Investigation of Target Sequencing of SARS-CoV-2 and Immunogenic Profiling in Host Cells of COVID-19 in Vietnam

Tham Hoang (✉ v.thamhh@vinbigdata.org)

Vingroup Big Data institute

Giang Vu

Vingroup Big Data institute

Mai Tran

Vingroup Big Data institute

Nam Vo

Vingroup Big Data institute

Quang Le

Hanoi University of Civil Engineering

Trang Tran

Vingroup Big Data institute

Research Article

Keywords: SARS-CoV-2, COVID-19 severity, Vietnam, clade, PRS

Posted Date: December 15th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1092321/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Investigation of target sequencing of SARS-CoV-2 and immunogenic profiling in host cells of COVID-19 in Vietnam

Tham H. Hoang^{1*}, Giang M. Vu¹, Mai H. Tran¹, Trang T.H. Tran¹, Quang D. Le² and Nam S. Vo^{1,3}

*Correspondence:

v.thamhh@vinbigdata.org

¹Center for Biomedical Informatics, Vingroup Big Data Institute, 458 Minh Khai, Hai Ba Trung, Hanoi, Vietnam
Full list of author information is available at the end of the article

Abstract

Background: A global pandemic has been declared for coronavirus disease 2019 (COVID-19), which has serious impacts on human health and healthcare systems in the affected areas, including Vietnam. None of the previous studies have a framework to provide summary statistics of the virus variants and assess the severity associated with virus proteins and host cells in COVID-19 patients in Vietnam.

Method: In this paper, we comprehensively investigated SARS-CoV-2 variants and immune responses in COVID-19 patients. We provided summary statistics of target sequences of SARS-CoV-2 in Vietnam and other countries for data scientists to use in downstream analysis for therapeutic targets. For host cells, we proposed a predictive model of the severity of COVID-19 based on public datasets of hospitalization status in Vietnam, incorporating a polygenic risk score. This score uses immunogenic SNP biomarkers as indicators of COVID-19 severity.

Result: We identified that the Delta variant of SARS-CoV-2 is most prevalent in southern areas of Vietnam and it is different from other areas in the world using various data sources. Our predictive models of COVID-19 severity had high accuracy (Random Forest AUC = 0.81, Elastic Net AUC = 0.7, and SVM AUC = 0.69) and showed that the use of polygenic risk scores increased the models' predictive capabilities.

Conclusion: We provided a comprehensive analysis for COVID-19 severity in Vietnam. This investigation is not only helpful for COVID-19 treatment in therapeutic target studies, but also could influence further research on the disease progression and personalized clinical outcomes.

Keywords: SARS-CoV-2; COVID-19 severity; Vietnam; clade; PRS

1 Introduction

The novel coronavirus disease 2019 (COVID-19) is a respiratory illness caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). COVID-19 was first reported as an outbreak in Wuhan, China, and proceeded to spread worldwide, resulting in the declaration of a pandemic. The presentation of COVID-19 can range from mild symptoms of fever, cough, headache, muscular pain, nausea, and vomiting to a severe illness characterized by pneumonia, acute respiratory distress syndrome, septic shock, and multi-organ failure [1]. COVID-19 continues to spread around the world, with over 234 million cases and almost 4.8 million deaths as of October 4th, 2021 according to Johns Hopkins university [2].

The ongoing fourth wave of COVID-19 infections in Vietnam is more serious than the previous three. According to the Vietnam Ministry of Health, despite drastic action, Ho Chi Minh City and other southern provinces of Vietnam in particular were still facing complex COVID-19 outbreaks, with more negative impacts on daily life and socio-economic development than in the previous waves. According to the report to WHO, from 3 January 2020 to 5:54pm CEST, 1 October 2021, Vietnam has 790,755 confirmed cases of COVID-19 with 19,301 deaths.

In term of genomic organization, SARS-CoV-2 genome sequence is approximately 27-30 kb in length. This includes two large genes—ORF1a and ORF1b—which encode 16 non-structural proteins (NSP1–NSP16), as well as genes encoding structural proteins S, E, M, and N. One mutation, D614G, is known to have first emerged in the spike protein S, which is responsible for the attachment of the virus to angiotensin-converting enzyme 2, the receptor for SARS-CoV-2 entry into human cells. This European origin variant was dominant in Vietnam in the early March 2020 [3]. There is also evidence of mutations in the receptor-binding domain of the S protein, which are of very high concern given that they can directly influence viral infectivity, transmissibility, and resistance to neutralizing antibodies and T cell responses [4]. Some variants rise rapidly in frequency and then collapse and disappear, while others rise and overtake the reference strain, becoming the new reference. Examples of these include B.1.1.7 (United Kingdom variant), B.1.351 (South African variant), B.1.1.28 (Brazilian variant), and B.1.617.2 (Indian variant) [5] [6].

In the blood atlas of COVID-19 hallmarks, Ahern et al. indicated several factors beneficial to the treatment of severe COVID-19 patients, including glucocorticoids (dexamethasone), inhibitors of the IL-6 receptor (tocilizumab/sarilumab), and Janus kinases (baricitinib) [7], [8], [9], [10], [11]. Blood-derived signatures that are associated with the disease's severity are immune suppression, myeloid dysfunction, lymphopenia, interferon-driven immunopathology, T cell activation/exhaustion, and immune senescence [12], [13], [14], [15], [16], [17]. In lung tissue, signs include neutrophil and macrophage infiltration, T cell cytokine production and alveolitis, as well as altered redox balance, endothelial damage, and thrombosis [18]. In addition, treatment of patients with corticosteroids, intravenous immunoglobulin, and selective cytokine blockades (tocilizumab) have been associated with higher risk of severe disease [19], [20], [21].

A recent study reported 13 genome-wide significant loci that are associated with SARS-CoV-2 infection or severe manifestations of COVID-19. Several of these loci correspond to previously documented associations with lung, autoimmune, and inflammatory diseases [22]. Downes et al. 2021 indicates LZTFL1 as a candidate effector gene at a COVID-19 risk locus in South Asian [23]. Prognostic factors combined with predictive risk models could lead to differentiation of COVID-19 patients based on their risk of severe disease or death. This risk stratification may subsequently guide better disease treatment and personalized outcomes [24]. A polygenic risk score (PRS) that aggregates the information of many common single-nucleotide polymorphisms (SNPs) weighted by the effect size obtained from large-scale discovery genome-wide association study (GWAS) is expected to improve the predictive power and performance of COVID-19 risk assessment [25], [26]. PRS using gene-panel SNPs to calculate associated risk is discussed [27].

2 Materials and Methods

2.1 Data Processing

Two workflows including a framework to align and annotate SARS-CoV-2 and a predictive model for COVID-19 patients have been developed. The first takes as input virus target sequence data from GISAID, the NCBI, and data collected in Vietnam to identify the virus genome sequence variants and provide summary statistics of these sequences. The second integrates PRSs into machine learning models from two sources: 1) GWAS with hospitalized COVID-19 patients and 2) a combination of immune biomarker variants that are associated with severity status and target data.

The data downloaded from GISAID in Vietnam consisted of 361 SARS-CoV-2 samples, shown in Figure 2. NCBI data and data collected from Vietnamese sites contained FASTA sequences of SARS-CoV-2, a summary of protein mutations, and patient metadata. Other datasets from different countries in Mekong regions were also downloaded. Nextclade (<https://docs.nextstrain.org>) is a tool that helps to identify the differences between target sequences and a reference sequence by Nextstrain to assign clades to these sequences [28]. Nextclade was used for alignment to reference SARS-CoV-2 Wuhan-1 (MN908947.3), and for detecting variants and protein mutations on these datasets.

With data from NCBI, we have gathered 322,101 samples collected in Q2 2021 (Quarter 2 of 2021) and 542,275 samples collected in Q3 2021. All samples had a length greater than 29,000 bases and number of Ns is less than 300. These samples

were also analyzed by Nextclade to find out which strains and mutations prevail among others.

COVID-19 data from 57,560 patients across 63 Vietnamese provinces (approximately as of July 20, 2021), and other data from almost 19,924 patients were downloaded from public source [29]. Age, sex, status, and other metadata of each patient were included. The patient's province of residence was a crucial parameter in the model as it represents the environment of coronavirus disease. Dataset from Whole-Genome Sequencing from Vietnam in The International Genome Sample Resource (IGSR) [30] will be used in the study. The target dataset is a cohort of 99 (of 124 samples) unrelated Vietnamese people in the project (whole-genome sequencing with 30x coverage) from Kinh ethnic group (100 KHV) on GRCh38 [31]. The χ^2 goodness-of-fit test for Hardy–Weinberg equilibrium was used on samples with related individuals, and missing genotypes were filtered. We also use the dataset from 1000 Vietnamese Genomes Project (1KVG), a source of genomic variants for Vietnamese population by sequencing the whole genome of 1008 unrelated healthy Vietnamese to a depth of at least 28x [32].

The most common method for calculating PRS is called clumping and thresholding (or pruning and thresholding), applies two filtering steps as shown in Figure 1. SNPs that weakly correlated with each other were retained. Clumps around SNPs were formed by using the linkage disequilibrium clumping procedure [33].

In PRS analyses can be characterized by the two input data sets: i) base (GWAS) data: summary statistics (e.g. betas, p-values) of genotype-phenotype associations at genetic variants and ii) target data: genotypes and phenotype(s) in individuals of the target sample [34]. We investigated the blood and lung biomarkers

incorporated into the model for 100 KHV and derived PRS for all individuals. Based on the result, we will reconstruct PRS for a larger dataset and apply several machine learning techniques to predict severity of COVID-19 patients in Vietnam.

2.2 Integrating biomarkers and target data variants into PRS computation

GWAS summary statistics of COVID-19 patient variants of Hospitalized vs. not hospitalized were downloaded. The summary was thresholded at $p = 5e - 8$ as standard in QC of GWAS. Summary statistics of Covid-19 were downloaded from open source COVID-19 HGI GWAS (<https://www.covid19hg.org>). These summary statistics are the result of a meta-analysis of 61 studies from 24 countries, and include the weights (effect sizes) and p-values of 13,498,845 variants, derived from a genotype-phenotype association study with 14,480 hospitalized patient samples and 73,191 control samples. GWAS QC excluded variants with p-value greater than .05. The PRS was derived for the training set using a pruning and thresholding method by Plink v1.9 [35]. The best model was selected based on R^2 . The PRS was calculated by the below equation of Plink.

$$PRS_j = \frac{\sum_i^N S_i * G_{ij}}{P * M_j} \quad (1)$$

The effect size of SNP i is S_i , effect allele j is G_{ij} , the ploidy of the sample is P (with humans, $P = 2$), the number of SNPs is N , the number of non-missing SNPs in sample j is M_j . In addition, individual phased genotype data of 100 KHV was from a VCF file. Standard quality controls were applied to the KHV VCF files, with missing genotype > 0.1 , Hardy-Weinberg Equilibrium $P > 1e^{-6}$, minor allele frequency $MAF < 0.01$.

2.3 Reconstructing training and testing data for machine learning models

The lack of direct PRS calculations for patients from Vietnam without a genotyping/sequencing profile posed a major challenge. Instead of directly predicting PRS using existing methods, we used a reconstruction method that applies a multivariate linear model to use the PRS calculations of an existing cohort (reference matrix *Ref* with PRS) with genotyping/sequencing to other cohorts, and showed the model can improve the prediction of severity. The model utilizes covariates captured age, gender, location... The predicted PRSs for 19,924 COVID-19 patients were then derived by a machine learning model to predict severity, and that correlated well with the measurements from clinical readouts.

$$ReconstructPRS = \sum_i^N Fraction_j^i \times PRS_{Ref}^j \quad (2)$$

ReconstructPRS is reconstructed PRS using the summation of all fraction ($Fraction_j^i$) measured by a covariate i and a sample j in the reference cohort and PRS of sample j in the cohort PRS_{Ref}^j . N is the number of covariates. SNPs relevant to COVID-19 were then ranked by probability of severity according to a COVID-19-related study from the GWAS catalog (downloaded in August

2021 from <https://www.ebi.ac.uk/gwas/>). The data show that provinces are highly correlated across datasets. Since the sequencing data of these 19,924 patients is not public, we used 100 KHV results to calculate and reconstruct PRS for the training and testing datasets. The PRS was based on GWAS on GRCh38 from [36], using the pruning and thresholding method as mentioned previously. The predictive model showed how certain non-genetic factors may impact the risk of hospitalization due to the virus. We used three machine learning models including SVM, Random Forest and Elastic Net to predict COVID-19 severity in Vietnamese patients based on PRS and other covariates such as age, sex, location, exercise, and underlying conditions. The training dataset contained 11,814 patients, with status of deceased, active and recovered. These statuses were assigned numeric values in the model. To simplify the models, this was converted to a binary class of 'Active' and 'Recovered'. Deceased patients were excluded.

3 Result

3.1 Comparative analysis for SARS-CoV-2 sequences by country using NCBI data

Figure 3 shows the percentage of SARS-CoV-2 clades in different countries for Q2 and Q3 2021. Overall, the strains in Q2 were quite diverse with common strains such as 20A, 20I (Alpha), 21A (Delta), 20B, 21F (Iota) but in the third quarter, strain 21A (Delta) predominated in most of the countries. In the case of the Delta variant, in the second quarter, a number of Asian countries such as India, Bahrain, Bangladesh, and Uzbekistan recorded the presence of this variant with a significant majority, while some European and American countries such as the US, Switzerland, Germany, this variant appeared but did not prevail. This indicates that the outbreak of the Delta variant took place first in Asian countries, then in European and American countries. Regarding mutations, our analysis results also show that the most common mutations in the third quarter are the typical mutations of the Delta variant such as S:D614G, S:P681R, S:L452R, S:T478, S:R158G. In summary, the data analyzed on NCBI show the emergence and the spread of the Delta variant and its mutations in recent times.

3.2 Comparative analysis for SARS-CoV-2 sequences in Vietnam and Thailand using GISAID data

In this part, 3,211 FASTA sequences from Thailand in GISAID have been used. We compared Vietnam and Thailand populations as they have similar genetic characteristics in other infectious diseases [37] and their data is widely available in Southeast Asia. Comparison of the number of sequences by each month shows that Thailand had a prevalence of Lineage B.1.1.7 (Alpha) in the second quarter of 2021, while Vietnam had a prevalence of Lineage B.1.167.2 (Delta) from May of 2021. In addition, lineage A.6, B.1.36.16 and AY.30, which first appeared in South East Asia, were detected mostly in Thailand (Figure 4). The analysis is consistent with the result from Chookajorn et al. 2021 [38] as the spread of the Alpha and Delta variants dominant over the region raised serious problems of the healthcare system. As an emerging epicenter of COVID-19 pandemic, Southeast Asian countries needs to take immediate collaborative actions to resolve these problems. More details of the analysis can be found in Supplementary Figure 1 and 2.

3.3 Comparative analysis for SARS-CoV-2 sequences in Vietnam between Hospitalized and Recovered patients

In GISAID datasets, there are 361 FASTA sequences of SARS-CoV 2 from Vietnam. The virus variants are divided into 10 Clades (G,GH,GK,GR,GRY,GV,L,O,S,V). More details on GISAID's clades can be found on the website (<https://www.gisaid.org/>). The result of comparison between all

clades from Vietnam shows two common variants D614G (in Spike region), P323L (in NSP12, known as ORF1a region) in almost all clades with prefix G in both groups of Hospitalized and Recovered patients. These 2 mutations overtake frequency of reference strain and can become new references. Furthermore, clade GK and GRY have more protein mutations than other clades that can be promising targets for for analyzing protein structure and designing COVID-19 vaccines or drugs (Figure 5).

Figure 5 Histogram of individuals by variants in Vietnam of Hospitalized and Recovered COVID-19 patients. Of the currently known SARS-CoV-2 clades, clade GR was the most prevalent worldwide, followed by GV and then GH.

3.4. Biomarker variants and target data variants associated with severity on Vietnamese cohorts

We formed nine gene sets associated with severity of COVID-19 patients, as introduced in Section I. Table 1 reports the immune gene sets, along with the number of genes in each set and the number of SNPs found in 100 KHV.

Table 1 Immune gene sets associated with severity of COVID-19

Geneset / FirstAuthor	Number of Genes	Number of SNPs	Reference
IL6 / Gordon	87	17653	[7]
dexamethasone / Horby	19	5322	[9]
immunesuppression / Bost	3	13612	[12]
myeloid dysfunction / Chen	4	534	[13]
lymphopenia / Diao	69	237	[14]
interferon immunopathology / Hadjadj	20	3724	[15]
T cell / Mann	200	34942	[16]
immune senescence / SchulteSchrepping	49	12406	[17]
endothelial / Grant	13	1584	[18]

Allele frequencies for SNPs of genes in each gene set were calculated for both 100 KHV and IKVG [32]. SNP allele frequencies for all gene panels were highly

correlated between sets (Pearson correlation p -value $< 2.2e^{-16}$, $R = 0.99$) (Supplementary Figure 3). 1KVG was able to detect some variants with much lower allele frequency compared with those frequencies of 100 KHV suggesting that using 1KVG (with much larger sample size) to increase the quality of variants, especially in immunogenic and drug targets used in Vietnamese people. These variants were added to the model as “causal” SNPs in the computation of PRS as illustrated in the second workflow in Figure 1.

3.5 Machine Learning models to predict severity outcomes in Vietnam

The two datasets from [29] (57,560 patients split by province and 19,924 patients with province information provided) were consistent in the distribution of patients between provinces (Pearson correlation $p = 1.2e^{-15}$, $R = 0.83$). This is an important result as location and other phenotypes were used to reconstruct PRS in the training and testing datasets for the machine learning models. The data has been divided by training 70% and testing 30% number of samples in the datasets.

With an average of 100 runs, Random Forest was the best model with $AUC = 0.81$, followed by Elastic Net with $AUC = 0.7$ and SVM with $AUC = 0.69$ in Figure 6.

4 Discussion

In this study, we used the genotype data of 99 samples from the 1000 Genomes Project, which were recruited only from Ho Chi Minh City. Although the Kinh

ethnic group is the main ethnic group in Vietnam, accounting for 86% of the country's population [39], these individuals may not represent the entire Vietnamese population. Therefore, we suggest that further investigation should be carried out with a 1000 Vietnamese Genomes Project dataset [32] recruited from 1008 unrelated individuals across the country, according to population distribution. We would expect this to increase the number of SNPs with allele frequency $> 1\%$. In this dataset, the metadata for these 1008 samples should include not only the basic health indices of BMI, blood pressure, glucose level, cholesterol level, and white blood cell count but also information about any chronic or hereditary diseases, as well as allergy factors (foods, drugs, or insects) and lifestyle factors (alcohol, cigarettes). These factors also influence the health and resilience of an individual against SARS-CoV-2 infection.

In addition to immune profiling, the prediction of COVID-19 severity in patients requires the evaluation of factors such as underlying disease [40], vaccination status, and the patient's intrinsic genetic response or adverse reactions to some drugs, especially some antibiotic therapies used for bacterial co-infection at ICU admission [41]. Allergy to β -lactam drugs like penicillin or amoxicillin, mainly caused by genetic factors from the interleukin and Human Leukocyte Antigen systems, is highly prevalent according to the National Centre of Drug Information and Adverse Drug Reactions [42]. We have studied numerous COVID-19 drugs, especially some used in Vietnam for COVID-19 outpatients and their PharmGKB IDs [43], [44] (Dexamethasone - PA449247, Methylprednisolone - PA450466, Prednisolone - PA451096, Rivaroxaban - PA165958360, Apixaban - PA166163740, and Remdesivir - PA166197141) (4109/Q-BYT issued by Vietnam Ministry of Health on August 26, 2021) that have allele frequency (for target gene variants in each drug) in 100 KHV in Supplementary Figure 4. This further investigation can be useful for the treatment benefit of Vietnamese patients when in hospital.

We initiated an effort to study the relationship between immunogenic profiling and SARS-CoV-2 infection severity by incorporating PRS based on immune gene sets. This approach is comprehensive as it incorporates PRS and immunogenic profiling of Vietnamese people. While providing novel scientific insights in Vietnam remains a major priority of this initiative study, we equally value learning from and collaborating with other countries in the Mekong regions (Cambodia, Laos, Myanmar, and Thailand) and other countries around the world. We expect to substantially contribute to the understanding of the variability of COVID-19 severity in Vietnam.

5 Conclusion

In this paper, we have investigated the SARS-CoV-2 profiling in Vietnam using various data sources and a predictive model of COVID-19 severity, using immunogenic profiling of the Vietnamese population based on investigation of SNPs in GWAS and metadata from 124 Vietnamese people (KHV) in the 1000 Genomes Project. Machine learning models showed high accuracy in predicting the hospitalization status of a very large dataset of Vietnamese COVID-19 patients. We expect to improve our model by using 1KVG dataset with both novel and known variants in order to have a better understanding of the immunogenic profiling of Vietnamese

people. This initial approach will not only be helpful in understanding susceptibility to SARS-CoV-2 infection, but could also inform how to control the disease, as well as treatment progression and recovery. By this way, we hope to make an impact on human health and healthcare systems in the areas of Vietnam affected by COVID-19 pandemic.

Ethics approval and consent to participate

In the 1KVG study, subjects provided informed consent and the study was approved by the Vinmec International Hospital Institutional Review Board with number 543/2019/Q-VMEC. All methods were carried out in accordance with the relevant guidelines and regulations (e.g. Helsinki Declaration).

Consent for publication

Not applicable

Availability of data and materials

Our wrapped pipeline for generating Polygenic Risk Score and all processed data are available for public use at <https://github.com/VMGiang/SeverityCOVID19>.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable

Author's contributions

T.H. performed bioinformatics analysis for human hosts, and developed the overall workflows. G.V. and Q.L. participated in developing the SARS-CoV-2 alignment and analysis. M.T. validated the outcomes related to biomedicine. T.T. processed 1KVG dataset. N.V. supervised the experiments and coordinated the progress of the overall project. T.H. and G.V. wrote the manuscript. Other authors revised the manuscript. All authors discussed the analysis results and contributed to bringing in innovative ideas into the manuscript.

Acknowledgements

We would like to express our deepest attitude to colleagues at Vingroup Big Data Institute to help us with comments for this paper. We are also thankful for using 1000 Vietnamese Genomes Project by Vingroup [32] for SNP validation.

Authors' information

Affiliations

Center for Biomedical Informatics, Vingroup Big Data institute, 458 Minh Khai, Hai Ba Trung, Hanoi, Vietnam
Tham H. Hoang, Giang M. Vu, Mai H. Tran, Trang T.H. Tran, Nam S. Vo
College of Engineering and Computer Science, VinUniversity
Nam S. Vo
Hanoi University of Civil Engineering, 55 Giai Phong, Hai Ba Trung, Hanoi, Vietnam
Quang D. Le

Corresponding author

Correspondence to Tham H. Hoang

Author details

¹Center for Biomedical Informatics, Vingroup Big Data institute, 458 Minh Khai, Hai Ba Trung, Hanoi, Vietnam.

²Hanoi University of Civil Engineering, 55 Giai Phong, Hai Ba Trung, Hanoi, Vietnam. ³College of Engineering and Computer Science, VinUniversity, 55 Giai Phong, Hai Ba Trung, Hanoi, Vietnam.

References

1. Zaim, S., Chong, J.H., Sankaranarayanan, V., Harky, A.: Covid-19 and multiorgan response. *Current problems in cardiology* **45**(8), 100618 (2020)
2. Dong, X., Cao, Y.-y., Lu, X.-x., Zhang, J.-j., Du, H., Yan, Y.-q., Akdis, C.A., Gao, Y.-d.: Eleven faces of coronavirus disease 2019. *Allergy* **75**(7), 1699–1709 (2020)
3. Nguyen, T.T., Pham, T.N., Van, T.D., Nguyen, T.T., Nguyen, D.T.N., Le, H.N.M., Eden, J.-S., Rockett, R.J., Nguyen, T.T.H., Vu, B.T.N., *et al.*: Genetic diversity of sars-cov-2 and clinical, epidemiological characteristics of covid-19 patients in hanoi, vietnam. *PloS one* **15**(11), 0242537 (2020)
4. Greaney, A.J., Starr, T.N., Gilchuk, P., Zost, S.J., Binshtein, E., Loes, A.N., Hilton, S.K., Huddleston, J., Eguia, R., Crawford, K.H., *et al.*: Complete mapping of mutations to the sars-cov-2 spike receptor-binding domain that escape antibody recognition. *Cell host & microbe* **29**(1), 44–57 (2021)
5. O'Toole, Á., Hill, V., Pybus, O.G., Watts, A., Bogoch, I.I., Khan, K., Messina, J.P., COVID, T., *et al.*: Tracking the international spread of sars-cov-2 lineages b. 1.1. 7 and b. 1.351/501y-v2. *Wellcome Open Research* **6** (2021)

6. Voloch, C.M., da Silva Francisco Jr, R., de Almeida, L.G., Cardoso, C.C., Brustolini, O.J., Gerber, A.L., Guimarães, A.P.d.C., Mariani, D., da Costa, R.M., Ferreira Jr, O.C., et al.: Genomic characterization of a novel sars-cov-2 lineage from rio de janeiro, brazil. *Journal of virology* **95**(10), 00119–21 (2021)
7. Gordon, A.C., Mouncey, P.R., Al-Beidh, F., Rowan, K.M., Nichol, A.D., Arabi, Y.M., Annane, D., Beane, A., Berry, L.R., Bhimani, Z., et al.: Interleukin-6 receptor antagonists in critically ill patients with covid-19. *The New England journal of medicine* (2021)
8. Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A.D., Rawlik, K., Pasko, D., Walker, S., Parkinson, N., Fourman, M.H., Russell, C.D., et al.: Genetic mechanisms of critical illness in covid-19. *Nature* **591**(7848), 92–98 (2021)
9. Horby, P.W., Campbell, M., Staplin, N., Spata, E., Emberson, J.R., Pessoa-Amorim, G., Peto, L., Brightling, C.E., Sarkar, R., Thomas, K., et al.: Tocilizumab in patients admitted to hospital with covid-19 (recovery): preliminary results of a randomised, controlled, open-label, platform trial. *Medrxiv* (2021)
10. Kalil, A.C., Patterson, T.F., Mehta, A.K., Tomashek, K.M., Wolfe, C.R., Ghazaryan, V., Marconi, V.C., Ruiz-Palacios, G.M., Hsieh, L., Kline, S., et al.: Baricitinib plus remdesivir for hospitalized adults with covid-19. *New England Journal of Medicine* **384**(9), 795–807 (2021)
11. Ahern, D.J., Ai, Z., Ainsworth, M., Allan, C., Allcock, A., Ansari, A., Arancibia-Carcamo, C.V., Aschenbrenner, D., Attar, M., Baillie, J.K., et al.: A blood atlas of covid-19 defines hallmarks of disease severity and specificity. *medRxiv* (2021)
12. Bost, P., De Sanctis, F., Canè, S., Ugel, S., Donadello, K., Castellucci, M., Eyal, D., Fiore, A., Anselmi, C., Barouni, R.M., et al.: Deciphering the state of immune silence in fatal covid-19 patients. *Nature communications* **12**(1), 1–15 (2021)
13. Chen, Z., Wherry, E.J.: T cell responses in patients with covid-19. *Nature Reviews Immunology* **20**(9), 529–536 (2020)
14. Diao, B., Wen, K., Chen, J., Liu, Y., Yuan, Z., Han, C., Chen, J., Pan, Y., Chen, L., Dan, Y., et al.: Diagnosis of acute respiratory syndrome coronavirus 2 infection by detection of nucleocapsid protein. *MedRxiv* (2020)
15. Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., Péré, H., Charbit, B., Bondet, V., Chenevier-Gobeaux, C., et al.: Impaired type i interferon activity and inflammatory responses in severe covid-19 patients. *Science* **369**(6504), 718–724 (2020)
16. Mann, D.M., Chen, J., Chunara, R., Testa, P.A., Nov, O.: Covid-19 transforms health care through telemedicine: evidence from the field. *Journal of the American Medical Informatics Association* **27**(7), 1132–1135 (2020)
17. Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., et al.: Severe covid-19 is marked by a dysregulated myeloid cell compartment. *Cell* **182**(6), 1419–1440 (2020)
18. Grant, R.A., Morales-Nebreda, L., Markov, N.S., Swaminathan, S., Querrey, M., Guzman, E.R., Abbott, D.A., Donnelly, H.K., Donayre, A., Goldberg, I.A., et al.: Circuits between infected macrophages and t cells in sars-cov-2 pneumonia. *Nature* **590**(7847), 635–641 (2021)
19. Mehta, P., McAuley, D.F., Brown, M., Sanchez, E., Tattersall, R.S., Manson, J.J.: Covid-19: consider cytokine storm syndromes and immunosuppression. *The lancet* **395**(10229), 1033–1034 (2020)
20. Zhao, B., Ni, C., Gao, R., Wang, Y., Yang, L., Wei, J., Lv, T., Liang, J., Zhang, Q., Xu, W., et al.: Recapitulation of sars-cov-2 infection and cholangiocyte damage with human liver ductal organoids. *Protein & cell* **11**(10), 771–775 (2020)
21. Buturovic, L., Zheng, H., Tang, B., Lai, K., Kuan, W.S., Gillett, M., Santram, R., Shojaei, M., Almansa, R., Nieto, J.A., et al.: A 6-mrna host response whole-blood classifier trained using patients with non-covid-19 viral infections accurately predicts severity of covid-19. *medRxiv* (2020)
22. Ganna, A., Initiative, C.-H.G., et al.: Mapping the human genetic architecture of covid-19 by worldwide meta-analysis. *MedRxiv* (2021)
23. Downes, D.J., Cross, A.R., Hua, P., Roberts, N., Schwesinger, R., Cutler, A.J., Munis, A.M., Brown, J., Mielczarek, O., de Andrea, C.E., et al.: Identification of lztfl1 as a candidate effector gene at a covid-19 risk locus. *Nature Genetics*, 1–10 (2021)
24. Izcovich, A., Ragusa, M.A., Tortosa, F., Lavena Marzio, M.A., Agnoletti, C., Bengolea, A., Ceirano, A., Espinosa, F., Saavedra, E., Sanguine, V., et al.: Prognostic factors for severity and mortality in patients infected with covid-19: A systematic review. *PloS one* **15**(11), 0241955 (2020)
25. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., Visscher, P.M.: Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics* **14**(7), 507–515 (2013)
26. Aragam, K.G., Dobbyn, A., Judy, R., Chaffin, M., Chaudhary, K., Hindy, G., Cagan, A., Finneran, P., Weng, L.-C., Loos, R.J., et al.: Limitations of contemporary guidelines for managing patients at high genetic risk of coronary artery disease. *Journal of the American College of Cardiology* **75**(22), 2769–2780 (2020)
27. Levine, M.E., Crimmins, E.M., Prescott, C.A., Phillips, D., Arpawong, T.E., Lee, J.: A polygenic risk score associated with measures of depressive symptoms among older adults. *Biodemography and social biology* **60**(2), 199–211 (2014)
28. Aksamentov, I., Roemer, C., Hodcroft, E.B., Neher, R.A.: Nextclade: clade assignment, mutation calling and
29. OpenDevelopmentMekong. <https://opendevelopmentmekong.net/>
30. Consortium, .G.P., et al.: A global reference for human genetic variation. *Nature* **526**(7571), 68 (2015)
31. Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P., Consortium, .G.P.: Alignment of 1000 genomes project reads to reference assembly grch38. *Gigascience* **6**(7), 038 (2017)
32. IKVG Data Portal. <https://genome.vinbigdata.org/>
33. Privé, F., Aschard, H., Blum, M.G.: Efficient implementation of penalized regression for genetic risk prediction. *Genetics* **212**(1), 65–74 (2019)
34. Choi, S.W., Mak, T.S.H., O'Reilly, P.F.: A guide to performing polygenic risk score analyses. *BioRxiv*, 416545 (2018)
35. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P.,

- De Bakker, P.I., Daly, M.J., *et al.*: Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**(3), 559–575 (2007)
36. Initiative, C.-H.G., *et al.*: 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* **28**(6), 715 (2020)
 37. Yen, H.-L., Webster, R.G.: Pandemic influenza as a current threat. vaccines for pandemic influenza, 3–24 (2009)
 38. Chookajorn, T., Kochakarn, T., Wilasang, C., Kotanan, N., Modchang, C.: Southeast asia is an emerging hotspot for covid-19. *Nature Medicine* **27**(9), 1495–1496 (2021)
 39. G.S.O.Vietnam. <https://www.gso.gov.vn/dan-so/thong-cao-bao-chi/>. Online
 40. Wolff, D., Nee, S., Hickey, N.S., Marschollek, M.: Risk factors for covid-19 severity and fatality: a structured literature review. *Infection* **49**(1), 15–28 (2021)
 41. Contou, D., Cally, R., Sarfati, F., Desaint, P., Fraissé, M., Plantefève, G.: Causes and timing of death in critically ill covid-19 patients. *Critical Care* **25**(1), 1–4 (2021)
 42. The National Centre of Drug information and Adverse drug reactions monitoring. <http://canhgiacduoc.org.vn/>. [Online]
 43. Whirl-Carrillo, M., McDonagh, E.M., Hebert, J., Gong, L., Sangkuhl, K., Thorn, C., Altman, R.B., Klein, T.E.: Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics* **92**(4), 414–417 (2012)
 44. Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C.F., Whaley, R., Klein, T.E.: An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics* **110**(3), 563–572 (2021)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [BMCI Infectious diseases supplementary.pdf](#)