BMC Infectious Diseases

RESEARCH Open Access

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

Tham H. Hoang[1*], Giang M. Vu[1], Mai H. Tran[1], Trang T. H. Tran[1], Quang D. Le[2], Khanh V. Tran[3], Tue T. Nguyen[3], Lan T. N. Nguyen[3], Thinh H. Tran[3], Van T. Ta[3] and Nam S. Vo[1,4]

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

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

*Correspondence: v.thamhh@vinbigdata.org

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

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 dominant strain. 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) [7–11]. 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 [12–17]. 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 [19–21].

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].

Materials and methods

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 Fig. 1. 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 identifies 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.

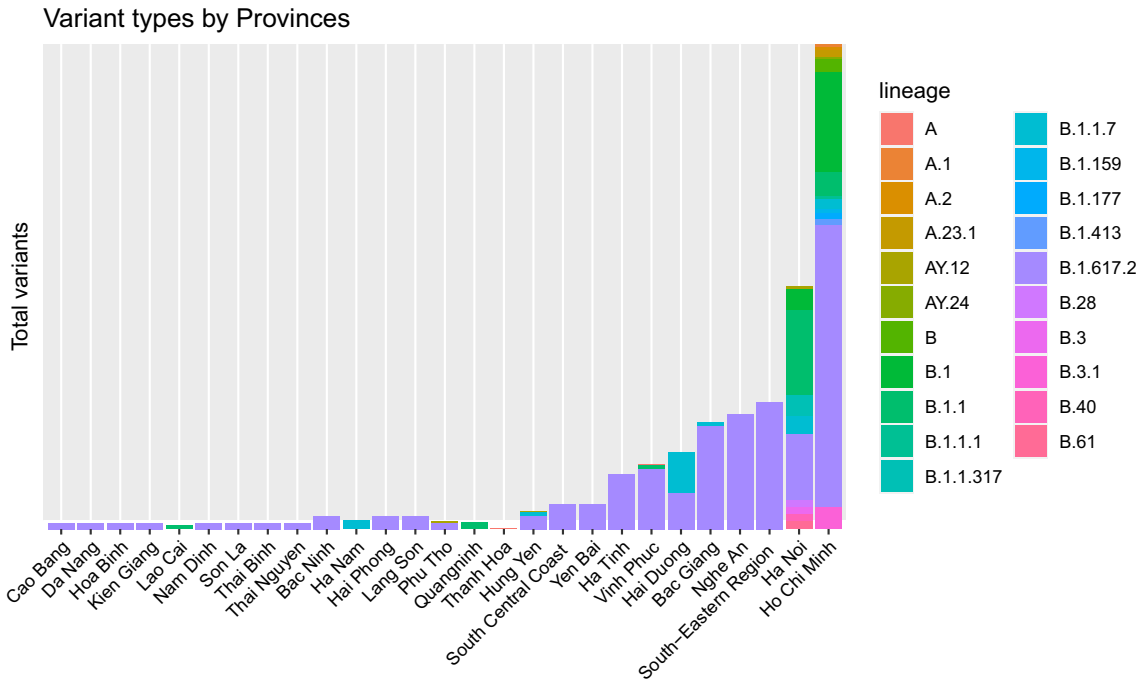

Fig. 1 Clade Pango lineage of 361 SARS-CoV-2 samples collected in Vietnam. The Delta variant (B.1.617.2) was the most prevalent variant as of GISAID data collection

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 30× 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 28× [32].

The most common method for calculating PRS is called clumping and thresholding (or pruning and thresholding), applies two filtering steps as shown in Fig. 2. 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.

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 0.05. 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.

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

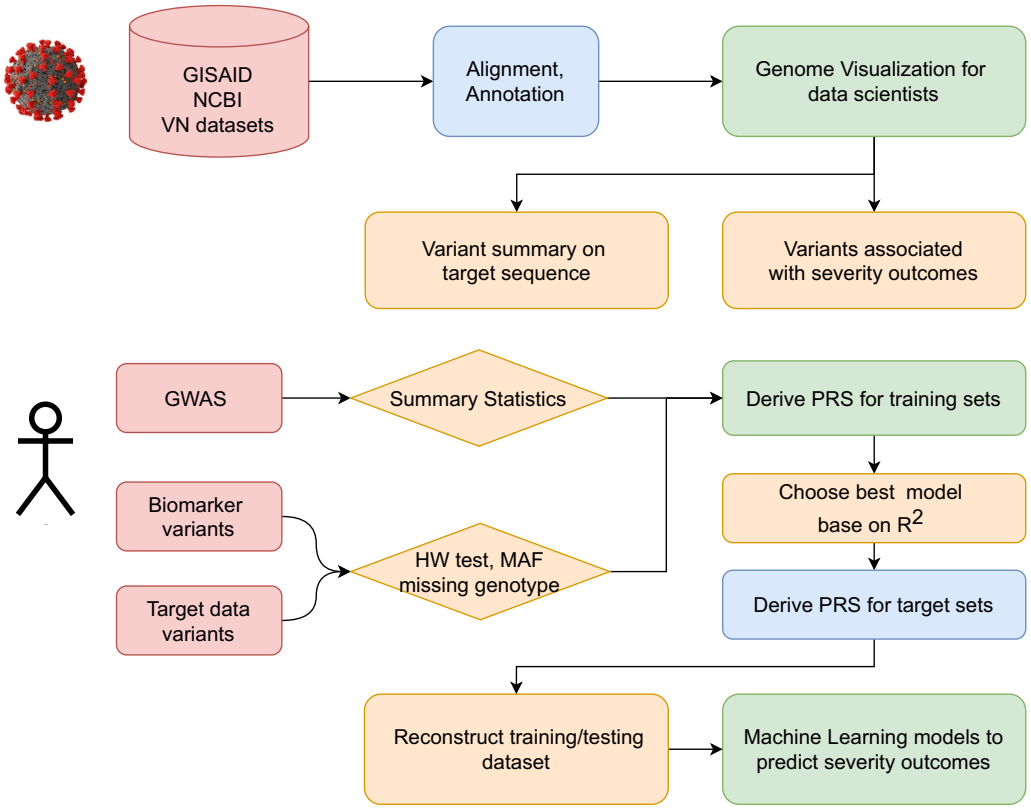

Fig. 2 Two workflows have been developed. The first takes as input virus target sequence data from GISAID, the NCBI, and data collected in Vietnam (VN dataset) to identify the virus genome sequence variants and provide summary statistics of these sequences. The second integrates PRSs from two sources including GWAS and a combination of immune biomarker variants associated with the severity of COVID-19 patients

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 $MAF < 0.01$.

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.

$$ReconstructPRS = \sum_{i}^{N} Fraction_j^i \times PRS_{Ref}^j \tag{2}$$

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.

Result

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.

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

In this part, 3211 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 (Fig. 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

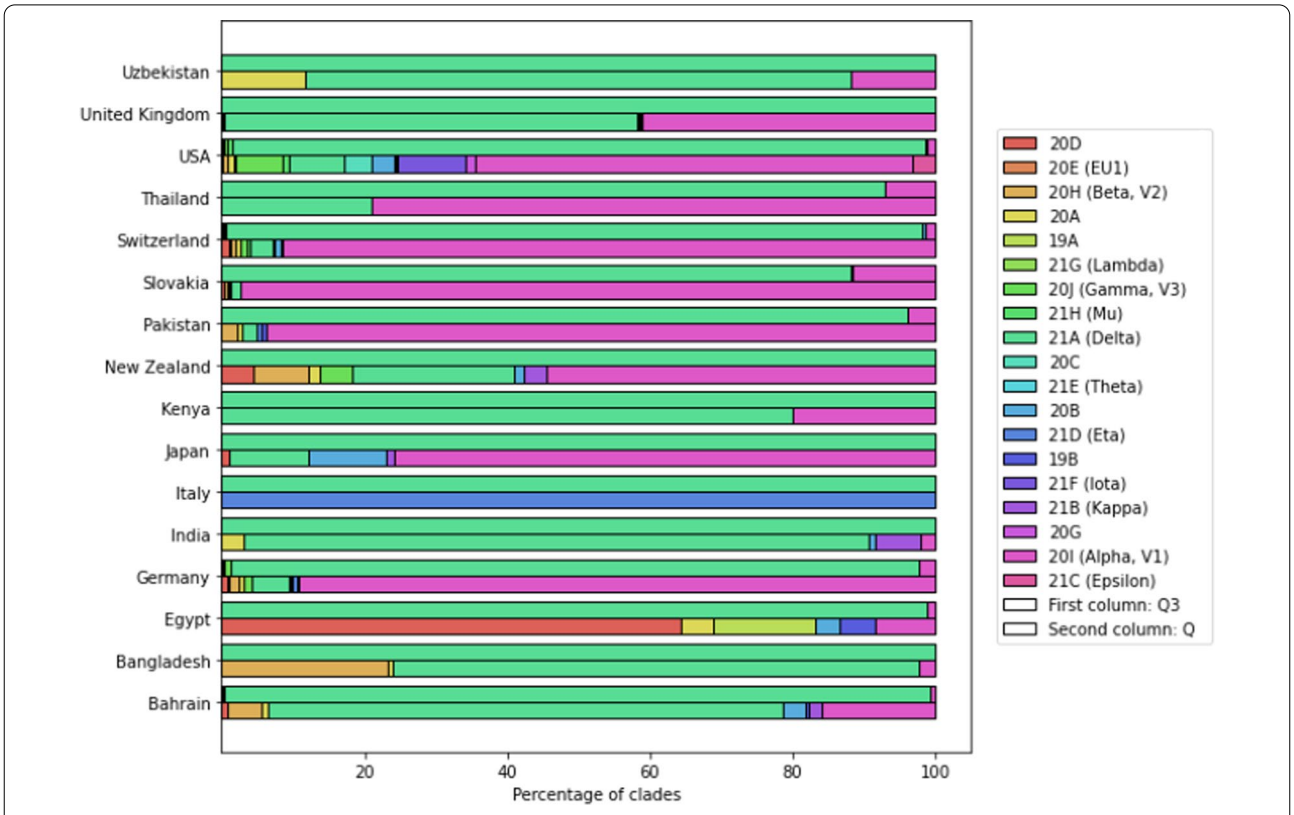

Fig. 3 Sequence analysis of SARS-CoV-2 among countries

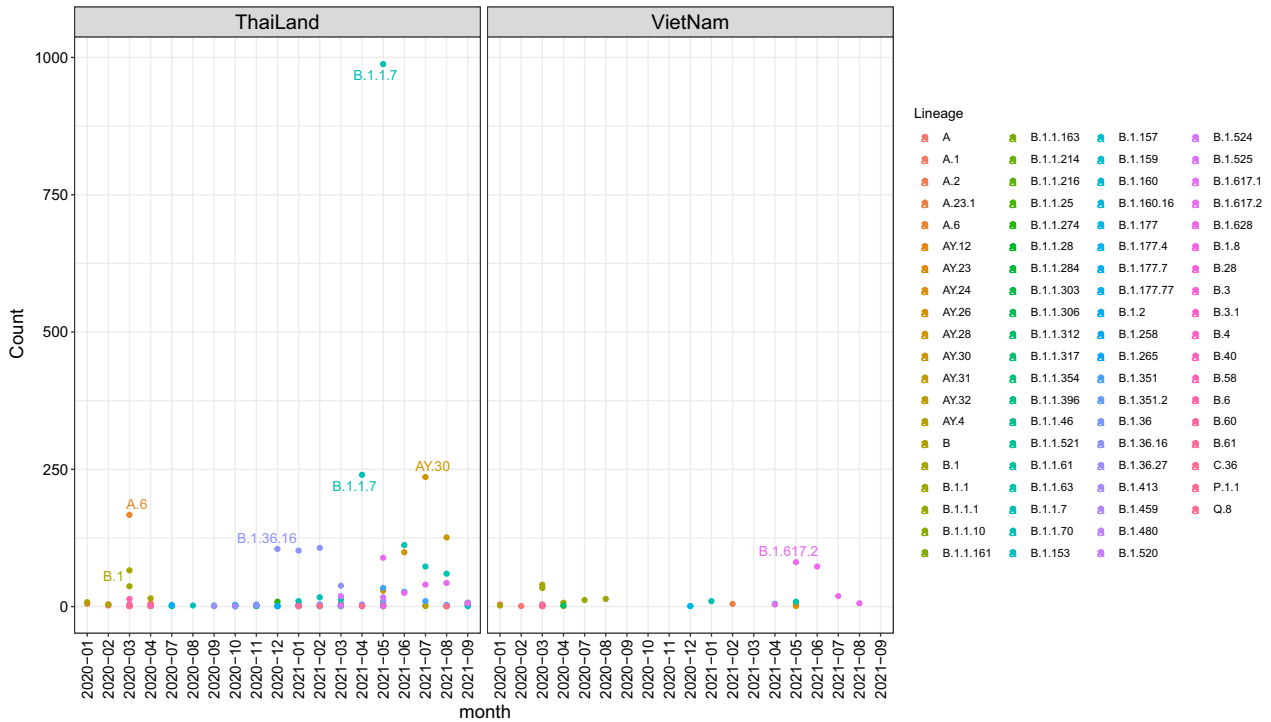

Fig. 4 Number of SARS-CoV-2 sequences with different clades from Vietnam and Thailand

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 Additional file 1: Figs. S1 and S2.

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 dominant strain. 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 (Fig. 5).

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 "Introduction" section. 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.

Allele frequencies for SNPs of genes in each gene set were calculated for both 100 KHV and 1KVG [32]. SNP allele frequencies for all gene panels were highly correlated between sets (Pearson correlation $p - value < 2.2e^{-16}, R = 0.99$) (Additional file 1: Fig. S3). 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 Fig. 1.

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

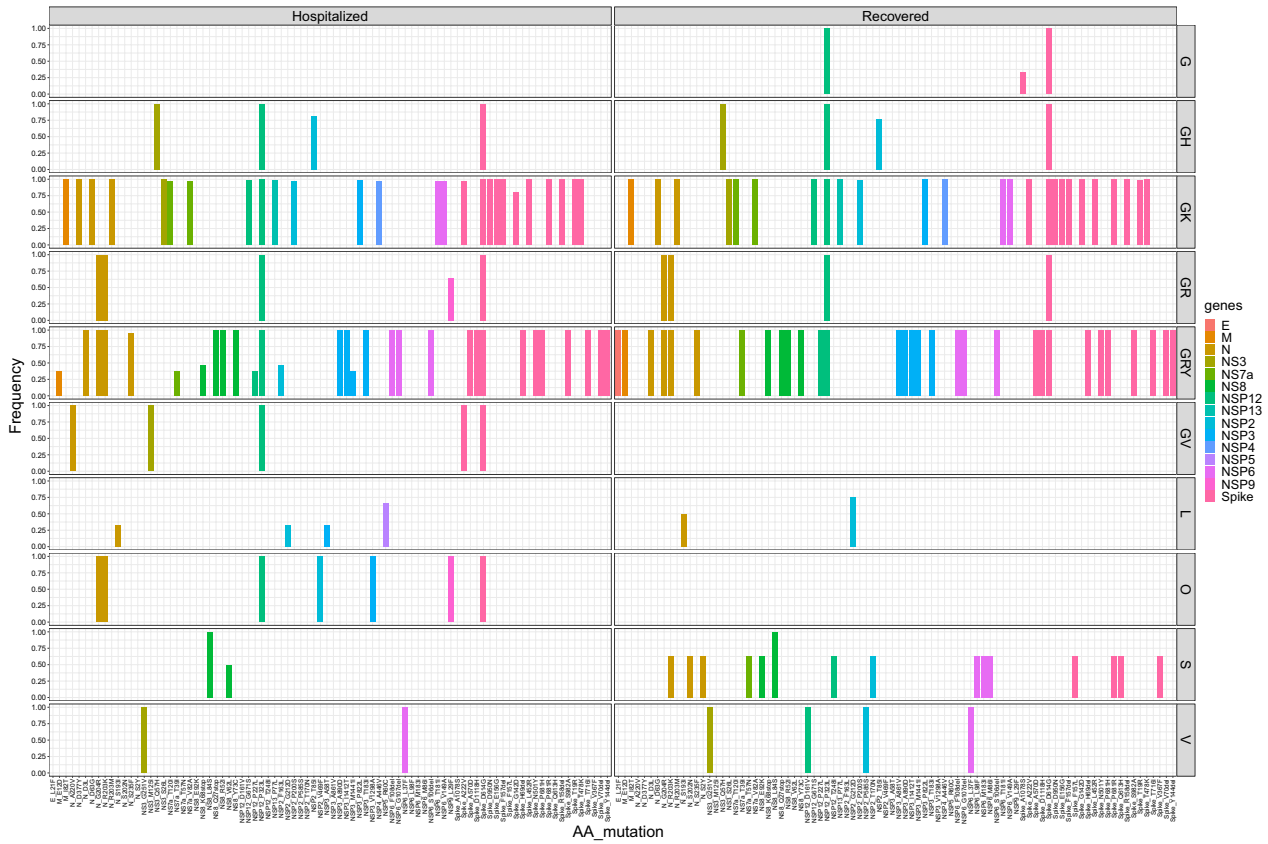

Fig. 5 Histogram of individuals by variants in Vietnam of Hospitalized and Recovered COVID-19 patients. Of the currently known SARS-CoV-2 clades, clade GR was the most prevalent worldwide, followed by GV and then GH

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

Geneset/first author	Number of genes	Number of SNPs	References
IL6/Gordon	87	17,653	[7]
Dexamethasone/Horby	19	5322	[9]
Immunesuppression/Bost	3	13,612	[12]
Myeloid dysfunction/Chen	4	534	[13]
Lymphopenia/Diao	69	237	[14]
Interferon immunopathology/Hadjadj	20	3724	[15]
Tcell/Mann	200	34,942	[16]
Immune senescence/SchulteSchrepping	49	12,406	[17]
Endothelial/Grant	13	1584	[18]

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 Fig. 6.

Discussion

Statistical analysis of the SARS-CoV-2 sequences obtained from the NCBI indicates that the predominant virus variants at a period of time may vary between countries and regions of the world. If the severity of COVID-19 is related to the variation of

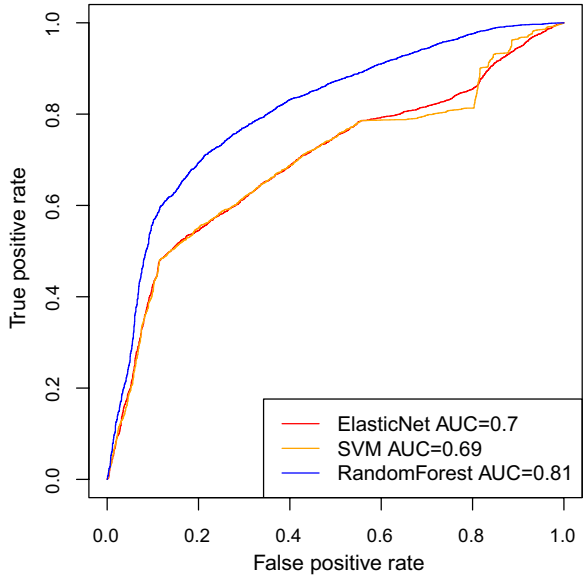

Fig. 6 High training accuracy varied by machine learning method to predict severity status of COVID-19 patients based on PRS and other covariates

the virus and the genetics of a population, then this association should be analyzed by country and ethnic group. On the other hand, the statistics for two time periods also show that the dominant strain in a country can change over time, the new dominant variant can replace existing variants. The dynamic change of SARS-CoV-2 variants requires prediction of COVID-19 severity in patients to be performed regularly to stay up to date with current prevailing variants.

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/QĐ-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 Additional file 1: Fig. S4. 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 (Additional files 2, 3, 4, 5, 6, 7).

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.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-022-07415-1.

Additional file 1: Fig. S1. Percentage of Clades by month in Vietnam and Thailand. **Fig. S2.** Percentage by Lineage by month in Vietnam and Thailand. **Fig. S3.** Scatter plots of allele frequency from datasets: one from WGS of Vietnamese people in the 1000 Genomes Project with high coverage, and the other from WGS of 1000 Vietnamese people [1] with nine immune gene sets associated with severity of COVID-19 in Vietnam. **Fig. S4.** Scatter plots of allele frequency from datasets: one from WGS of Vietnamese people in the 1000 Genomes Project with high coverage, and the other from WGS of 1000 Vietnamese people [1] with 6 target gene sets of COVID-19 drugs used in Vietnam.

Additional file 2. Table S1. SarsCov2Q2Q3: The number of sequences included in the Q3 and Q2 columns for each country.

Additional file 3. Table S2. IL6 Gordon: The overlapping SNPs between IL6 gene set (Gordon et al.) with GWAS significant set (e.g., p-value < 5e-8).

Additional file 4. Table S3. Myeloid dysfunc Chen: The overlapping SNPs between myeloid dysfunction gene set (Chen et al.) with GWAS significant set.

Additional file 5. Table S4. Interferon Hadjadj: The overlapping SNPs between Interferon gene set (Hadjadj et al.) with GWAS significant set.

Additional file 6. Table S5. Tcell Mann: The overlapping SNPs between Tcell gene set (Mann et al.) with GWAS significant set.

Additional file 7. Table S6. All genset: The overlapping SNPs between all gene set with GWAS significant set.

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 thankful for using 1000 Vietnamese Genomes Project by Vingroup [32] for the validation of SNPs. We are also thankful for using the VN dataset which is SARS-CoV-2 RNA-Seq samples from Hanoi Medical University for the validation of SARS-CoV-2 framework. This work was partly supported by the Vietnam Ministry of Science and Technology, grant number 1539/ QD-BKHCN.

Author contributions

T.H.H. performed bioinformatics analysis for human hosts, and developed the overall workflows. G.M.V. and Q.D.L. participated in developing the SARS-CoV-2 alignment and analysis. M.H.T. validated the outcomes related to biomedicine. T.T.H.T. processed 1KVG dataset. T.T.N., L.N.T.N., K.V.T. and T.H.T. contributed to collect the Vietnam data and perform sequencing. V.T.T. supervised the experiments and coordinated the progress of the project in Hanoi Medical University. N.V. supervised the experiments and coordinated the progress of the overall project. T.H.H. and G.M.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. All authors read and approved the final manuscript.

Funding

This work was partly supported by the Vietnam Ministry of Science and Technology, Grant Number 1539/ QD-BKHCN.

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.

Declarations

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/QĐ-VMEC. out in accordance with the relevant guidelines and regulations (e.g. Helsinki Declaration).

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Center for Biomedical Informatics, Vingroup Big Data Institute, 458 Minh Khai Street, Hai Ba Trung, Hanoi, Vietnam. [2]Hanoi University of Civil Engineering, 55 Giai Phong Street, Hai Ba Trung, Hanoi, Vietnam. [3]Hanoi Medical University, 1 Ton That Tung Street, Dong Da, Hanoi, Vietnam. [4]College of Engineering and Computer Science, VinUniversity, Vinhomes Ocean Park, Gia Lam, Hanoi, Vietnam.

Received: 18 November 2021 Accepted: 20 April 2022
Published online: 19 June 2022

References

1. Zaim S, Chong JH, Sankaranarayanan V, Harky A. COVID-19 and multiorgan response. Curr Probl Cardiol. 2020;45(8): 100618.
2. Dong X, Cao Y-y, Lu X-x, Zhang J-j, Du H, Yan Y-q, Akdis CA, Gao Y-d. Eleven faces of coronavirus disease 2019. Allergy. 2020;75(7):1699–709.
3. Nguyen TT, Pham TN, Van TD, Nguyen TT, Nguyen DTN, Le HNM, Eden J-S, Rockett RJ, Nguyen TTH, Vu BTN, et al. Genetic diversity of SARS-COV-2 and clinical, epidemiological characteristics of COVID-19 patients in Hanoi, Vietnam. PloS One. 2020;15(11):0242537.
4. Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, Hilton SK, Huddleston J, Eguia R, Crawford KH, et al. Complete mapping of mutations to the SARS-COV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe. 2021;29(1):44–57.
5. O'Toole ´A, Hill V, Pybus OG, Watts A, Bogoch II, Khan K, Messina JP, COVID T, UK G, et al. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. Wellcome open research 2021;6.
6. Voloch CM, da Silva Francisco R Jr, de Almeida LG, Cardoso CC, Brustolini OJ, Gerber AL, Guimarães APdC, Mariani D, da Costa RM, Ferreira OC Jr, et al. Genomic characterization of a novel SARS-COV-2 lineage from Rio de Janeiro, Brazil. J Virol. 2021;95(10):00119–21.
7. Gordon AC, Mouncey PR, Al-Beidh F, Rowan KM, Nichol AD, Arabi YM, Annane D, Beane A, Berry LR, Bhimani Z, et al. Interleukin-6 receptor antagonists in critically ill patients with COVID-19. N Engl J Med. 2021.
8. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, Walker S, Parkinson N, Fourman MH, Russell CD, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021;591(7848):92–8.
9. Horby PW, Campbell M, Staplin N, Spata E, Emberson JR, Pessoa-Amorim G, Peto L, Brightling CE, 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 AC, Patterson TF, Mehta AK, Tomashek KM, Wolfe CR, Ghazaryan V, Marconi VC, Ruiz-Palacios GM, Hsieh L, Kline S, et al. Baricitinib plus Remdesivir for hospitalized adults with COVID-19. N Engl J Med. 2021;384(9):795–807.
11. Ahern DJ, Ai Z, Ainsworth M, Allan C, Allcock A, Ansari A, Arancibia-Carcamo CV, Aschenbrenner D, Attar M, Baillie JK, 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 RM, et al. Deciphering the state of immune silence in fatal COVID-19 patients. Nat Commun. 2021;12(1):1–15.
13. Chen Z, Wherry EJ. T cell responses in patients with COVID-19. Nat Rev Immunol. 2020;20(9):529–36.

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. 2020;369(6504):718–24.
16. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inf Assoc. 2020;27(7):1132–5.
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. 2020;182(6):1419–40.
18. Grant RA, Morales-Nebreda L, Markov NS, Swaminathan S, Querrey M, Guzman ER, Abbott DA, Donnelly HK, Donayre A, Goldberg IA, et al. Circuits between infected macrophages and t cells in SARS-COV-2 pneumonia. Nature. 2021;590(7847):635–41.
19. Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ. COVID-19: consider cytokine storm syndromes and immunosuppression. Lancet. 2020;395(10229):1033–4.
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. 2020;11(10):771–5.
21. Buturovic L, Zheng H, Tang B, Lai K, Kuan WS, Gillett M, Santram R, Shojaei M, Almansa R, Nieto JA, 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. 2020.
22. Ganna A, Initiative C-HG, et al. Mapping the human genetic architecture of covid-19 by worldwide meta-analysis. MedRxiv, 2021.
23. Downes DJ, Cross AR, Hua P, Roberts N, Schwessinger R, Cutler AJ, Munis AM, Brown J, Mielczarek O, de Andrea CE, et al. Identification of lztfl1 as a candidate effector gene at a covid-19 risk locus. Nat Genet. 2021;53(11):1606–15.
24. Izcovich A, Ragusa MA, Tortosa F, Lavena Marzio MA, 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. 2020;15(11):0241955.
25. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPS. Nat Rev Genet. 2013;14(7):507–15.
26. Aragam KG, Dobbyn A, Judy R, Chaffin M, Chaudhary K, Hindy G, Cagan A, Finneran P, Weng L-C, Loos RJ, et al. Limitations of contemporary guidelines for managing patients at high genetic risk of coronary artery disease. J Am Coll Cardiol. 2020;75(22):2769–80.
27. Levine ME, Crimmins EM, Prescott CA, Phillips D, Arpawong TE, Lee J. A polygenic risk score associated with measures of depressive symptoms among older adults. Biodemogr Soc Biol. 2014;60(2):199–211.
28. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021;6(67):3773.
29. OpenDevelopmentMekong. https://opendevelopmentmekong.net/
30. Consortium GP, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68.
31. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, Consortium GP. Alignment of 1000 genomes project reads to reference assembly grch38. Gigascience. 2017;6(7):038.
32. 1KVG Data Portal. https://genome.vinbigdata.org/
33. Privé F, Aschard H, Blum MG. Efficient implementation of penalized regression for genetic risk prediction. Genetics. 2019;212(1):65–74.
34. Choi SW, Mak TSH, O'Reilly PF. A guide to performing polygenic risk score analyses. BioRxiv, 2018;416545.
35. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
36. Initiative C-HG, 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. Eur J Hum Genet. 2020;28(6):715.
37. Yen H-L, Webster RG. Pandemic influenza as a current threat. vaccines for pandemic influenza, 2009:3–24.
38. Chookajorn T, Kochakarn T, Wilasang C, Kotanan N, Modchang C. Southeast Asia is an emerging hotspot for COVID-19. Nat Med. 2021;27(9):1495–6.
39. G.S.O.Vietnam. https://www.gso.gov.vn/dan-so/thong-cao-bao-chi/.
40. Wolff D, Nee S, Hickey NS, Marschollek M. Risk factors for COVID-19 severity and fatality: a structured literature review. Infection. 2021;49(1):15–28.
41. Contou D, Cally R, Sarfati F, Desaint P, Fraissé M, Plantefève G. Causes and timing of death in critically ill COVID-19 patients. Crit Care. 2021;25(1):1–4.
42. The National Centre of Drug information and Adverse drug reactions monitoring. http://canhgiacduoc.org.vn/.
43. Whirl-Carrillo M, McDonagh EM, Hebert J, Gong L, Sangkuhl K, Thorn C, Altman RB, Klein TE. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92(4):414–7.
44. Whirl-Carrillo M, Huddart R, Gong L, Sangkuhl K, Thorn CF, Whaley R, Klein TE. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2021;110(3):563–72.

Publisher's Note