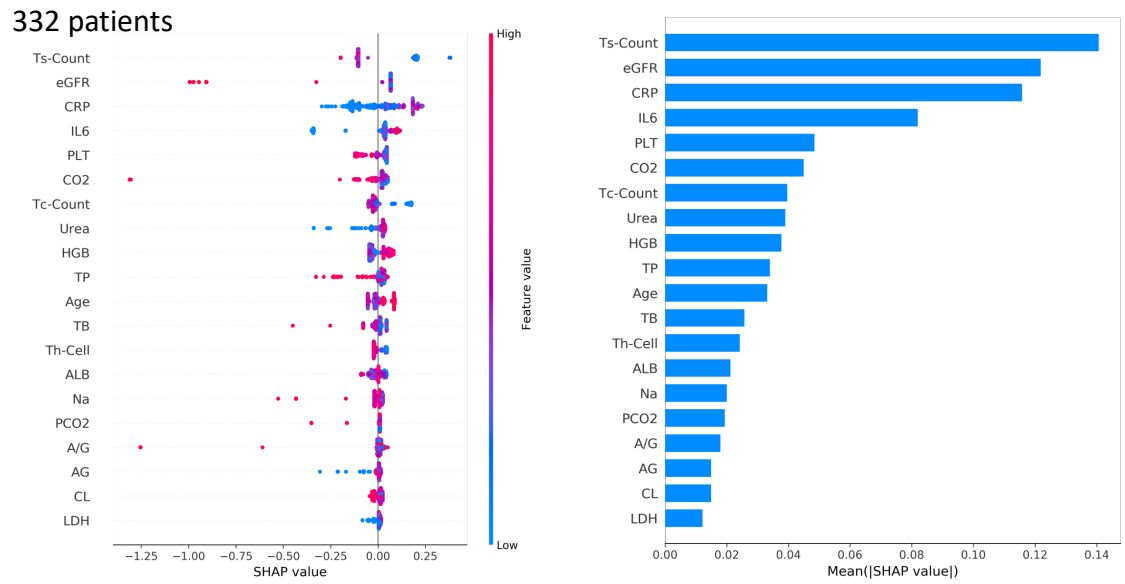


Figure S5. Importance of the sixty four laboratory assessment features to classify the five groups of patients. Shown are the top twenty features of the greatest importance. Shown are the distribution of the SHAP prediction value (which indicates the effect of each feature on the classification of the patient severity) for each laboratory assessment feature (y-axis) for each patient (each dot). Minus and positive values indicate negative and positive effects. Top: distribution for the 284 unrelated patients. Bottom: distribution for all the 332 patients.

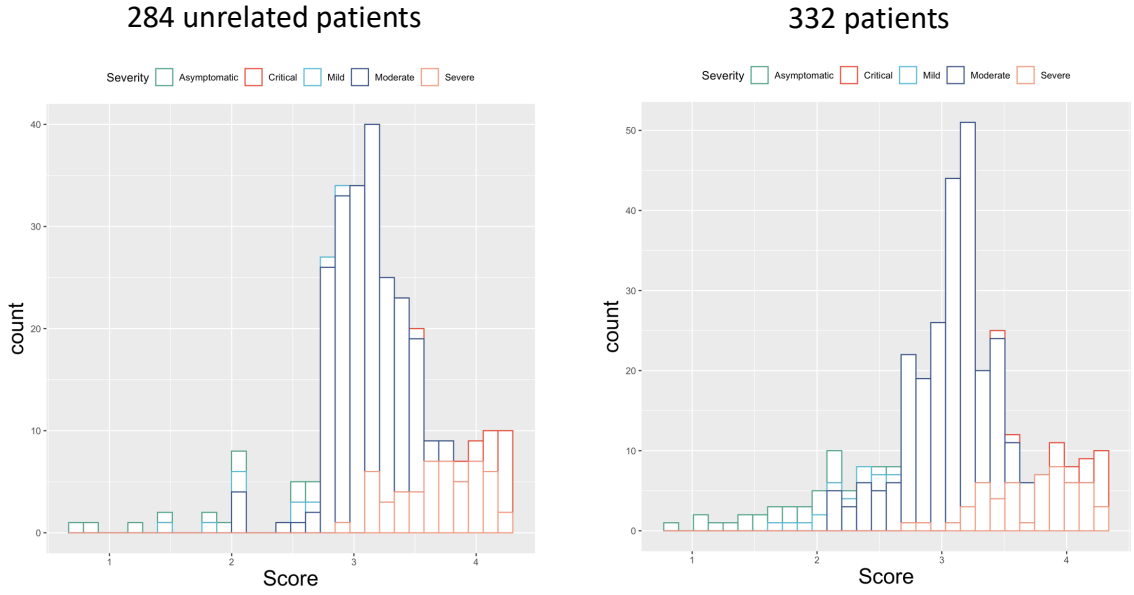


Figure S6. Distribution of the severity score according to the five severity categories

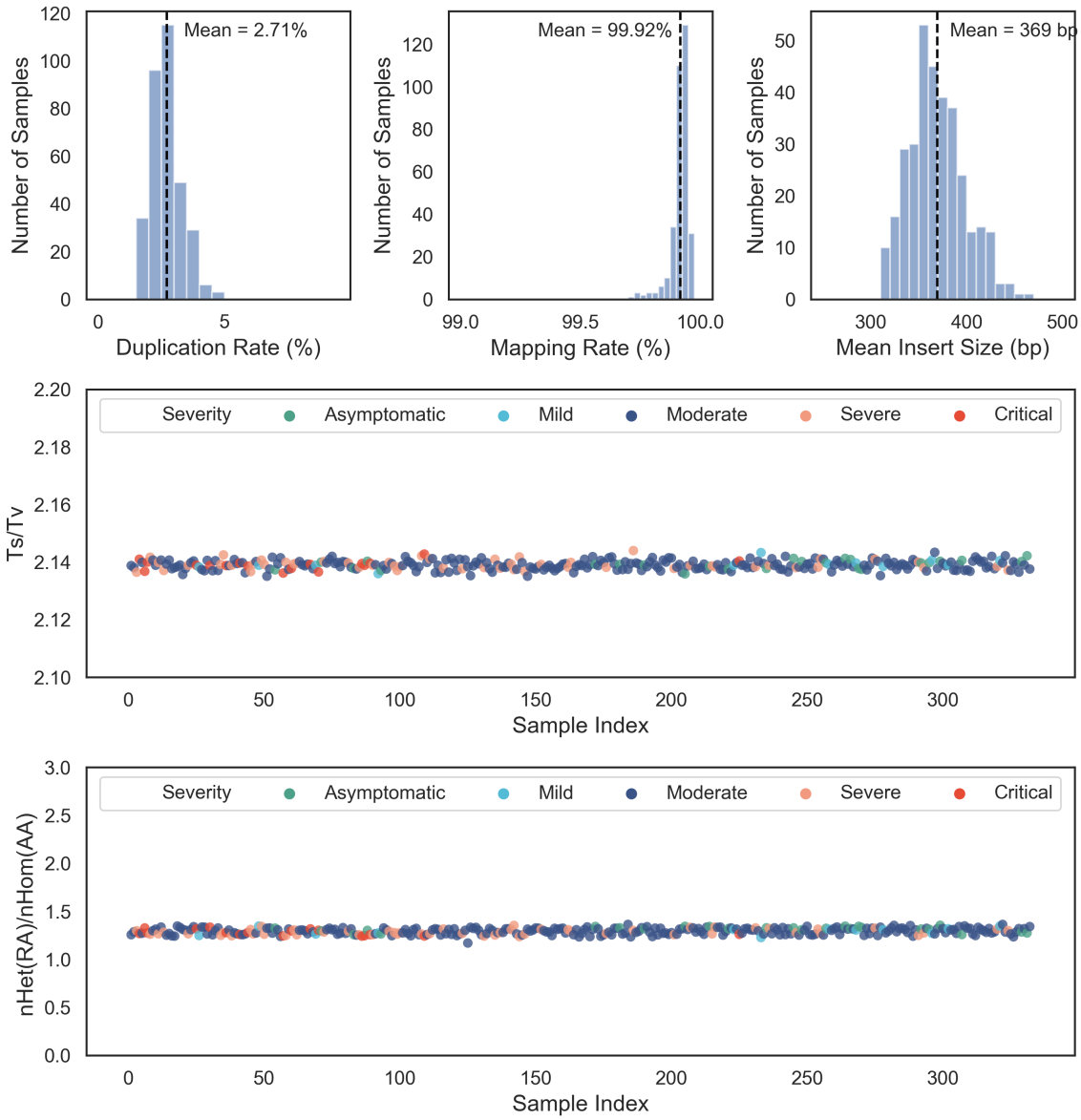


Figure S7. Quality control of individual sequenced genome

Figure S8. Estimated number of loss of function variants for each patient. The label of the y-axis indicates the asymptomatic, mild, moderate, severe and critical group, respectively. The label of the x-axis indicate sample ID of the patient. Frameshift, splice acceptor, splice donor and stop gain variants are shown in green, skype blue, dark blue and red, respectively.

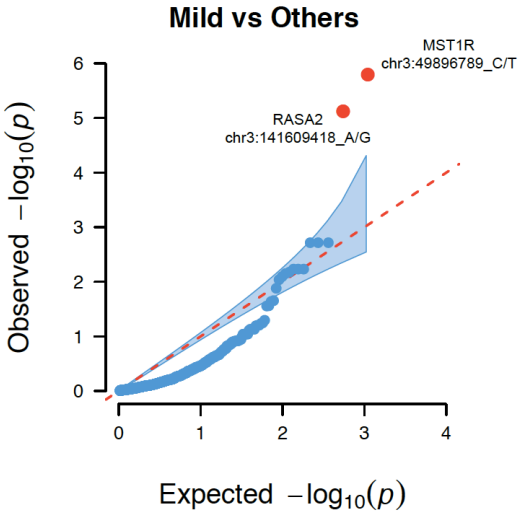


Figure S9. Loss of function variants between the severe and the critically severe group of patients versus the rest of the patients.

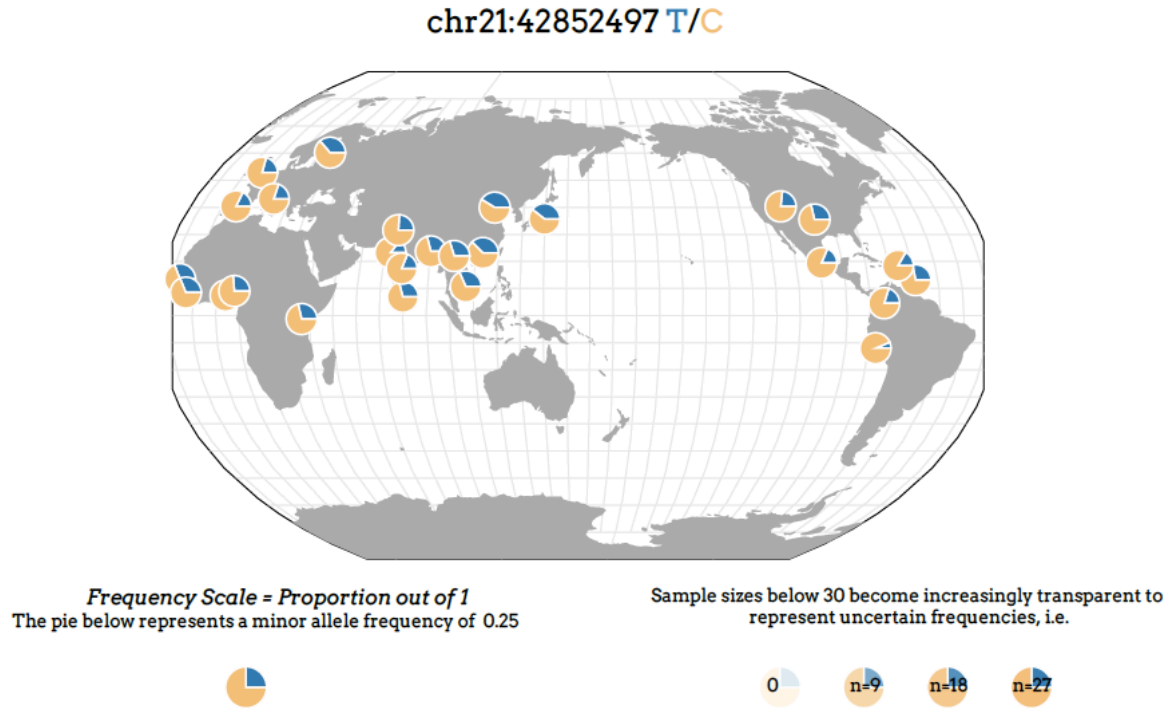


Figure S10. Allele frequency of the p.Val197Met variant in TMPRSS2 among the 1000 genomes populations. The allele frequency of the reference and alternative allele is visualized by the geography of genetic variants browser developed by the university of Chicago. p.Val197Met is located at the 42852497 position in chr21 with rsID rs12329760.

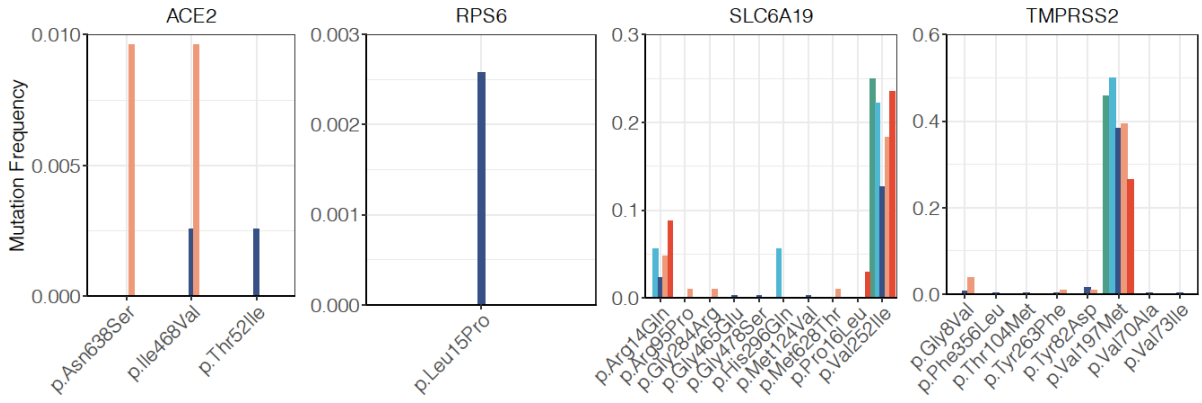
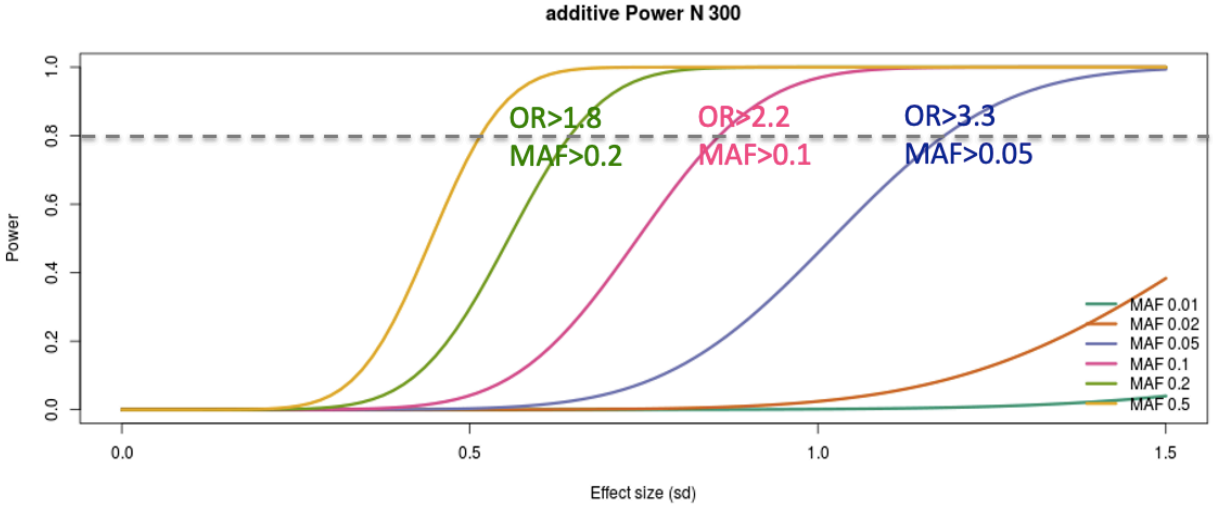


Figure S11. Allele frequency distribution for the functional variants present in nine genes related to the host-pathogen interaction by Sharma et al., 2020 BioRxiv among the five patient severity groups. Genes *ADAM17*, *HNRNPA1*, *SUMO1*, *NACA* and *BTF3* doesn't contain any missense and loss of function variants among the patients. Shown are the allele frequency of the missense variants found in *ACE2*, *PRS6*, *SLC6A19* and *TMPRSS2* among the five patient group. No loss of function variants were present in those genes.



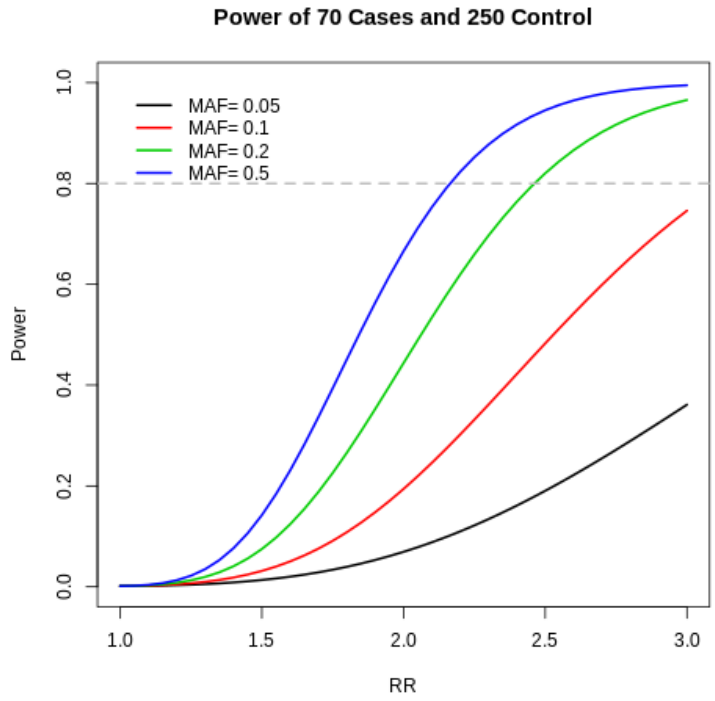
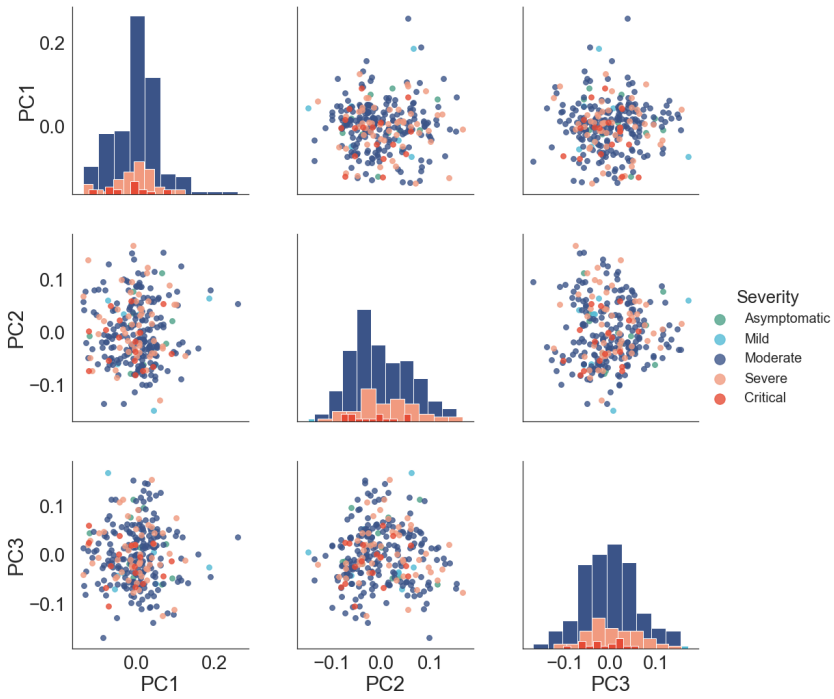


Figure S12. Power for single variant association test. Top: linear regression Bottom: Logistic regression. Significant threshold=1e-7



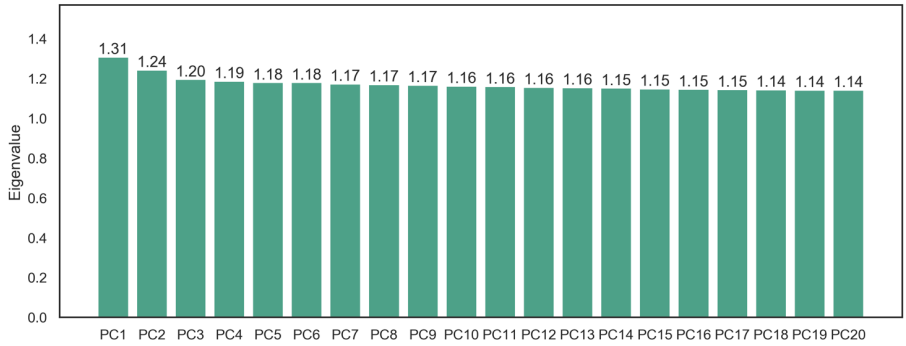
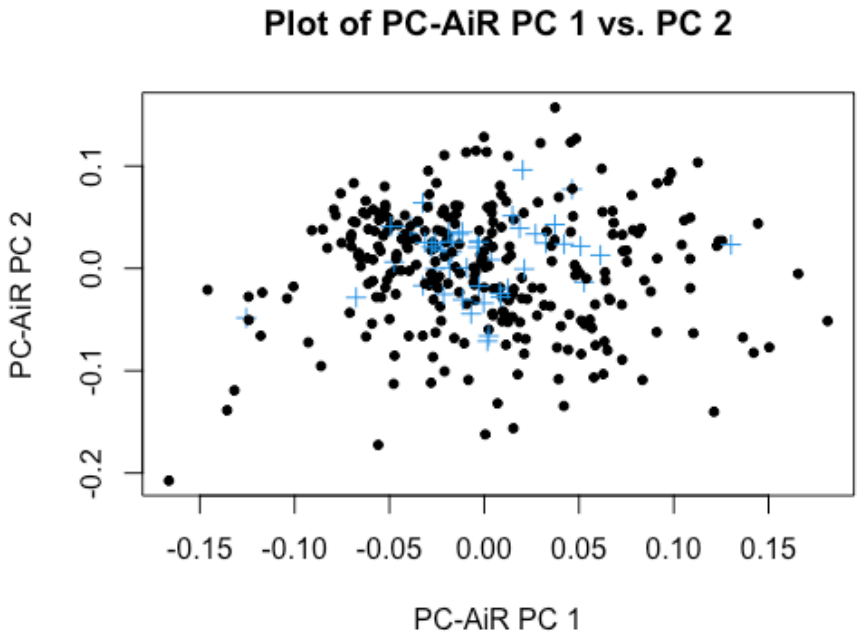


Figure S13. Principle component analysis of the unrelated COVID-19 patients (N=284) recruited in the study. Shown are the top 3 eigenvectors and the top 20 eigenvalues for the principle component analysis using plink.



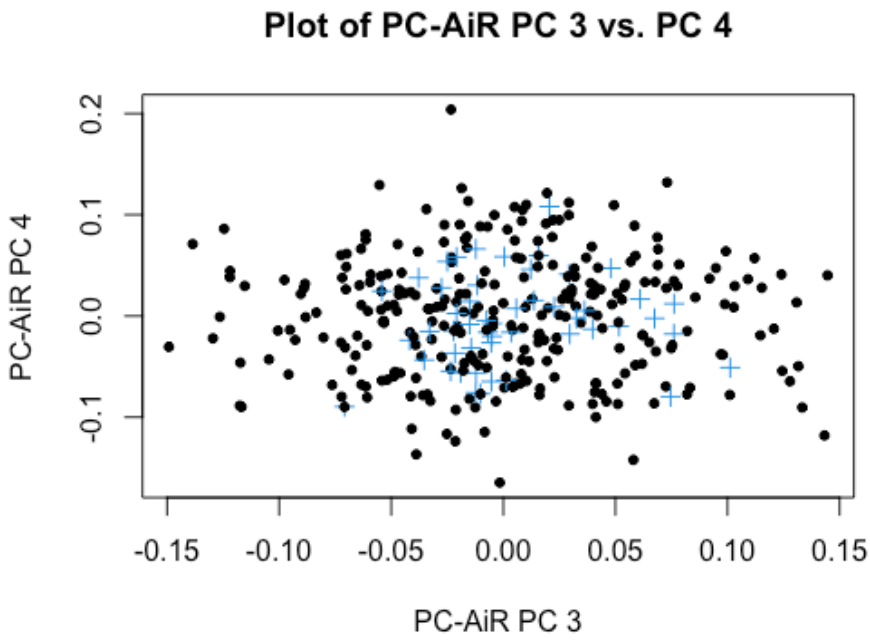


Figure S14. Principle component analysis of the all the COVID-19 patients (N=332) recruited in the study. Shown are the top 4 eigenvectors and the top 20 eigenvalues in the principle component analysis using PC-AiR in Genesis R package. Black dot indicates unrelated individual. Blue cross indicates related individuals.

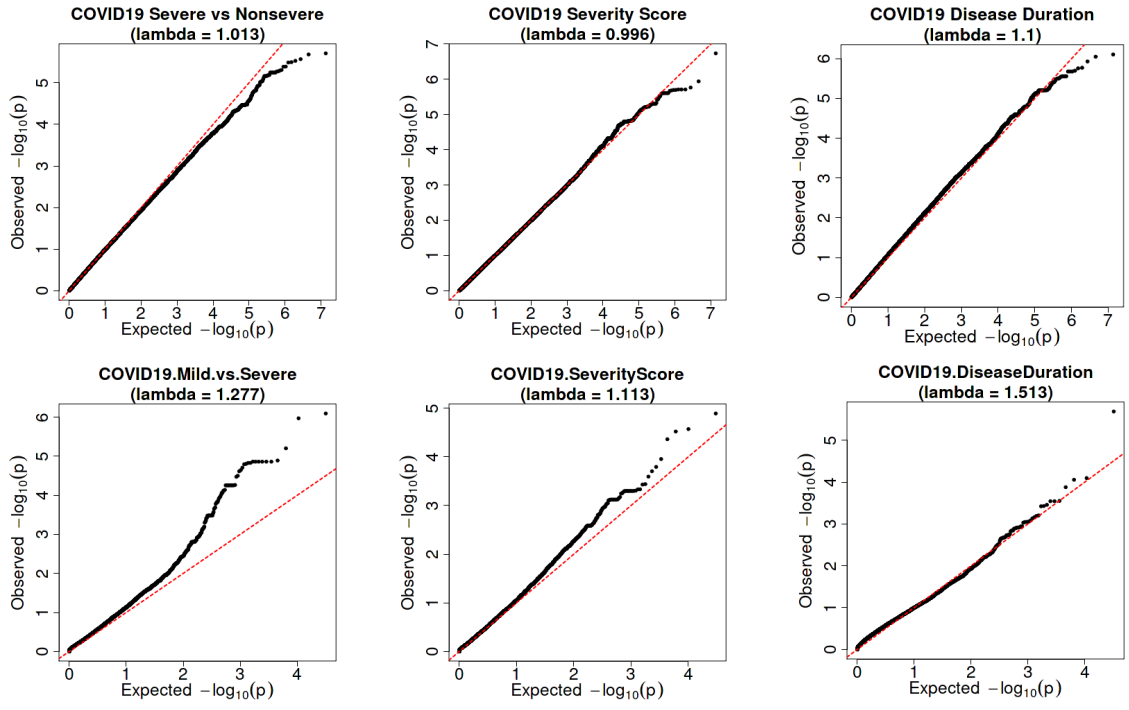


Figure S15. Quantile-quantile plots for the three traits, linked to Figure 3.

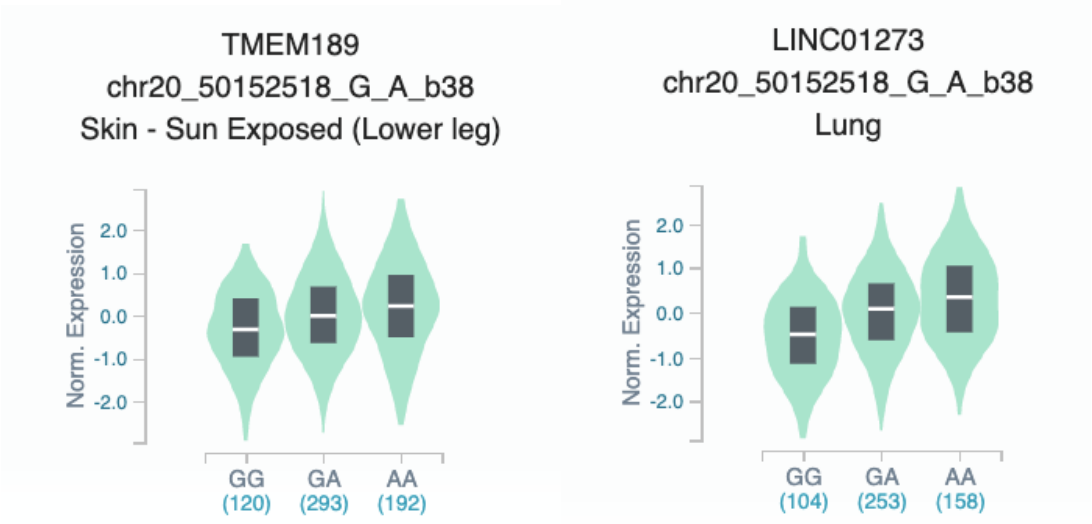
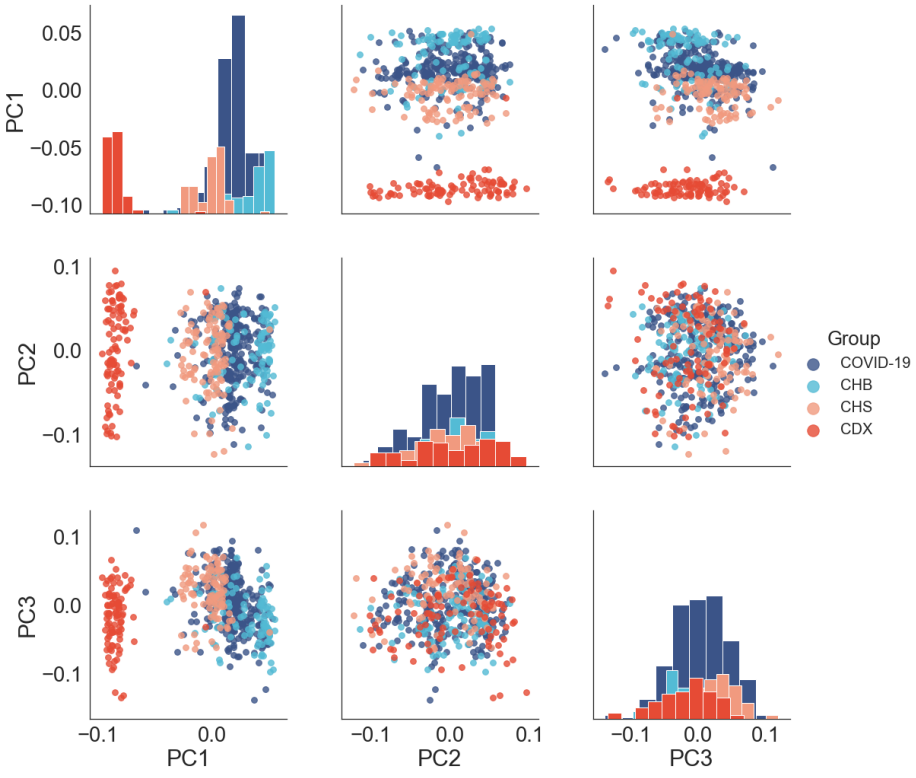


Figure S16. Altered gene expression given the three genotypes at the lead SNP rs6020298. Figures from Gtex portal using the GTEx Analysis Release V8. The A allele increases *LINC01273*, *TMEM189* expression over almost all the tissues. Only two representative plots were shown here.



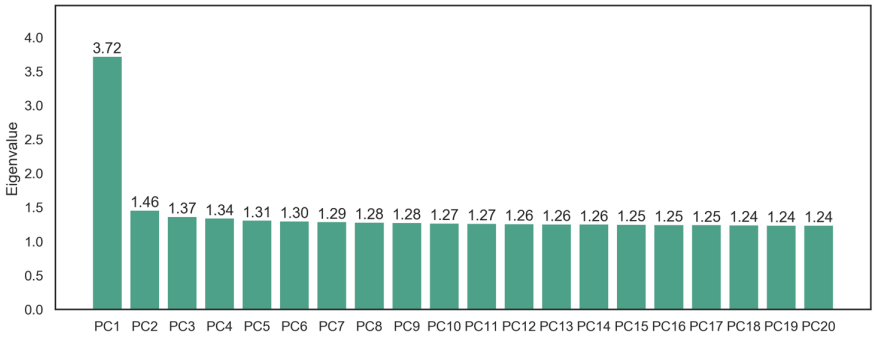
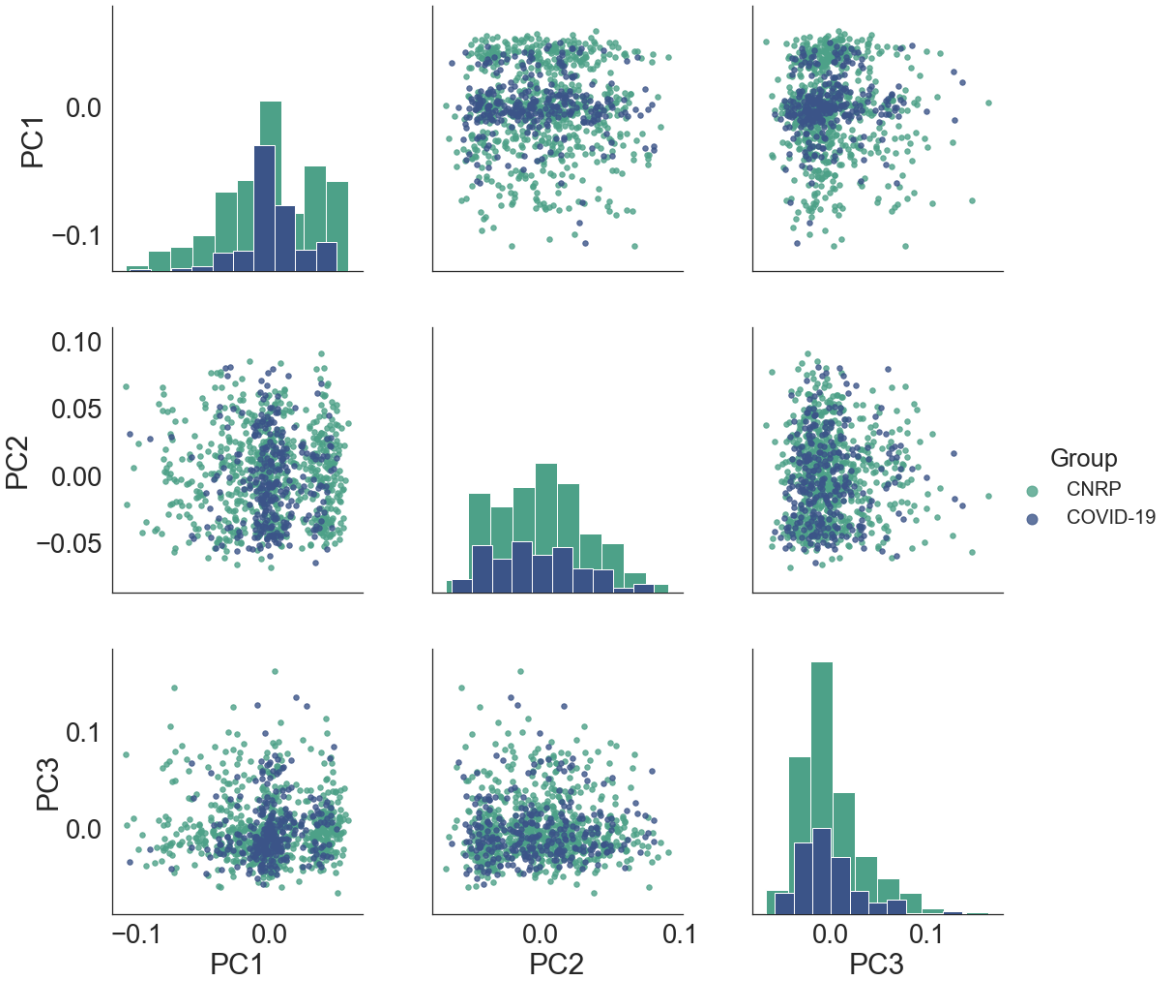


Figure S17. Principle component analysis for the unrelated patients (N=284) and the 1KGP Chinese population (N=301). CHB: Chinese from Beijing, CHS: Chinese from the South, CDX: Chinese Dai in Xishuangbanna, China.



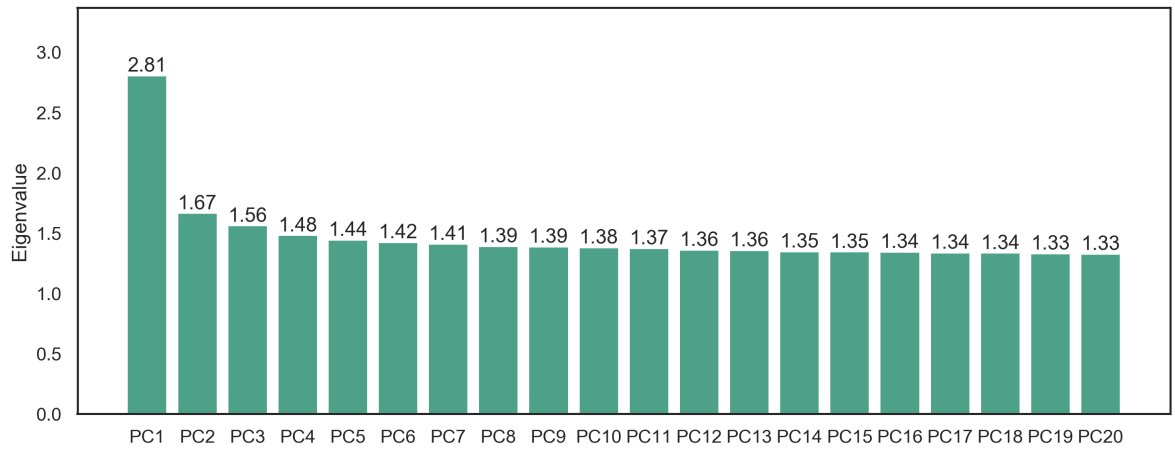


Figure S18. Principle component analysis for the 284 unrelated patients and the 665 control individuals in the Chinese Reference Panel program.

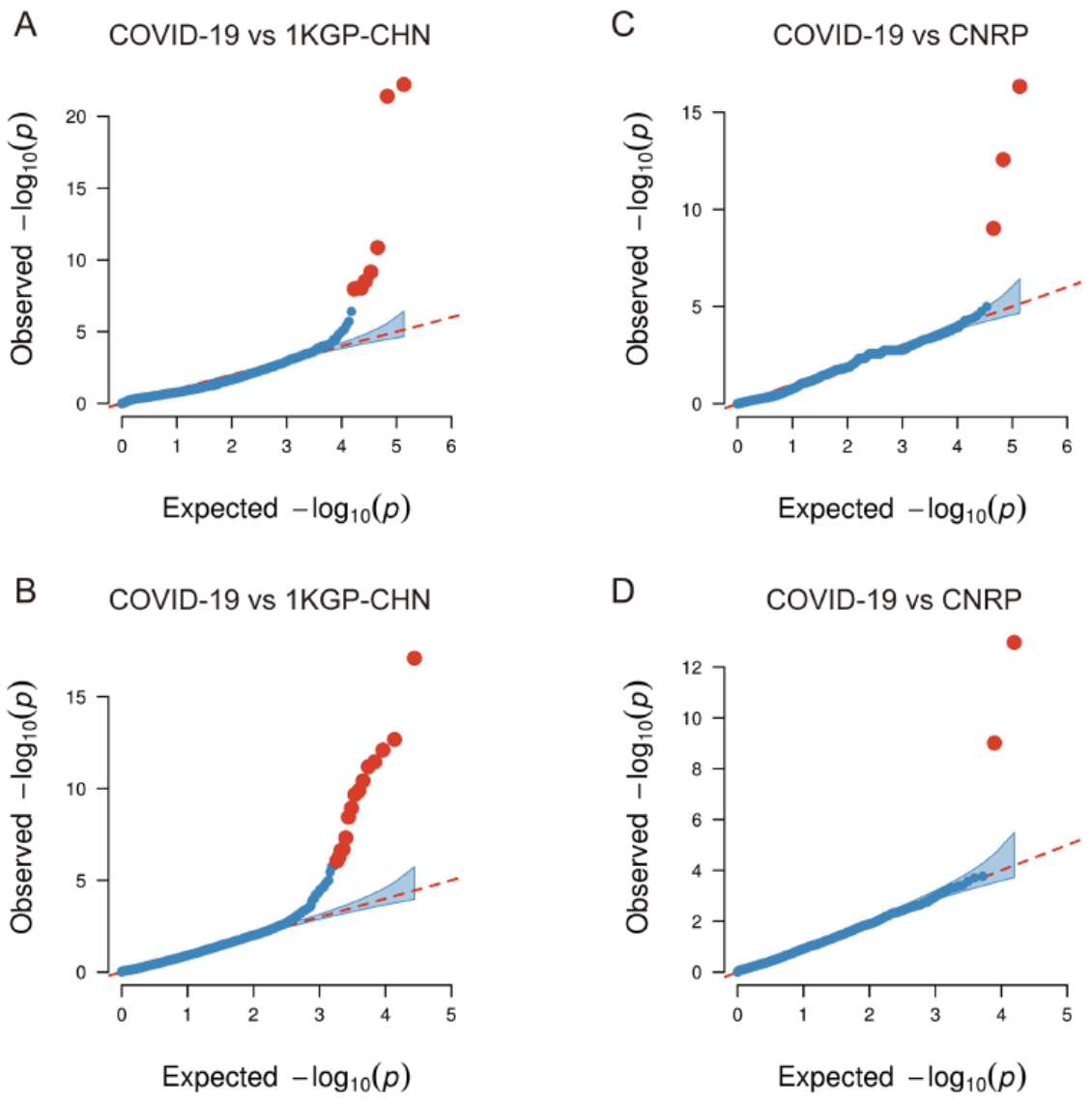


Figure S19. Quantile-quantile plot for single variant and gene-based association test between COVID-19 patients and the general populations. (A) single variant association test and (B) gene-based association test between the unrelated COVID-19 patients (N=284) and the 1KGP Chinese population (N=301) (C) single variant association test and (D) gene-based association test between the unrelated COVID-19 patients (N=284) and the CNRP Chinese population (N=271). Only variants with moderate or high impacts by variant effect predictor were shown in (A) and (C).