Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0356.v1

Initial review and analysis of COVID-19 host genetics and associated phenotypes

Yosuke Tanigawa[1] and Manuel A. Rivas[1]

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, 94305

Abstract

The global pandemic of COVID-19 accounts for more than 14,000 deaths worldwide. However, little is known about the host genetics interaction with infection and COVID-19 progression. To better understand the role of host genetics, we review the current literature, aggregate readily available genetic resources, and provide some updated analysis relevant to COVID-19 and associated phenotypes. Using the unrelated individuals in UK Biobank (total n = 337,579 across 5 populations), we aggregate human leukocyte antigen and ABO blood type frequencies. We find significant and consistent risk reduction of blood group O reported in Zhao et al. and encourage broad sharing of ABO blood type frequencies that are readily accessible across COVID-19 with mild, moderate, and severe/critical symptoms for robust inferences at https://tinyurl.com/abo-covid19. In addition, we generate polygenic risk scores (PRSs) weights for 29 blood measurements, including clinically relevant haematological measurements for COVID-19, such as lymphocyte count and percentage. Focusing on the 8 most COVID-19 clinically relevant blood measurements, we performed PRS-PheWAS analysis across 44 disease antigen measurements (n = 6,643 unrelated individuals in White British group), infectious diseases and acute respiratory infections (n = 20,928 cases and 349,000 controls across 3 population groups) and deaths (n = 1,846 cases and 368,082 controls), recorded in hospital inpatient record and death registry data, respectively, in UK Biobank, and find host genetic PRS associations with disease risk. Taken together, we anticipate these resources

Correspondence to ytanigaw@stanford.edu and mrivas@stanford.edu

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0356.v1

(https://github.com/rivas-lab/covid19) will aid in improving our understanding of host genetic risk factors playing a role in SARS-CoV-2 infection and COVID-19 disease severity.

Keywords: COVID-19, SARS-CoV-2, host genetics, genetics, polygenic risk score

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0356.v1

Introduction

Here, we explore the literature available for COVID-19 disease and aggregate reference datasets like UK Biobank data to better understand the role host genetics may play in affecting viral infection predisposition and COVID-19 progression. Furthermore, we provide polygenic risk score weights for haematological measurements that play a role in the body's immune system that help the body fight viral infection.

HLA frequencies

Motivated by the HIV Controllers Study to define host genetic effects on the outcome of a viral infection where they analyzed the effects of individual amino acids in the human leukocyte antigen (HLA) region of the genome and identified 300 single nucleotide polymorphisms (SNPs) within the region and none elsewhere (Study & The International HIV Controllers Study, 2010), and by the 23andMe study that found that the HLA region is significantly associated with 13 of the 23 common infectious diseases studied(Tian et al., 2017), we sought to aggregate all the available HLA allelotype references including across five groups in UK Biobank(Bycroft et al., 2018) including South Asian (n= 7,885), East Asian (n = 1,154), African (n = 6,497), White British (n = 337,138), and non British European (n = 24,905) with the aim to provide more fine-scale estimates of allelotype frequencies in specific populations by further integrating with country of origin data and external reference data from the 17th International HLA & Immunogenetics workshop (http://17ihiw.org/17th-ihiw-ngs-hla-data/). We made these data publicly available at https://github.com/rivas-lab/covid19/tree/master/UKB_HLA_freq.

ABO blood type

Recently, Zhao et al. investigated the relationship between the ABO blood group from 1775 patients infected with SARS-CoV-2 and compared the distribution to surveys of ABO blood group distribution of 3,694 normal people from Wuhan city and 23,386 people from Shenzhen City (Zhao et al., 2020). In the study they found that **blood group A** had a

significantly higher risk for COVID-19 compared with non-blood group A groups (albeit modest effect size OR = 1.20, *p* = 0.02), whereas **blood group O** had a ***significantly lower*** risk for COVID-19 compared with non-O blood groups (OR = 0.67, p < 0.001). Given this result we sought to aggregate ABO blood group frequencies across population groups we have genotype data available for by using the combination of alleles at three different SNPs that represent the four major ABO antigens (rs8176746, rs687289, rs507666) clearly demonstrated by the Uppsala team (Johansson et al., 2015) (we thank Mike Inouye (@minouye271) for pointing us to this reference). It does seem like the data for rs8176746 is reported on the (-) strand as https://gnomad.broadinstitute.org/variant/9-136131322-G-T?dataset=gnomad_r2_1 shows it is a G/T allele and not C/A. First we generated additional population definitions based on more fine-scale resolution provided by UK Biobank (https://github.com/rivas-lab/covid19/blob/master/ABO/sample_qc_v3.2.self_reported_pop_def.ipynb) using Data Field 21000 (Ethnic background) and compared the frequencies observed in UK Biobank to control groups and COVID-19 patients frequencies in both Shenzhen and Wuhan. Overall, we observed similar ABO blood type inferred frequencies using the three marker haplotype analysis between UK Biobank Chinese and Shenzhen. When we compared O blood group frequencies between Shenzhen controls and UK Biobank Chinese group we did not find differences in frequencies (p = 0.978; OR = 1.003, 95% CI: 0.898 - 1.122), but did find differences in frequencies between Shenzhen 3rd Hospital patients and UK Biobank Chinese group ($p = 9.76 \times 10^{-4}$; OR = 0.629, 95% CI: 0.470 - 0.837) consistent with Zhao et al. observation. Nonetheless, we did find that the frequency of O blood group was different between UK Biobank Chinese group and Wuhan controls (p = 0.00121) suggesting that we should carefully consider inferences regarding ABO blood group differences and hope that these data be made available immediately.

Polygenic risk scores

Lower lymphocyte count has been associated with more severe disease in COVID-19 (Yang

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0356.v1

et al., 2020). Lymphocytopenia occured in more than 80% of critically ill patients in a cohort, and due to targeted invasion by SARS-CoV viral particles damaging the cytoplasmic component of the lymphocyte. In a separate study with 452 COVID-19 patients, severe cases tended to have lower lymphocyte counts and higher leukocyte counts, as well as lower percentage of monocytes, eosinophils, and basophils (**Table 1**)(Qin et al., 2020). To what extent the lower lymphocyte counts seen with disease are related to baseline lymphocyte or other hematologic indices is unclear. Here, we used 29 haematological assays that were performed on whole blood from UK Biobank (Category 10081), to fit multivariate predictive models for phenotypes using the **snpnet** package(Qian et al., 2019). The idea is that there are individuals that have altered counts of white cell types that are crucial to our immune system due to germline genetic factors, which can aid in potentially identifying individuals that are at risk of progressing to severe outcomes. Here, we trained genetic risk prediction models using a genotype data set that is a combination of the directly genotyped array variants, imputation, HLA alleles, and copy number variants for a total of 5,182,706 variants in the analysis(Aguirre et al., 2019). Overall, we find improved predictive performance for haematological measurements of immune cell subtypes involved in viral defense including lymphocyte count and percentage beyond standard covariates like age, sex, and principal components (**Table 1**, **Figure 1**). Furthermore, we find that PRS captures individuals that are likely to have low lymphocyte count and percentage levels that may play a role in susceptibility to viral infection and disease progression. For example, we find that individuals at the bottom 5% of lymphocyte count PRS are likely to have lymphocyte count levels below 1.67 $x10^9$ per litre, and those at the bottom 5% of lymphocyte percentage PRS are likely to have lymphocyte % levels below 25% (**Figure 2**), which similarly can be conducted for neutrophil count and percentage levels (**Supplementary Data 1**). Remarkably, only a few studies have been conducted assessing the relationship between lymphopenia and risk of infection and infection-related death. One such study in large population numbers came from a prospective Danish population-based study in 98,344 individuals (Warny et al., 2018) where they find evidence of individuals with lymphopenia had multivariable-adjusted hazard ratios of 1.41 for any

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0356.v1

infection, 1.31 for pneumonia, and 1.70 for infection-related death. As we begin aggregating host genetic data from COVID-19 patients with clinical trajectories one specific aim will be to assess the association of haematological measurements PRS, which may be better powered than initial genome-wide association scans(International Schizophrenia Consortium et al., 2009). To begin the efforts we provide PRS weights across 29 haematological measurements including the 8 lab measurements found to have significant association between mild and severe COVID-19 disease patients (Qin et al., 2020), which may serve as a useful companion to resources currently available in PGS catalog(Xu et al., 2020) (http://www.pgscatalog.org/).

phenotype	PRS + covariates r-squared	PRS r-squared	covariates r-squared	Direction for severity	Severe range	Normal range
Monocyte %	0.124	0.084	0.042	Lower	6.6 (4.3 - 8.8)	3.0-10.0
Monocyte count	0.103	0.068	0.038	NS		
Eosinophill count	0.096	0.088	0.009	Lower	0.0 (0.0 - 0.0)	0.02-0.52
Lymphocyte %	0.091	0.072	0.02	Lower	14.1 (8.8 - 21.4)	20.0-50.0
Eosinophill %	0.085	0.079	0.008	Lower	0.0 (0.0 - 0.2)	0.4-8.0
Neutrophill count	0.084	0.08	0.004	Higher	4.3 (2.9 - 7.0)	1.8-6.3
White blood cell count	0.072	0.065	0.008	Higher	5.6 (4.3 - 8.4)	3.5 - 9.5
Neutrophill %	0.07	0.067	0.002	Higher	77.6 (68.9 - 86.5)	40.0 - 75.0
Lymphocyte count	0.03	0.019	0.013	Lower	0.8 (0.6 - 1.1)	1.1 - 3.2

Table 1. Blood laboratory genetic prediction and severe COVID-19 patient reference levels. r-squared for prediction model with covariates (covariates r-squared) and polygenic risk scores (PRS r-squared) and combined (PRS + covariates r-squared). Direction for severity comparing non-severe to severe patients according to Qin et al. 2020. Severe COVID-19 patient range (Severe range). Normal range according to Qin et al. 2020.

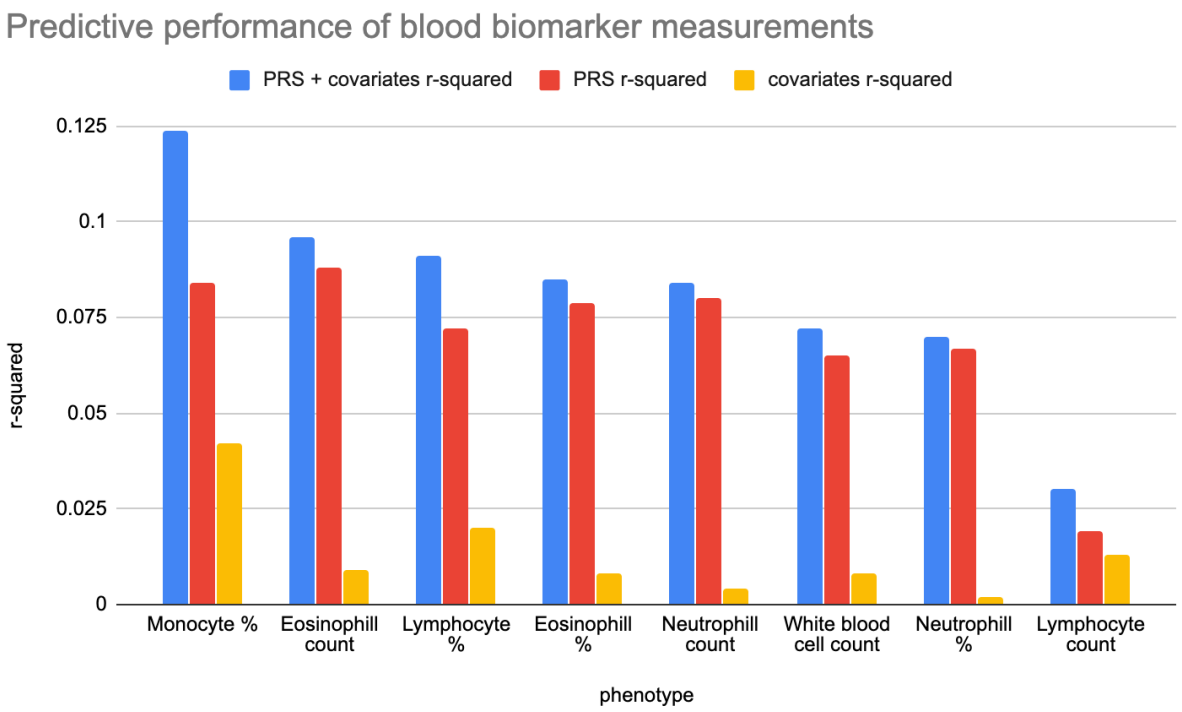


Figure 1. Predictive performance of blood biomarker measurements with covariates (age, sex, and principal components) and genetics (polygenic risk scores [PRS]). Blood laboratory phenotypes with established association between mild and severe COVID-19 patients (x-axis). (blue) PRS + covariates r-squared value in a held-out test set (y-axis), (red) PRS only r-squared, and covariates (orange; age, sex, Array, principal components) r-squared.

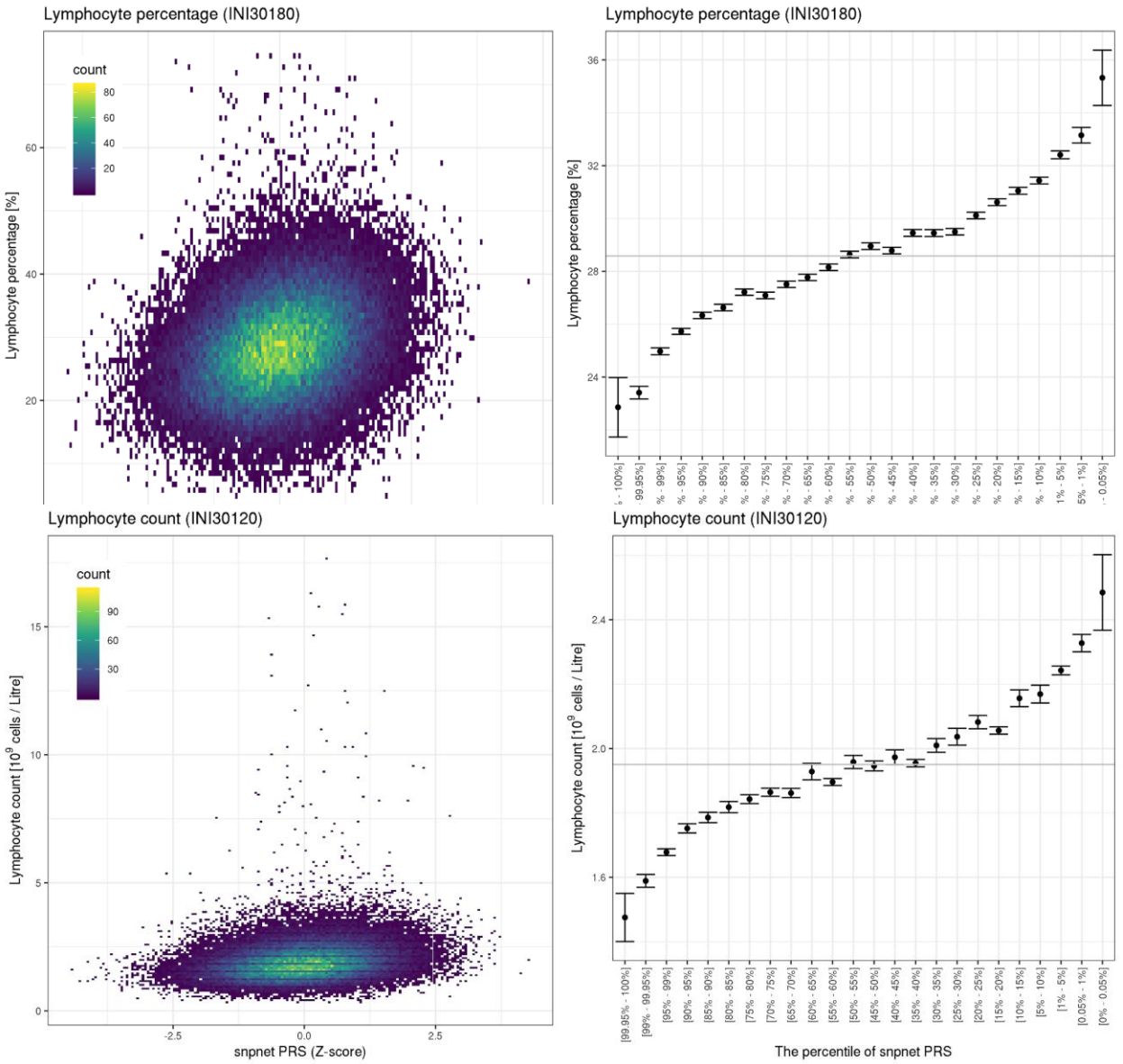


Figure 2. Polygenic risk scores and their relationship to predicted lymphocyte measurements. (top, left) Relationship between PRS for lymphocyte percentage (%) and lymphocyte percentage (%) in a held out test set. (top, right) Lymphocyte percentage (%) and its corresponding standard error at each PRS quantile of lymphocyte percentage PRS. (bottom, left) Relationship between PRS for lymphocyte count and lymphocyte count measurements in a held out test set. (bottom, right) Lymphocyte count and its corresponding standard error at each PRS quantile of lymphocyte count PRS.

Blood count polygenic risk score associations to infectious diseases

As proof of principle, we used the polygenic risk score weights trained with the snpnet package across the eight laboratory biomarkers in Table 1 and Figure 1 to assess association with 44 measured levels of antigens for selected infectious diseases measured in approximately 10,000 individuals in UK Biobank (Category 1307). We found 3 associations at Bonferroni threshold of significance $p < 10^{-4}$ including associations between Neutrophil percentage (%) PRS and IE1A antigen for Human Herpesvirus-6 ($p = 4.4 \times 10^{-5}$; $\beta = -0.05$ per SD, 95% CI: -0.062 - 0.038), White blood cell count PRS and 1gG antigen for Herpes Simplex-virus-1 ($p = 3.34 \times 10^{-5}$; $\beta = 0.05$ per SD, 95% CI: 0.038 - 0.062) and Pp 28 antigen for Human Cytomegalovirus ($p = 5.98 \times 10^{-5}$; $\beta = 0.049$ per SD, 95% CI: 0.037 - 0.061) (**Figure 3, Supplementary Data 2**).

Blood count polygenic risk score associations to respiratory infections, acute respiratory distress syndrome, influenza and pneumonia and death as a result

Finally, we assessed association between blood count polygenic risk scores and respiratory infections, acute respiratory distress syndrome, influenza and pneumonia and death as a result. We aggregated hospital in-patient and death register data from over 337,000 individuals in UK Biobank for ICD codes corresponding to J00-J06, J09-J18, J80, and J20-J22 (UK Biobank Fields 41202, 41204, 40001, 40002, 41201, and 41270 for diseases and 40001 and 40002 for deathes). We find evidence of association between Neutrophil count PRS and risk to disease ($p = 1.04 \times 10^{-33}$; OR = 1.095 per PRS SD, 95% CI: 1.087 - 1.103) and death ($p = 1.33 \times 10^{-6}$, OR = 1.125 per PRS SD; 95% CI: 1.098 - 1.153); white blood cell count PRS and risk to disease ($p = 1.02 \times 10^{-37}$; OR = 1.101 per PRS SD, 95% CI: 1.092 - 1.109) and death ($p = 4.27 \times 10^{-6}$, OR = 1.118 per PRS SD; 95% CI: 1.091 - 1.146); and lymphocyte count and lymphocyte percentage ($p = 3.20 \times 10^{-12}$; OR = 1.053 [1.045 - 1.060]; p = 0.0074, OR = 0.98 [0.973 - 0.988]) to disease (**Figure 4**) with consistent effects in non-British European and South Asian group in UK Biobank (**Supplementary Data 3**). These results suggest that the

PRS are positioned to assess associations with immune response to SARS-CoV-2.

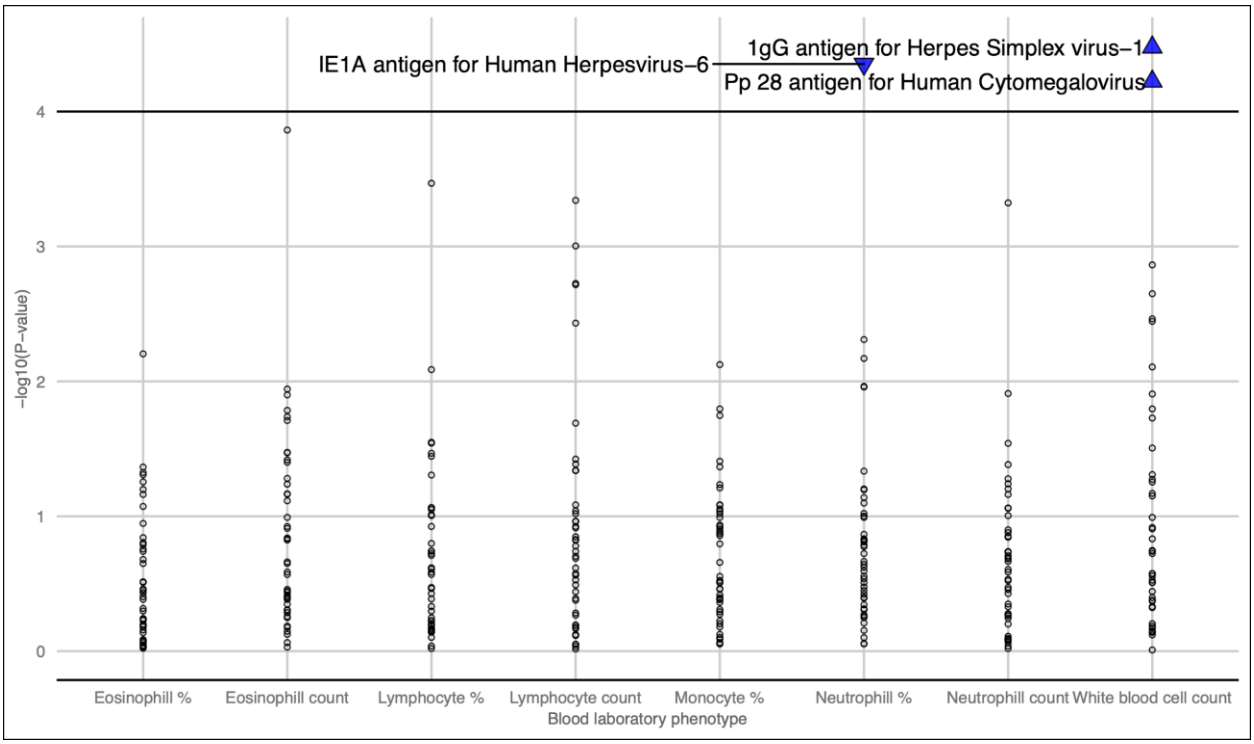


Figure 3. Blood biomarker polygenic risk score association to infectious diseases in UK Biobank. (bottom) Association (-log10(P-value), y-axis) between blood biomarker PRS (x-axis) and infectious disease antigen levels in UK Biobank. Sign of effect is shown by an arrow.

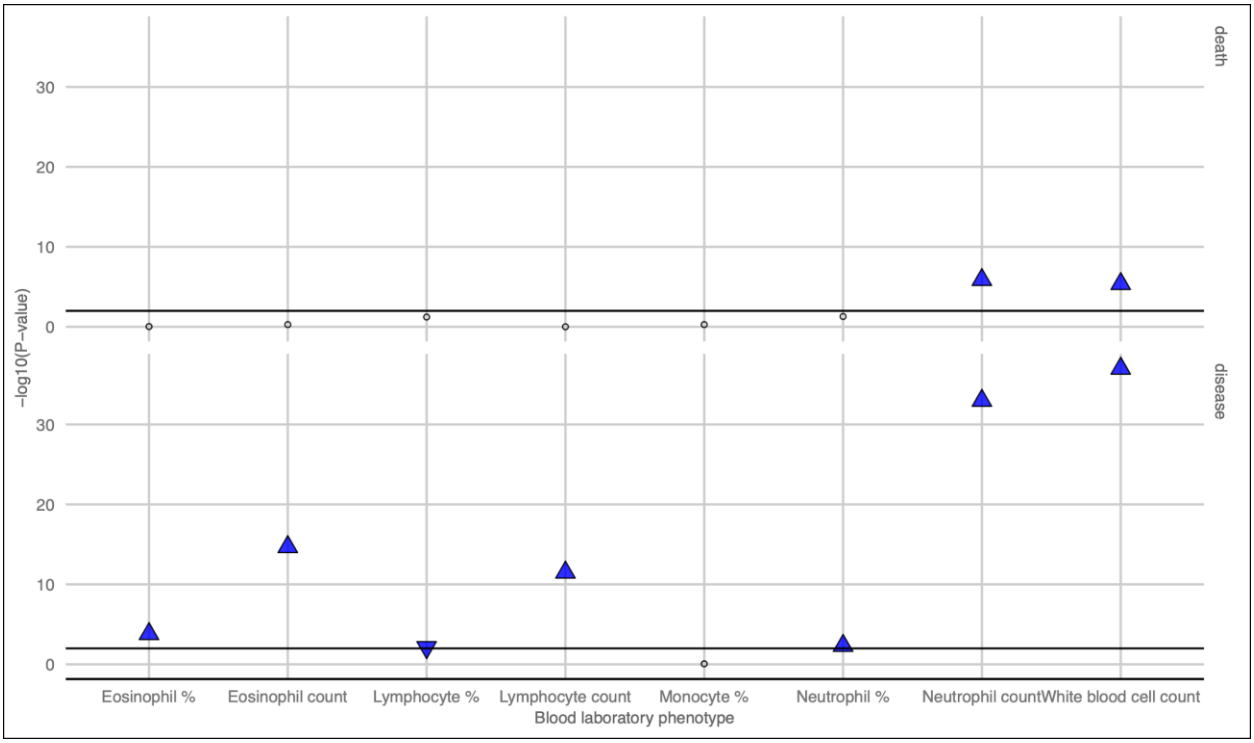


Figure 4. Blood biomarker polygenic risk score association to respiratory infections, acute respiratory distress syndrome, influenza and pneumonia and death as a result in UK Biobank. Association (-log10(P-value), y-axis) between blood biomarker PRS (x-axis) and respiratory infections, acute respiratory distress syndrome, influenza and pneumonia (bottom) and death as a result (top) in UK Biobank. Sign of effect is shown by an arrow.

Discussion

We present a review and analysis of COVID-19 host genetics and related phenotypes. We present some reference datasets that we hope will aid ongoing studies to improve our understanding of the role that host germline genetics plays in infection predisposition and COVID-19 disease progression including: 1) HLA allelotype and 2) ABO blood group frequencies across different population groups, and 3) polygenic risk score weights across haematological measurements.

Based on our review and analysis we find some support for O blood group protection. However, we caution that further data aggregation across independent populations will be

needed to make robust inferences. One such effort is the International COVID-19 Host Genetics Initiative (https://covid-19genehostinitiative.net/), where we are intimately involved. These efforts will include the generation of host genotype data via arrays, exome, and whole genome sequencing combined with analysis of phenotypes. Given that blood group information is accessible we encourage all areas around the globe to make those data readily available. Simple summary statistics like those provided in Zhao et al. (distribution of blood group frequencies across all COVID-19 patients, mild COVID-19 patients, and severe COVID-19 patients) should expedite integration of host genetics. We make a publicly accessible Google Spreadsheet available for personnel to input ABO blood group summary level data **https://tinyurl.com/abo-covid19** .

Insights gained from the PRS analysis include the association between blood count measurement PRS to infectious disease, respiratory infections, acute respiratory distress syndrome, influenza and pneumonia and death as a result. These results suggest we should consider PRS analysis of blood count measurements in the context of SARS-CoV-2 infection and COVID-19 disease severity.

Acknowledgements

We thank Mijail Rivas for useful advice on the clinical manifestations observed in COVID-19 pathogenesis and progression. We thank Vijay Sankaran for providing feedback on the PRS analysis and qualifying that baseline lymphocyte count may or may not relate to alterations upon severe infections. We thank Marcelo Fernandez-Vina for highlighting the international HLA reference and Euan Ashley for highlighting the ABO blood group work on COVID19 risk. We thank Mike Inouye and Alex Dilthey for their continued help in identifying relevant references. We thank Stanford Research Computing Center for providing a prioritized queue for COVID-19 research. Y.T. is supported by a Funai Overseas Scholarship from the Funai Foundation for Information Technology and the Stanford University School of Medicine.

 doi:10.20944/preprints202003.0356.v1

Online resources

Analysis scripts and notebooks

https://github.com/rivas-lab/covid19

HLA allelotype frequencies

https://github.com/rivas-lab/covid19/tree/master/UKB_HLA_freq

ABO blood group frequencies

https://github.com/rivas-lab/covid19/tree/master/ABO

Publicly accessible Spreadsheet to share ABO blood group counts

https://tinyurl.com/abo-covid19

Polygenic risk score (PRS) weights

https://github.com/rivas-lab/covid19/tree/master/snpnet

 doi:10.20944/preprints202003.0356.v1

References

Aguirre, M., Rivas, M. A., & Priest, J. (2019). Phenome-wide Burden of Copy-Number Variation in the UK Biobank. *American Journal of Human Genetics*, *105*(2), 373–383.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209.

International Schizophrenia Consortium, Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752.

Johansson, Å., Alfredsson, J., Eriksson, N., Wallentin, L., & Siegbahn, A. (2015). Genome-Wide Association Study Identifies That the ABO Blood Group System Influences Interleukin-10 Levels and the Risk of Clinical Events in Patients with Acute Coronary Syndrome. *PloS One*, *10*(11), e0142518.

Qian, J., Du, W., Tanigawa, Y., Aguirre, M., Tibshirani, R., Rivas, M. A., & Hastie, T. (2019). *A Fast and Flexible Algorithm for Solving the Lasso in Large-scale and Ultrahigh-dimensional Problems*. https://doi.org/10.1101/630079

Qin, C., Zhou, L., Hu, Z., Zhang, S., Yang, S., Tao, Y., Xie, C., Ma, K., Shang, K., Wang, W., & Tian, D.-S. (2020). Dysregulation of immune response in patients with COVID-19 in Wuhan, China. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*. https://doi.org/10.1093/cid/ciaa248

Study, T. I. H. C., & The International HIV Controllers Study. (2010). The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. In *Science* (Vol. 330, Issue 6010, pp. 1551–1557). https://doi.org/10.1126/science.1195271

Tian, C., Hromatka, B. S., Kiefer, A. K., Eriksson, N., Noble, S. M., Tung, J. Y., & Hinds, D. A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nature Communications*, *8*(1), 599.

Warny, M., Helby, J., Nordestgaard, B. G., Birgens, H., & Bojesen, S. E. (2018). Lymphopenia and risk of infection and infection-related death in 98,344 individuals from a prospective Danish population-based study. *PLoS Medicine*, *15*(11), e1002685.

Xu, Y., Vuckovic, D., Ritchie, S. C., Akbari, P., Jiang, T., Grealey, J., Butterworth, A. S., Ouwehand, W. H., Roberts, D. J., Di Angelantonio, E., Danesh, J., Soranzo, N., & Inouye, M. (2020). Learning polygenic scores for human blood cell traits. In *bioRxiv* (p. 2020.02.17.952788). https://doi.org/10.1101/2020.02.17.952788

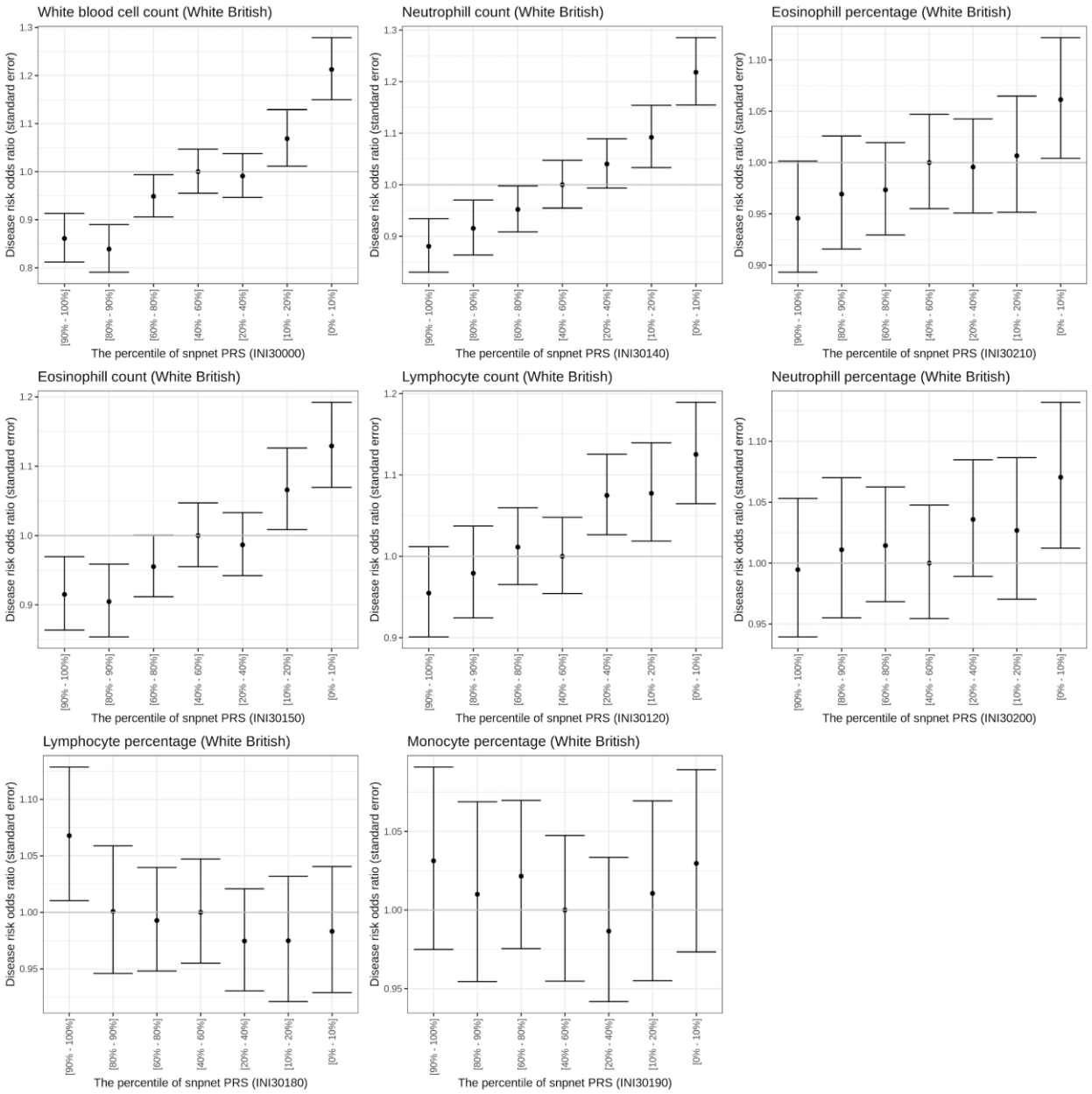
Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J. 'an, Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., Yu, T., Wang, Y., Pan, S., Zou, X., Yuan, S., & Shang, Y. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. In *The Lancet Respiratory Medicine*. https://doi.org/10.1016/s2213-2600(20)30079-5

Zhao, J., Yang, Y., Huang, H.-P., Li, D., Gu, D.-F., Lu, X.-F., Zhang, Z., Liu, L., Liu, T., Liu, Y.-K., He, Y.-J., Sun, B., Wei, M.-L., Yang, G.-Y., Wang, X., Zhang, L., Zhou, X.-Y., Xing, M.-Z., & Wang, P. G. (2020). Relationship between the ABO Blood Group and the COVID-19 Susceptibility. *medRxiv*, 2020.03.11.20031096.

 doi:10.20944/preprints202003.0356.v1

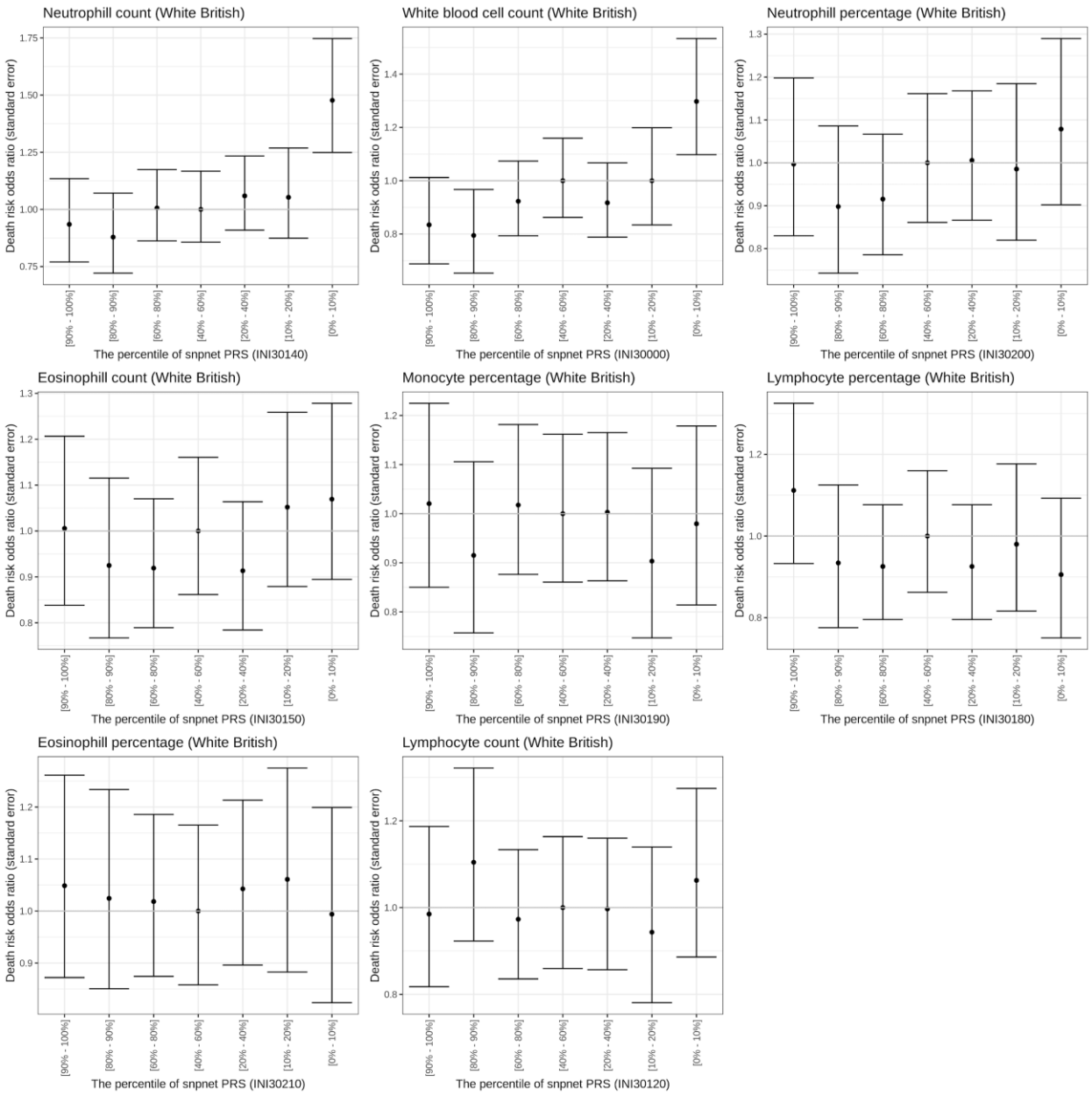
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0356.v1

Supplementary Figures



Supplementary Figure 1. Polygenic predictions for the disease risks of infectious diseases and acute respiratory infections. Each panel represents the percentile bins (x-axis) for one of the 8 blood measurements and the risks of infectious diseases and acute respiratory infections (yaxis).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0356.v1



Supplementary Figure 2. Polygenic predictions for the mortality risks of infectious diseases and acute respiratory infections. Each panel represents the percentile bins (x-axis) for one of the 8 blood measurements and the mortality risks of infectious diseases and acute respiratory infections (yaxis).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0356.v1

Supplementary Data

Supplementary Data 1. Polygenic prediction of 8 blood measurements. For each of the 8 blood measurements (each column), we report the mean and standard error stratified by percentile bin (Percentile bin column).

Supplementary Data 2. Polygenic predictions performance across 44 disease antigen measurements. For the unrelated individuals in White British population in UK Biobank, we report the BETA, its standard error (BETA column) and p-value (P column) of the scaled polygenic risk score for one of the 8 blood measurements (PRS and PRS_pheno columns) for disease antigen measurements (GBE_ID and "disease antigen measurements" columns).

Supplementary Data 3. Polygenic predictions performance for the disease and mortality risks of infectious diseases and acute respiratory infections. For each of the three populations (White British, Non-British white, and South Asian, recorded in population column), we report the odds ratio, its standard error (OR column) and p-value (P column) of the scaled polygenic risk score for one of the 8 blood measurements (PRS and PRS_pheno columns) for disease risk or the mortality risk (phenotype column).