RESEARCH ARTICLE

REVISED Prediction of the effects of the top 10 nonsynonymous variants from 30229 SARS-CoV-2 strains on their proteins [version 2; peer review: 2 approved]

Previously titled: Prediction of the Effects of Nonsynonymous Variants on SARS-CoV-2 Proteins

Boon Zhan Sia, Wan Xin Boon, Yoke Yee Yap, Shalini Kumar, Chong Han Ng

Faculty of Information Science and Technology, Multimedia University, Bukit Beruang, Melaka, 75450, Malaysia

v2 **First published:** 06 Jan 2022, **11**:9
https://doi.org/10.12688/f1000research.72904.1

Latest published: 18 May 2022, **11**:9
https://doi.org/10.12688/f1000research.72904.2

Abstract

Background: SARS-CoV-2 virus is a highly transmissible pathogen that causes COVID-19. The outbreak originated in Wuhan, China in December 2019. A number of nonsynonymous mutations located at different SARS-CoV-2 proteins have been reported by multiple studies. However, there are limited computational studies on the biological impacts of these mutations on the structure and function of the proteins.

Methods: In our study nonsynonymous mutations of the SARS-CoV-2 genome and their frequencies were identified from 30,229 sequences. Subsequently, the effects of the top 10 highest frequency nonsynonymous mutations of different SARS-CoV-2 proteins were analyzed using bioinformatics tools including co-mutation analysis, prediction of the protein structure stability and flexibility analysis, and prediction of the protein functions.

Results: A total of 231 nonsynonymous mutations were identified from 30,229 SARS-CoV-2 genome sequences. The top 10 nonsynonymous mutations affecting nine amino acid residues were ORF1a nsp5 P108S, ORF1b nsp12 P323L and A423V, S protein N501Y and D614G, ORF3a Q57H, N protein P151L, R203K and G204R. Many nonsynonymous mutations showed a high concurrence ratio, suggesting these mutations may evolve together and interact functionally. Our result showed that ORF1a nsp5 P108S, ORF3a Q57H and N protein P151L mutations may be deleterious to the function of SARS-CoV-2 proteins. In addition, ORF1a nsp5 P108S and S protein D614G may destabilize the protein structures while S protein D614G may have a more open conformation compared to the wild type.

Conclusion: The biological consequences of these nonsynonymous mutations of SARS-CoV-2 proteins should be further validated by in vivo and in vitro experimental studies in the future.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 18 May 2022	✓ view	✓ view
	↑	↑
version 1 06 Jan 2022	? view	? view

1. **Ujwal Ranjit Bagal**, Centers for Disease Control and Prevention, Atlanta, USA

   Vishal Nayak, Centers for Disease Control and Prevention, Atlanta, USA
   Frederick National Laboratory for Cancer Research, Frederick, USA

2. **In-Hee Lee**, Boston Children's Hospital, Boston, USA

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
SARS-CoV-2, nonsynonymous mutation, co-mutation, COVID-19

This article is included in the Cell & Molecular Biology gateway.

This article is included in the Research Synergy Foundation gateway.

Corresponding author: Chong Han Ng (chng@mmu.edu.my)

Author roles: Sia BZ: Investigation, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Boon WX**: Investigation, Methodology; **Yap YY**: Investigation, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kumar S**: Investigation, Methodology; **Ng CH**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research is supported by Multimedia University, Malaysia, IRFund 2.0 (grant number MMUI/210119 awarded to Chong Han, Ng). The funder has no role in study design, data analysis, decision to publish or manuscript preparation.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Sia BZ, Boon WX, Yap YY *et al.* **Prediction of the effects of the top 10 nonsynonymous variants from 30229 SARS-CoV-2 strains on their proteins [version 2; peer review: 2 approved]** F1000Research 2022, **11**:9 https://doi.org/10.12688/f1000research.72904.2

First published: 06 Jan 2022, **11**:9 https://doi.org/10.12688/f1000research.72904.1

REVISED **Amendments from Version 1**

The title of the paper and the abstract have been revised to improve the clarity of the writing. The introduction section with the examples of the effect of nonsynonymous mutations affecting the function of the viral protein has been included in the last paragraph. The information about the geographical distribution of the SARS-CoV-2 dataset with the date range has been added as extended data. The detail of the high coverage parameter has been included in the methods section. Additional information on GTF and VCF files are added in the methods. Both references for the scoring method of SIFT4G and PROVEAN have been included. The GTF and VCF files for SIFT4G analysis are deposited in Figshare. The frequency percentage of the top 10 nonsynonymous mutations in the primary lineages associated with the past and present variant of concern has been included in Table 3. Figure 2 has been revised with the additional information on amino acid changes. Co-mutation analysis in the discussion section has been expanded. Additional paragraph in the discussion section has been added to discuss the impact of our study and to explain why the study of some of the identified mutations remain relevant for the newer variants. Some citations have been deleted, added or updated accordingly.

Any further responses from the reviewers can be found at the end of the article

Introduction

A new coronavirus disease known as COVID-19, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first reported in Wuhan, China in December 2019.[1] SARS-CoV-2 is a positive-sense single stranded RNA virus with a helical nucleocapsid. The genome size of SARS-CoV-2 is about 30 kilobases. There are 11 protein-coding genes from the SARS-CoV-2 genome including four structural genes (spike (S), envelope (E), membrane (M), and nucleocapsid (N) genes) and seven nonstructural genes (ORF1ab, ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10).[2]

The SARS-CoV-2 virus can rapidly mutate to bypass the immune response of the host.[3] These mutations can be synonymous, nonsynonymous, deletions, insertions, or others. Nonsynonymous mutations are expected to have a greater impact than synonymous mutations since nonsynonymous mutations affect the amino acid sequences of a protein, subsequently they may change their structures and functions. According to Kim et al. (2020), a total of 767 synonymous and 1352 nonsynonymous mutations have been identified from SARS-CoV-2 genomes.[4] In another study, a total of 119 SNPs were identified using 11,183 SARS-CoV-2 genomes, in which there were 74 nonsynonymous mutations and 43 synonymous mutations.[5] From a study on the analysis of nonsynonymous mutations in structural proteins of SARS-CoV-2, it has been shown that S and N proteins have higher mutation rate per gene compared to that of E and M proteins.[6] The mutations of SARS-CoV-2 proteins may affect viral transmission, host immune evasion, and disease severity. Many SARS-CoV-2 S protein mutations, for example, D614G, N501Y have been heavily studied in detail because S protein facilitates viral entry into human host through the binding of ACE2 receptor.[7] S protein D614G mutation increases viral transmission in animal model and human cell line study[8] while S protein N501Y mutation results in an increase in binding to ACE2 receptor[9] and an increase in viral replication in human upper-airway cells.[10] However, the biological consequences of some of these mutations on the functions and structures of SARS-CoV-2 proteins remain unclear. In our study, computational analysis of the nonsynonymous mutations of SARS-CoV-2 proteins were performed using different bioinformatics tools including co-mutation analysis, protein structure stability and flexibility analysis, and protein function analysis to predict the effects of the mutations on the structures and functions of proteins.

Methods

Sequences and structures retrieval

The SARS-CoV-2 genomes data were downloaded from GISAID database (Global Initiative on Sharing All Influenza Data, RRID:SCR_018251).[11] In this study, a total number of 30,229 SARS-CoV-2 virus genomes data with collection dates ranging from 2020-01-01 to 2021-03-21 were retrieved. The information on the geographical distribution of the SARS-CoV-2 dataset with the date range were summarized in a table as extended data.[12] To make sure that only high-quality sequences were used, the filters including complete genome, high coverage (<1% Ns and <0.05% unique amino acid mutations (not seen in other sequences in database) and no insertion/deletion unless verified by submitter) and patient status, excluding low coverage were applied. The reference strain NC_045512.2 with a total number of 29903 bases was retrieved from NCBI database (NCBI, RRID:SCR_006472). The wild type protein structures obtained from RCSB PDB (Research Collaboratory for Structural Bioinformatics Protein Data Bank, RRID:SCR_012820) are listed in Table 1.[13] Since N protein R203 and G204 are located at a disordered region which does not have a well-defined three-dimensional structure, no experimental structural data was available for the prediction analysis. A predicted model of N protein model (QHD43423, estimate TM-score = 0.97) generated with D-I-TASSER/C-I-TASSER pipeline was used.[14]

Multiple sequence alignment of SARS-CoV-2 genomes

Multiple sequence alignment (MSA) was performed using rapid calculation in MAFFT (MAFFT, version 7.467, RRID: SCR_011811) which supports alignment for more than 20,000 sequences.[15] After all SARS-CoV-2 sequences were

Table 1. SARS-CoV-2 protein structures used in this study.

Protein	Nucleotide changes	Amino acid changes	Template structure (PDB ID)
ORF1a nsp5	C10376T	P108S	7KPH
ORF1b nsp12	C14408T	P323L	6YYT
	C14708T	A423V	6YYT
S	A23063T	N501Y	7A92
	A23403G	D614G	7A92
ORF3a	G25563T	Q57H	6XDC
N	C28725T	P151L	6VYO
	G28881A	R203K	QHD43423
	G28882A	R203K	QHD43423
	G28883C	G204R	QHD43423

aligned to the reference genome, the multiple sequence alignment file was visualized under MEGA X software, version 10.2.5 build 10210330 (MEGA Software, RRID:SCR_000667).

Identification of nonsynonymous mutations and the statistics of the mutation in the SARS-CoV-2 proteins

The 11 different coding sequences were extracted from these 30,229 strains according to their genomic positions in the reference strain (fasta file format) in NCBI, which is NC_045512.2. Inappropriate sequences of base calling errors, "N" unresolved nucleotides, and undefinable gaps were omitted. Then, the frequency and number of nonsynonymous mutations in these 30,229 strains were identified using a Python script. The frequency percentage of the top 10 nonsynonymous mutations in the primary lineages associated with the past and present variant of concern (VOC) were obtained from COVID CG (COVID CG, RRID:SCR_022266).[16]

Co-mutation analysis of SARS-CoV-2 proteins

The concurrence ratio of each nonsynonymous mutation in the SARS-CoV-2 genome was determined using GESS database (The Global Evaluation of SARS-CoV-2/hCoV-19 Sequences, RRID:SCR_021847)[17] derived from GISAID web server. The concurrence search used for the analysis of the concurrence ratio in the top 10 nonsynonymous mutations is listed in Table 2. The frequency for each SNV in the concurrence search is greater than 0.1%. The chord diagram for co-mutations of nonsynonymous mutations in the SARS-CoV-2 genome was generated using Circos table viewer (Circos, RRID:SCR_011798).[18]

Prediction of mutation effect on protein stability and flexibility

To predict the effects of the mutations on the stability and flexibility of the protein structure, the protein structures were analyzed with DynaMut server (DynaMut, RRID:SCR_021849).[19] The free energy change between the wild type and mutant protein structure ($\Delta\Delta G$) predicts the status of protein stability, in which the values of $\Delta\Delta G$ above zero indicate a good stabilization while any values below zero or negative indicate a destabilizing outcome. The difference in entropic energy between the wild type and mutant structures ($\Delta\Delta S_{Vib}$ ENCoM) predicts the status of protein flexibility, in which the values of $\Delta\Delta S_{Vib}$ ENCoM above zero indicate an increase in flexibility while any values below zero or negative indicate a decrease in flexibility.

Prediction of mutation effect on protein function

SIFT 4G (Sorting Tolerant From Intolerant For Genomes, RRID:SCR_021850)[20] and PROVEAN (Protein Variation Effect Analyzer, RRID:SCR_002182)[21] were used to predict the deleteriousness of the nonsynonymous single nucleotide polymorphisms (nsSNPs) on SARS-CoV-2 protein structure. SIFT 4G predicts the effects of the mutations based on the sequence conservation and amino acid properties. For SIFT 4G analysis, gene annotation file (GTF), fasta file containing the SARS-CoV-2 genome sequences was obtained from Ensembl (Ensembl, RRID:SCR_002344).[22] Subsequently a variant call format file (VCF) comprising all the SNP of SARS-CoV-2 was obtained using SNP-sites tool (RRID:SCR_02226).[12] After that, the SARS-CoV-2 genome database, built with the SIFT 4G algorithm, was created. Lastly, SIFT 4G annotator was applied to annotate the VCF file with SARS-CoV-2 genome database. Mutations with a SIFT 4G score of less than 0.05 were considered deleterious.[20] PROVEAN predicts the effects of the mutations based on the principle of alignment-based score. For PROVEAN analysis, the amino acid sequence along with the amino acid

Table 2. **Concurrence ratio of top 10 nonsynonymous mutations in SARS-CoV-2 proteins.**

Coding region and amino acid change		ORF1a nsp5 P108S	ORF1b nsp12 P323L	ORF1b nsp12 A423V	S protein N501Y	S protein D614G	ORF3a Q57H	N protein P151L	N protein R203K	N protein R203K	N protein G204R
Coding region and amino acid change	**Nucleotide change**	**C10376T**	**C14408T**	**C14708T**	**A23063T**	**A23403G**	**G25563T**	**C28725T**	**G28881A**	**G28882A**	**G28883C**
ORF1a nsp5 P108S	**C10376T**	0	99.9	98.3	0.2	99.9	4.6	97.2	91.6	91.6	91.7
ORF1b nsp12 P323L	**C14408T**	99.9	0	99.9	98.8	99.8	99.7	99.9	99.7	99.8	99.8
ORF1b nsp12 A423V	**C14708T**	98.3	99.9	0	0.1	99.9	0.8	98.3	98.6	98.6	98.6
S protein N501Y	**A23063T**	0.2	98.8	0.1	0	99.2	25.7	0.2	67.3	65.6	66.7
S protein D614G	**A23403G**	99.9	99.8	99.9	99.2	0	99.9	100	99.6	99.7	99.7
ORF3a Q57H	**G25563T**	4.6	99.7	0.8	25.7	99.9	0	1.1	0.4	0.3	0.2
N protein P151L	**C28725T**	97.2	99.9	98.3	0.2	100	1.1	0	98.1	98.1	98.1
N protein R203K	**G28881A**	91.6	99.7	98.6	67.3	99.6	0.4	98.1	0	99.9	99.9
N protein R203K	**G28882A**	91.6	99.8	98.6	65.6	99.7	0.3	98.1	99.9	0	99.9
N protein G204R	**G28883C**	91.7	99.8	98.6	66.7	99.7	0.2	98.1	99.9	99.9	0

variation were processed in the PROVEAN server to get the prediction result. Mutations with a value less than −2.5 were considered as deleterious.[21]

Results

The statistics of nonsynonymous mutations in SARS-CoV-2 proteins

From the multiple alignment analysis, we identified 231 nonsynonymous mutations from 30,229 SARS-CoV-2 genome sequences. Figure 1 shows the numbers of the nonsynonymous mutations found in 11 coding sequences of SARS-CoV-2 proteins. ORF1a has the highest numbers of nonsynonymous mutations, followed by S protein and N protein. The top 10 highest frequency nonsynonymous mutations affecting 9 amino acids residues including ORF1a nsp5 P108S, ORF1b nsp12 P323L and A423V, S protein N501Y and D614G, ORF3a Q57H, N protein P151L, R203K and G204R and their frequency percentage in the primary lineages associated with the past and present VOCs are shown in Table 3.

Co-mutation analysis of SARS-CoV-2 proteins

Some nonsynonymous mutations may be random and have no or little biological impact on viral transmission and pathogenesis. If a single nonsynonymous mutation co-mutates with other mutations, they may evolve together and interact functionally. To study co-mutation between different nonsynonymous mutations, the concurrence ratio of co-mutations in the top 10 nonsynonymous mutations was retrieved from GESS database website as shown in Table 2. The visualization of co-mutations in the top 10 nonsynonymous mutations generated with Circos table view is shown in Figure 2. In this chord diagram, connection ribbons represent co-mutations and each ribbon between row and column segments represents the value of concurrence ratio in each top 10 nonsynonymous mutations. Single colours encoded in circular arranged segments represent its own specific mutation whereas rainbow colours represent co-mutation in each mutation. The size of circular arrangement segments is proportional to the total value of concurrence ratio in a row or column. The circular size segment of ORF3a Q57H (G25563T) with the smallest segment size means the total value of concurrence ratio in row or column of ORF3a Q57H (G25563T) having the lowest concurrence ratio. A high concurrence ratio shows high co-mutation between each mutation with thicker ribbon size. S protein D614G (A23403G) with all other nine nonsynonymous mutations had concurrence ratios greater than 99%. On the other hand, low concurrence ratio shows low co-mutation with thinner ribbon size, for example, mutation ORF3a Q57H (G25563T) had the lowest concurrence ratio, only having a high concurrence ratio with S protein D614G (A23403G) and ORF1b nsp12 (P323L) C14408T, the top 2 nonsynonymous mutations which were present in more than 90% of the reported sequences.

Prediction of mutation effect on protein stability and flexibility

Table 4 summarizes the results of predicted effects of mutations on protein stability and flexibility obtained from DynaMut. Only two mutations, namely ORF1a nsp5 P108S and S protein D614G were predicted to be destabilizing with ΔΔG values of −0.288 and −0.072, respectively. For the prediction of protein flexibility, only S protein D614G was predicted to have an increase in flexibility with an $\Delta\Delta S_{\text{Vib}}$ ENCoM value of 0.523.

Prediction of mutation effect on protein function

The prediction results of nonsynonymous mutations in the SARS-CoV-2 proteins using SIFT 4G and PROVEAN are shown in Table 5. SIFT 4G functional missense mutation score predicted that the P108S mutation in ORF1a nsp5 was deleterious (score 0.00) while four mutations S protein D614G, ORF3a Q57H, N protein R203K and G204R were tolerated (>0.05). However, the SIFT 4G results of ORF1b nsp12 P323L and A423V, S protein N501Y and N protein

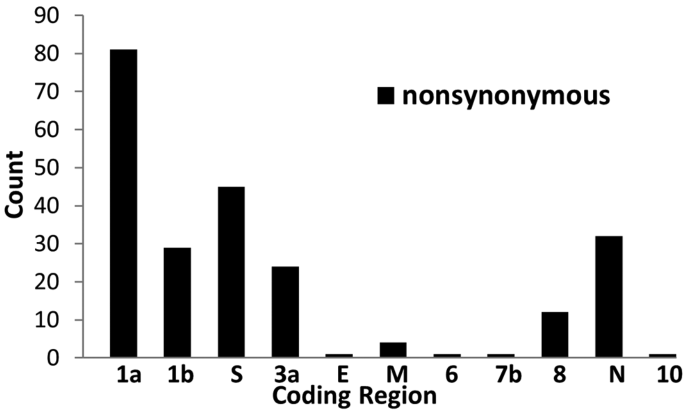

Figure 1. The numbers of nonsynonymous mutations in 11 coding sequences of SARS-CoV-2 proteins.

Table 3. Top 10 nonsynonymous mutations of SARS-CoV-2 proteins and their frequency percentage in the primary lineages of VOCs.

Protein	Nucleotide changes	Amino acid changes	Frequency	Lineage (VOC)				
				B.1.1.7 (α)	B.1.351 (β)	P.1 (γ)	B.1.617.2 (δ)	BA.1 (ο)
ORF1a nsp5	C10376T	P108S	4024	-	-	-	-	-
ORF1b nsp12	C14408T	P323L	27953	100%	90%	99%	100%	100%
	C14708T	A423V	3988	-	-	-	-	-
S	A23063T	N501Y	4218	99%	91%	97%	0%	94%
	A23403G	D614G	28022	100%	100%	100%	100%	100%
ORF3a	G25563T	Q57H	5274	0%	98%	0%	0%	0%
N	C28725T	P151L	4007	-	-	-	-	-
	G28881A	R203K	18116	99%	0%	97%	0%	100%
	G28882A	R203K	18092	99%	0%	97%	0%	100%
	G28883C	G204R	18090	91%	0%	97%	0%	99%

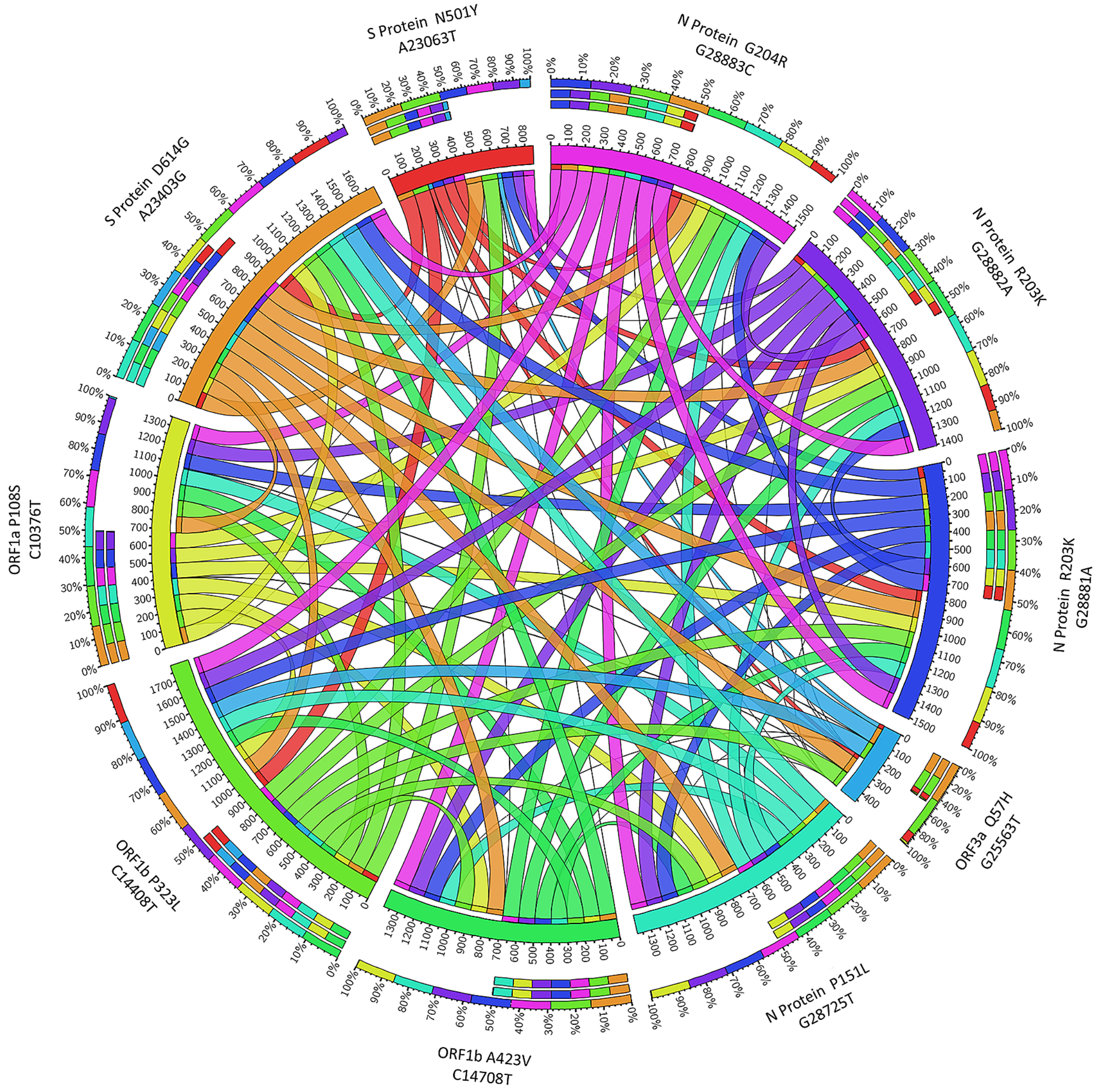

Figure 2. Visualization of co-mutation in top 10 nonsynonymous mutations in SARS-CoV-2 proteins.

Table 4. Prediction of nonsynonymous mutation effect on SARS-CoV-2 proteins stability.

Protein	Mutation	$\Delta\Delta G$ (kcal/mol)	Prediction outcome	$\Delta\Delta S_{Vib}$ ENCoM (kcal.mol^{-1}.K^{-1})	Molecule flexibility
ORF1a nsp5	P108S	−0.288	Destabilizing	−0.208	Decrease
ORF1b nsp12	P323L	1.784	Stabilizing	−0.432	Decrease
	A423V	0.776	Stabilizing	−0.348	Decrease
S	N501Y	0.013	Stabilizing	−0.088	Decrease
	D614G	−0.072	Destabilizing	0.523	Increase
ORF3a	Q57H	0.275	Stabilizing	−0.160	Decrease
N	P151L	1.111	Stabilizing	−0.325	Decrease
	R203K	0.749	Stabilizing	−0.107	Decrease
	G204R	1.064	Stabilizing	−2.522	Decrease

Table 5. Prediction of nonsynonymous mutation effect on SARS-CoV-2 proteins function.

Protein	Mutation	SIFT 4G	Provean
ORF1a nsp5	P108S	0.00 (deleterious)	−3.71 (deleterious)
ORF1b nsp12	P323L	-	−0.91 (neutral)
	A423V	-	1.21 (neutral)
S	N501Y	-	−0.09 (neutral)
	D614G	1.00 (tolerated)	0.60 (neutral)
ORF3a	Q57H	0.61 (tolerated)	−3.29 (deleterious)
N	P151L	-	−4.93 (deleterious)
	R203K	0.11 (tolerated)	−1.60 (neutral)
	G204R	0.08 (tolerated)	−1.66 (neutral)

P151L mutations cannot be obtained due to missing data in the Ensembl database. For the PROVEAN score, three nonsynonymous mutations, namely ORF1a nsp5 P108S, ORF3a Q57H and N protein P151L were predicted to be deleterious (score < -2.5). However, six nonsynonymous mutations, namely ORF1b nsp12 P323L and A423V, S protein N501Y and D614G, N protein R203K and G204R were predicted to be neutral (score > -2.5).

Discussion

The top 10 highest frequency nonsynonymous mutations of SARS-CoV-2 identified from 30,229 SARS-CoV-2 genome sequences were further analyzed with co-mutation analysis, prediction of the protein structure stability and flexibility analysis, and prediction of the protein function analysis. To determine if two nonsynonymous mutations of SARS-CoV-2 proteins co-mutate, concurrence ratio was calculated. Many nonsynonymous mutations showed a high concurrence ratio, suggesting these mutations may evolve together and interact functionally. The top 2 nonsynonymous mutations, S protein D614G and ORF1b nsp12 P323L (as known as RNA-dependent RNA polymerase) showed very high concurrence ratio with other mutations since they emerged in the early phase of the pandemic.[23] Previously it has been shown that S protein D614G co-evolved with ORF1b nsp12 P323L.[23] The combination of both mutations may enhance viral fitness based on epidemiological data, although the molecular mechanisms of this evolutionary advantage remain elusive.[23] Interestingly other mutations with high or medium concurrence ratio are found in more than 90% of the lineage B.1.1.7 in the alpha variant, in which its frequency peaked between March-May 2021. However, the ORF3a Q57H mutation with the lowest concurrence ratio is absent in the lineage B.1.1.7. Unfortunately, the information on the frequency percentage of ORF1a nsp5 P108S, ORF1b nsp12 A423V and N protein P151L mutations in the primary lineages of different VOCs are not available. In another study, it has been predicted that multiple SARS-CoV-2 genes may have epistatic interactions linked to viral fitness.[24] The effects of a mutation can be neutral, harmful, or beneficial to the virus. It is expected that most single mutations have a small effect on viral fitness. It remains an arduous task to associate a specific phenotype with a single viral mutation since it is possible that a specific phenotype is contributed to by the effects of multiple mutations.

There are huge numbers of single nucleotide polymorphisms (SNPs) present in the SARS-CoV-2 genome, hence evaluating the biological functions of all SNPs using experimental approaches is not feasible. Therefore, prediction of the effects of SNPs allows us to prioritize variants which may have some significant biological functions. Our study used the meta-prediction approach to perform functional predictions of nonsynonymous mutations to minimize the false positive rate. When two or three tools are combined, the prediction accuracy increases and reaches greater performance, however, the sensitivity is subsequently decreased as more tools are combined.[25]

Of all these nine protein mutations, only two mutations namely ORF1a nsp5 P108S and S protein D614G were predicted to reduce their stability whereas only S protein D614G may have more a flexible conformation compared to the wild type. S protein binds to human ACE2 receptors to gain access to the host cell.[7] D614G mutation is found at S1 domain which is involved in receptor binding.[26] Two independent studies of S protein D614G mutant structures derived from cryo-electron microscopy analysis has demonstrated that the G614 mutant adopts a more open conformation compared to D614 wild type.[27,28] Interestingly, an *in vitro* study has shown that S protein D614G mutation may enhance virus infectivity by promoting the packing of S protein into the virion, not by enhancing the binding of S protein to the ACE2 receptor.[29] On the other hand, ORF1a nsp5, also known as 3C-like protease is responsible for cleaving viral polypeptides during replication.[2] A study by Abe et al., (2021) has showed that ORF1a nsp5 protein P108S mutation associated with the clade, 20B-T (lineage B.1.1.284), diminished its activity, possibly leading to a reduction in disease severity.[30]

Since the protein function depends directly on the three-dimensional structure of the protein, we wanted to see if these mutations may affect the function of the protein using SIFT 4G and PROVEAN prediction tools. The PROVEAN tool is applicable for all organisms. SIFT4G, instead of SIFT was used since it allows us to build a SARS-CoV-2 genome database with variant annotation. Interestingly ORF1a nsp5 protein P108S mutation was the only mutation found to be deleterious from both SIFT4G and PROVEAN functional analysis. Together with the DynaMut stability result, it has been demonstrated that this mutation may be harmful to the virus itself, and can be less damaging to the human host as reported by Abe et al. (2022).[30] On the other hand, ORF3a Q57H and N protein P151L mutations are predicted to be deleterious by the PROVEAN tool only. ORF 3a is an ion channel (viroporin) which is involved in viral egress steps through lysosomal trafficking.[31,32] ORF3a Q57H mutation not only causes a change in amino acid in ORF3a, but also produces a truncated ORF3b due to the overlapping protein-coding sequences shared by ORF3a and ORF3b.[33] However, there are conflicting results about the effect of the ORF3a Q57H mutation on the human host immune response.[33,34] N protein is involved in the liquid-liquid phase separation for the viral genome packaging.[35] N protein P151L mutation is located at the RNA binding domain. It has been proposed that this mutation may disrupt the protein-drug interaction.[36] Although another two N protein mutations, R203K and G204R were not predicted to be deleterious in our study, they have been identified in the alpha variant, B.1.1.7, gamma variant, P.1, lambda variant, C.37 and omicron variant, BA.1.[37] While N protein, T205I mutation has been reported in the beta variant, B.1.351 and Mu variant, B.1.621.[37] More recently, another N protein mutation, R203M has been reported in the delta variant, B.1.617.2.[37] Interestingly mutants with N protein S202R or R203M mutations can pack more RNA material compared to the wild type based on *in vitro* studies.[38] These observations and experimental results suggest that N protein residues, S202, R203, G204 and T205 may play some role on viral RNA replication.

The SARS-CoV-2 virus genome data from GISAID database ranging from 1st January 20 to 22 March 21 were analyzed in this study. The frequency of alpha variant peaked around March-May 21 in most countries.[37] Hence, it is not surprised that five nonsynonymous mutations, including ORF1b nsp12 P323L, S protein N501Y, S protein D614G, N protein R203K and N protein G204R identified in this study were also part of the defining mutations in the alpha variant.[37] Note that ORF1b nsp12 P323L is the same mutation as ORF1b nsp12 P314L, nsp12 is located between ORF1a and ORF1b, and the last 9 overlapping amino acid residues from ORF1a, SADAQSFLN were included in ORF1b nsp12 P323L. The mutational profile of SARS-CoV-2 genome is changing very rapidly. However, it is out of the scope of this paper to monitor the mutational changes of SARS-CoV-2 genome since it is impossible to keep up with the exponential growth of these data. Interestingly two preprints on genomic surveillance analysis using specimens collected in late 2021 have showed that a small number of wild deer in North America carry the alpha or alpha-like variant.[39,40] Although the alpha variant circulating in human population has been replaced by other variants, it remains to be seen if the alpha variant would jump back from animal to human. Furthermore, these five nonsynonymous mutations found in the alpha variant were also found in the current dominant variant, the omicron variant as shown in Table 3. Therefore, it is still relevant to study the consequences of these mutations.

Conclusion

In this study, ORF1a nsp5 P108S, S protein D614G, ORF3a Q57H and N protein P151L mutations have been predicted to alter their structures and/or functions. Since all the reported variants of concern contain multiple mutations present in multiple SARS-CoV-2 proteins, it is necessary to evaluate the impact of these mutations in combination on viral transmission and pathogenicity. The biological consequences of these nonsynonymous mutations of SARS-CoV-2 proteins should be further validated with *in vivo* and *in vitro* experimental studies in the future.

Ethics and dissemination

No ethical approval is required for data analysis in this study (EA0802021).

Author contribution

CHN contributes to the concept, design, supervision of the project. SBZ, WXB, YYY and SK contribute to the design, methodology, and data collection. SBZ, WXB, YYY and SK contributed to the analysis, and interpretation of data.

All authors were involved in drafting and revising the manuscript and approved the final version.

Data availability

Underlying data

SARS-CoV-2 virus genome sequence data were downloaded from the GISAID Database.

The geographical distribution of the SARS-CoV-2 genome dataset with the date range from 2020-01-01 to 2021-03-21 was summarized in a table and deposited in Figshare.

Figshare: Geographical Distribution (SARS-CoV-2). https://doi.org/10.6084/m9.figshare.19721716.v1[41]

The additional multiple alignment data can be obtained from Figshare

Figshare: MSA (SARS-CoV-2). https://doi.org/10.6084/m9.figshare.16681900.v4[42]

This project contains the following underlying data.

- MSA_0 (31-12-2019 to 31-05-2020).fasta file contains multiple sequence alignment data of SARS-CoV-2 genome sequences ranging between 31-12-2019 and 31-05-2020.

- MSA_1 (01-06-2020 to 15-10-2020).fasta file contains multiple sequence alignment data of SARS-CoV-2 genome sequences ranging between 01-06-2020 and 15-10-2020.

- MSA_2 (16-10-2020 to 31-01-2021).fasta file contains multiple sequence alignment data of SARS-CoV-2 genome sequences ranging between 16-10-2020 and 31-01-2021.

- MSA_3 (01-02-2021 to 22-03-2021).fasta file contains multiple sequence alignment data of SARS-CoV-2 genome sequences ranging between 01-02-2021 to 22-03-2021.

In SIFT4G analysis, we used GTF file containing the SARS-CoV-2 genome sequences, and VCF file comprising all the SNP of SARS-CoV-2.

Figshare: SIFT4G (SARS-CoV-2). https://doi.org/10.6084/m9.figshare.19697365[43]

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

The python script used for the identification of SARS-CoV-2 genome mutations can be obtained through GitHub (https://github.com/wxboon98/Mutations-Identification).

Acknowledgments

This research is supported by Multimedia University, Malaysia, IRFund 2.0 (grant number MMUI/210119 awarded to Chong Han, Ng). The funder has no role in study design, data analysis, decision to publish or manuscript preparation.

References

1. Huang C, *et al.*: **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.** *Lancet.* Feb. 2020; **395**(10223): 497–506.
PubMed Abstract | Publisher Full Text

2. Mousavizadeh L, Ghasemi S: **Genotype and phenotype of COVID-19: Their roles in pathogenesis.** *J. Microbiol. Immunol. Infect.* 2020; **54**(2): 159–163.
Publisher Full Text

3. Harvey WT, *et al.*: **SARS-CoV-2 variants, spike mutations and immune escape.** *Nat. Rev. Microbiol.* 2021; **19**(7): 409–424.
PubMed Abstract | Publisher Full Text

4. Kim J-S, Jang J-H, Kim J-M, *et al.*: **Genome-Wide Identification and Characterization of Point Mutations in the SARS-CoV-2 Genome.** *Osong. Public Health Res. Perspect.* 2020; **11**(3): 101–111.
PubMed Abstract | Publisher Full Text

5. Yuan F, Wang L, Fang Y, *et al.*: **Global SNP analysis of 11,183 SARS-CoV-2 strains reveals high genetic diversity.** *Transbound. Emerg. Dis.* Nov. 2021; **68**(6): 3288–3304.
PubMed Abstract | Publisher Full Text

6. Das JK, Roy S: **A study on non-synonymous mutational patterns in structural proteins of SARS-CoV-2.** *Genome.* 2021; **64**(7): 665–678.
PubMed Abstract | Publisher Full Text

7. Walls AC, Park YJ, Tortorici MA, *et al.*: **Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.** *Cell.* 2020; **181**(2): 281–292.e6.
PubMed Abstract | Publisher Full Text

8. Plante JA , *et al.*: **Spike mutation D614G alters SARS-CoV-2 fitness.** *Nature.* 2021; **592**(7852): 116–121.
PubMed Abstract | Publisher Full Text

9. Starr TN, *et al.*: **Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.** *Cell.* Sep. 2020; **182**(5): 1295–1310.e20.
Publisher Full Text | PubMed Abstract

10. Liu Y, *et al.*: **The N501Y spike substitution enhances SARS-CoV-2 infection and transmission.** *Nature.* 2022; **602**(7896): 294–299.
PubMed Abstract | Publisher Full Text

11. GISAID Initiative: [Accessed: 23-Sep-2021].
Reference Source

12. Page AJ, *et al.*: **SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments.** *Microb. Genomics.* Apr. 2016; **2**(4): e000056.
Publisher Full Text | PubMed Abstract

13. RCSB PDB: Homepage: [Accessed: 01-Dec-2021].
Reference Source

14. Modeling of the SARS-COV-2 Genome using D-I-TASSER: [Accessed: 01-Dec-2021].
Reference Source

15. MAFFT - a multiple sequence alignment program: [Accessed: 23-Sep-2021].
Reference Source

16. Chen AT, Altschuler K, Zhan SH, *et al.*: **Covid-19 cg enables sars-cov-2 mutation and lineage tracking by locations and dates of interest.** *Elife.* Feb. 2021; **10**: 1–15.
Publisher Full Text

17. Fang S, *et al.*: **GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences.** *Nucleic Acids Res.* Jan. 2021; **49**(D1): D706–D714.
PubMed Abstract | Publisher Full Text

18. Krzywinski M, *et al.*: **Circos: An information aesthetic for comparative genomics.** *Genome Res.* Sep. 2009; **19**(9): 1639–1645.
PubMed Abstract | Publisher Full Text

19. Rodrigues CHM, Pires DEV, Ascher DB: **DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability.** *Nucleic Acids Res.* Jul. 2018; **46**(W1): W350–W355.
PubMed Abstract | Publisher Full Text

20. Vaser R, Adusumalli S, Leng SN, *et al.*: **SIFT missense predictions for genomes.** *Nat. Protoc.* Dec. 2015; **11**(1): 1–9.
PubMed Abstract | Publisher Full Text

21. Choi Y, Chan AP: **PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels.** *Bioinformatics.* 2015; **31**(16): 2745–2747.
PubMed Abstract | Publisher Full Text

22. Howe KL, *et al.*: **Ensembl 2021.** *Nucleic Acids Res.* Jan. 2021; **49**(D1): D884–D891.
Publisher Full Text | PubMed Abstract

23. Ilmjärv S, *et al.*: **Concurrent mutations in RNA-dependent RNA polymerase and spike protein emerged as the epidemiologically most successful SARS-CoV-2 variant.** *Sci. Rep.* Jul. 2021; **11**(1): 1–13.
Publisher Full Text

24. Zeng H-L, Dichio V, Horta ER, *et al.*: **Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes.** *Proc. Natl. Acad. Sci.* Dec. 2020; **117**(49): 31519–31526.
PubMed Abstract | Publisher Full Text

25. Sun H, Yu G: **New insights into the pathogenicity of non-synonymous variants through multi-level analysis.** *Sci. Rep.* Feb. 2019; **9**(1): 1–11.

26. Korber B, *et al.*: **Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus.** *Cell.* Aug. 2020; **182**(4): 812–827.e19.
PubMed Abstract | Publisher Full Text

27. Yurkovetskiy L, *et al.*: **Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant.** *Cell.* Oct. 2020; **183**(3): 739–751.e8.
PubMed Abstract | Publisher Full Text

28. Benton DJ, *et al.*: **The effect of the D614G substitution on the structure of the spike glycoprotein of SARS-CoV-2.** *Proc. Natl. Acad. Sci.* Mar. 2021; **118**(9): e2022586118.
PubMed Abstract | Publisher Full Text

29. Zhang L, *et al.*: **SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity.** *Nat. Commun.* Nov. 2020; **11**(1): 1–9.
Publisher Full Text

30. Abe K, *et al.*: **Pro108Ser mutation of SARS-CoV-2 3CL pro reduces the enzyme activity and ameliorates the clinical severity of COVID-19** *Sci. Reports.* 2022; **12**: 123AD, 1299.
Publisher Full Text

31. Miao G, *et al.*: **ORF3a of the COVID-19 virus SARS-CoV-2 blocks HOPS complex-mediated assembly of the SNARE complex required for autolysosome formation.** *Dev. Cell.* Feb. 2021; **56**(4): 427–442.e5.
PubMed Abstract | Publisher Full Text

32. Ghosh S, *et al.*: **β-Coronaviruses Use Lysosomes for Egress Instead of the Biosynthetic Secretory Pathway.** *Cell.* Dec. 2020; **183**(6): 1520–1535.e14.
PubMed Abstract | Publisher Full Text

33. Lam JY, *et al.*: **Loss of orf3b in the circulating SARS-CoV-2 strains.** *Emerg. Microbes Infect.* 2020; **9**(1): 2685–2696.
PubMed Abstract | Publisher Full Text

34. Chu DKW, *et al.*: **Introduction of ORF3a-Q57H SARS-CoV-2 Variant Causing Fourth Epidemic Wave of COVID-19, Hong Kong, China - Volume 27, Number 5—May 2021 - Emerging Infectious Diseases journal - CDC.** *Emerg. Infect. Dis.* May 2021; **27**(5): 1492–1495.
PubMed Abstract | Publisher Full Text

35. Savastano A, Ibáñez de Opakua A, Rankovic M, *et al.*: **Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates.** *Nat. Commun.* Nov. 2020; **11**(1): 1–10.
Publisher Full Text

36. Azad GK: **Identification and molecular characterization of mutations in nucleocapsid phosphoprotein of SARS-CoV-2.** *PeerJ.* Jan. 2021; **9**: e10666.
PubMed Abstract | Publisher Full Text

37. CoVariants: [Accessed: 03-Dec-2021].
Reference Source

38. Syed AM, *et al.*: **Rapid assessment of SARS-CoV-2 evolved variants using virus-like particles.** *Science (80-.).* 2021; **6184**.

39. Marques AD, *et al.*: **Evolutionary Trajectories of SARS-CoV-2 Alpha and Delta Variants in White-Tailed Deer in Pennsylvania.** *medRxiv.* Feb. 2022; 2022.02.17.22270679.

40. Pickering B, *et al.*: **Highly divergent white-tailed deer SARS-CoV-2 with potential deer-to-human transmission.** *bioRxiv.* Feb. 2022; **17**: 2022.02.22.481551.

41. Han NC, Boon WX: **Geographical Distribution (SARS-CoV-2). figshare.** *Dataset.* 2022.
Publisher Full Text

42. Boon WX, Ng CH: **MSA (SARS-CoV-2). figshare.** *Dataset.* 2021.
Publisher Full Text

43. Han NC, Sia Z: **SIFT4G (SARS-CoV-2). figshare.** *Dataset.* 2022.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 09 June 2022

https://doi.org/10.5256/f1000research.133846.r138228

✓

Ujwal Ranjit Bagal
Centers for Disease Control and Prevention, Atlanta, GA, USA

Vishal Nayak
[1] Centers for Disease Control and Prevention, Atlanta, USA
[2] Frederick National Laboratory for Cancer Research, Frederick, MD, USA

The authors have addressed all our concerns in the revised version of the manuscript. The current version looks almost ready for indexing.

Below are a few grammatical errors we want them to correct or rephrase the sentence.

1. Since N protein R203 and G204 are located at a disordered region which does not have a .....

   Rephrase the sentence.

2. A predicted model of N protein model (QHD43423, e.....

   Rephrase the sentence

3. Inappropriate sequences of base calling errors, "N" unresolved nucleotides, and undefinable gaps were omitted.

   with basecalling errors

4. Hence, it is not surprised that five nonsynonymous mutations, including ORF1b nsp12 P323L, S protein N501Y, S protein D614G, N protein R203K and N protein G204R identified in this study were also part of the defining mutations in the alpha variant

   change to surprising instead of surprised

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics & Evolutionary Biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 27 May 2022

https://doi.org/10.5256/f1000research.133846.r138227

✓ **In-Hee Lee**
Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA

The revised article properly addresses the points raised by the reviewers.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 10 March 2022

https://doi.org/10.5256/f1000research.76515.r125647

? **In-Hee Lee**
Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA

The authors examined 10 selected nonsynonymous mutations in SARS-CoV-2 genome for their predicted effect on protein stability as well as their co-mutations. Overall the paper is written well to understand the experiments and analysis results.

Methods:

1. Authors mentioned that there were 231 nonsynonymous mutations from the 30,229 SARS-CoV-2 genome sequences used in the analysis. However, only 10 were intensively investigated throughout the paper. Can you explain the criteria why these mutations were selected?

2. The number of nonsynonymous mutations in coding sequences (Figure 1) may need to be adjusted by the length of each coding sequence.

Results:

3. Given the diverse nature of sequences collected by GISAID, it would be helpful to understand if authors could provide more details about 30,229 sequences used in the study: geographic information for the origin of collection, genetic nomenclature (Nextstrain clade, PANGO lineage, variants of concern or interest by WHO).

4. Co-mutation analysis was particularly intriguing because the co-mutation frequencies were high for most mutations. Can you discuss more about this in the Discussion? Also, I wonder if it will persist when co-mutation analysis were done by genetic nomenclature.

5. Mutations are as both nucleotide changes and amino acid changes in most figures and tables, but Figure 2 only shows nucleotide changes while Table 4 and 5 show only amino acid changes. Can you put amino acid changes on Figure 2 for easy cross-match with other figures and tables?

Others:

6. Specifying 10 nonsynonymous variants in the title may help readers from misinterpreting that the paper conducted an intensive investigation of all possible nonsynonymous variants.

7. The findings reported by the paper might have been limited to the sequences collected from a time-period almost a year ago (2020-01-01 ~ 2021-03-21). Adding discussion about the impact of the study with the advent of omicron variants would be interesting to readers of wide backgrounds.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 09 May 2022

Chong Han Ng, Multimedia University, Bukit Beruang, Malaysia

The authors examined 10 selected nonsynonymous mutations in SARS-CoV-2 genome for their predicted effect on protein stability as well as their co-mutations. Overall the paper is written well to understand the experiments and analysis results.

Methods:

1. Authors mentioned that there were 231 nonsynonymous mutations from the 30,229 SARS-CoV-2 genome sequences used in the analysis. However, only 10 were intensively investigated throughout the paper. Can you explain the criteria why these mutations were selected?

RE: The 10 nonsynonymous mutations are identified based on the mutations with the top 10 highest frequency identified from this study.

2. The number of nonsynonymous mutations in coding sequences (Figure 1) may need to be adjusted by the length of each coding sequence.

RE: Figure 1 shows the total number of the mutations in each gene, eg D614G, N501Y in S protein. It is not relevant to include extra information.

Results:

3. Given the diverse nature of sequences collected by GISAID, it would be helpful to understand if authors could provide more details about 30,229 sequences used in the study: geographic information for the origin of collection, genetic nomenclature (Nextstrain clade, PANGO lineage, variants of concern or interest by WHO).

RE: We included the information on primary lineages for the past and present VOCs associated with the top 10 nonsynonymous mutations and their mutation frequency in Table 3. We also included the information on the geographical distribution of the SARS-CoV-2 dataset with the date range summarized in a table as extended data. While we can see diverse genome dataset coming from different regions, we don't know if there is a good correlation between the reported COVID case number and the number of SARS-CoV-2 genome data deposited to GISAID database. There can be some disparity in genomic surveillance in different countries due to these possible reasons, such as the quality of the sequencing data, the accessibility to research funding resource, the socioeconomic status, the government policy. Therefore, it is less relevant for our study since we are not aimed to monitor the SARS-CoV-2 mutation profile in different regions.

4. Co-mutation analysis was particularly intriguing because the co-mutation frequencies were high for most mutations. Can you discuss more about this in the Discussion? Also, I wonder if it will persist when co-mutation analysis were done by genetic nomenclature.

RE: The GESS database we used does not support co-mutation analysis by the clades or lineages. If we use other tools, we may get different results. Therefore, we didn't do co-mutation analysis by the clades or lineages. However, we expand the discussion part on co-mutation analysis based on the information of mutation frequency percentage by the lineages derived from COVID CG database.

5. Mutations are as both nucleotide changes and amino acid changes in most figures and tables, but Figure 2 only shows nucleotide changes while Table 4 and 5 show only amino acid changes. Can you put amino acid changes on Figure 2 for easy cross-match with other figures and tables?

RE: Figure 2 has been revised with the additional information on amino acid changes.

Others:

6. Specifying 10 nonsynonymous variants in the title may help readers from misinterpreting that the paper conducted an intensive investigation of all possible nonsynonymous variants.

RE: To reflect the scope of the study better, the title of paper has been revised to “Prediction of the effects of the top 10 nonsynonymous variants from 30229 SARS-CoV-2 strains on their proteins.”

7. The findings reported by the paper might have been limited to the sequences collected from a time-period almost a year ago (2020-01-01 ~ 2021-03-21). Adding discussion about the impact of the study with the advent of omicron variants would be interesting to readers of wide backgrounds.

RE: Different SARS-CoV-2 variants of concern have specific sets of defining mutations; some are common among these VOCs while some are unique. Additional paragraph in the discussion section has been added to discuss the impact of our study and to explain why the study of some of the identified mutations remain relevant for the newer variants.

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 February 2022

https://doi.org/10.5256/f1000research.76515.r123168

?

Ujwal Ranjit Bagal
Centers for Disease Control and Prevention, Atlanta, GA, USA

Vishal Nayak
[1] Centers for Disease Control and Prevention, Atlanta, USA
[2] Frederick National Laboratory for Cancer Research, Frederick, MD, USA

The authors in this paper have tried to use an extensive set of SARS-COV-2 genomes to identify and select the top 10 non-synonymous mutations for analyzing the co-mutation effect, as well as its effect on the stability and flexibility of the protein structure. Overall, the paper is well written in terms of language and grammar, experiments were well conducted. The title suggests all the variants were analyzed. Hence, it can be modified to show that only the top 10 non-synonymous mutations were studied. Abstract is well written giving all details of the papers finding.

SARS-COV-2 is an excellent example where there are public repositories with highly curated whole genome datasets and associated metadata are available for analysis. The authors need to use more data from 2020 to 2021 as the number of curated genomes available in GISAID is large (200,000+). This will make the analysis less obsolete.

The authors have applied meta-prediction methods for variant analysis. If relevant data in terms of genetic clade, and region where the sample was collected from, can be added to this analysis the results will be more relevant and can be verified using laboratory techniques. Hence, we suggest the authors try to incorporate these points and resolve a few issues mentioned below and resubmit the paper with additional tables and content.

Below are a few comments for the authors we as reviewers suggest:

1. The introduction seems too small. More details about the virus and work showing the effect of non-synonymous mutations affecting the viral efficacy need to be mentioned.

2. In the Methods section
    1. For the downloaded datasets, what was the threshold used for coverage? A table showing the number of genomes with date (range should do), coverage above threshold, genetic nomenclature (clade name), and geographical information will be helpful to understand the diversity within the dataset.

    2. You have performed a Co-mutation analysis using the GESS database. Does it provide information about the mutation frequency, which genetic nomenclature it was observed? If you can provide that information, it will be useful. The concurrence table is good, but with knowledge of the above information it will become more relevant.

    3. What was the criteria used for “top 10 nonsynonymous mutations”?

    4. For prediction of mutation effect on protein function, where was the GTF file, as well

   as the VCF files, obtained from? There is no mention of whole genome SNP analysis. This part is a bit confusing. Clarification is required.

   5. “Mutations with a value less than-2.5 were considered as deleterious”. Can you provide a reference showing why -2.5 is used as a threshold? Same with the SIFT 4G score threshold of 0.05.

3. In the Results section:
   1. Figure 2: Is it possible to add the amino acid changes (e.g., D614G) instead of just nucleotide mutations for better understanding?

   2. Is it possible to show the genetic nomenclature associated with the top 10 nonsynonymous mutations?

   3. Also, if possible, can you add figures showing the domain or the position on a 3D protein structure? This is optional as you have discussed the domain for few proteins in the discussion section.

4. In the Discussion section:
   1. You write “showed very high concurrence ratio with other mutations since they emerged in the early phase of the pandemic”. How did you come to this conclusion? With reference to our comments in the results and methods section, if you can add this information in a tabular format it will be more informative.

   2. “The combination of both mutations may enhance viral fitness based on epidemiological data”. There is no mention of the epidemiological data in the results section. If it's in the supplementary files, mention it.

   3. “P108S mutation diminished its activity, possibly leading to a reduction in disease severity.” Can you mention in which genetic clade it was observed?

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics & Evolutionary Biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 09 May 2022

Chong Han Ng, Multimedia University, Bukit Beruang, Malaysia

"The authors in this paper have tried to use an extensive set of SARS-COV-2 genomes to identify and select the top 10 non-synonymous mutations for analyzing the co-mutation effect, as well as its effect on the stability and flexibility of the protein structure. Overall, the paper is well written in terms of language and grammar, experiments were well conducted. The title suggests all the variants were analyzed. Hence, it can be modified to show that only the top 10 non-synonymous mutations were studied. Abstract is well written giving all details of the papers finding."

RE: To reflect the scope of the study better, the title of paper has been revised to “Prediction of the effects of the top 10 nonsynonymous variants from 30229 SARS-CoV-2 strains on their proteins.”

"SARS-COV-2 is an excellent example where there are public repositories with highly curated whole genome datasets and associated metadata are available for analysis. The authors need to use more data from 2020 to 2021 as the number of curated genomes available in GISAID is large (200,000+). This will make the analysis less obsolete."

RE: We used SARS-CoV-2 virus genome data ranging from 1st January 20 to 22 March 21. The first draft of the manuscript was prepared in late June 21 and submitted in late August 21. Due to some delay in the editorial review, the paper was only published on 6th January 2022. Five nonsynonymous mutations, including ORF1b nsp12 P323L, S protein N501Y, S protein D614G, N protein R203K and N protein G204R identified in this study were also part of the defining mutations in the alpha variant. The mutational profile of SARS-CoV-2 genome is changing very rapidly. However, we are not aiming to monitor the mutational changes of SARS-CoV-2 genome since it is impossible to keep up with the exponential growth of these data. As of 26 April 2022, there are more than 10 million SARS-CoV-2 virus genomes sequences deposited in GISAID database. Although it is useful to get more recent dataset, it is out of the scope of our study to update the data. Both identification of the SARS-CoV-2 mutations and the prediction analysis of the biological consequences of these mutations are very time-consuming processes. Additional paragraph in the discussion section has been added to discuss the impact of our study and to explain why the study of some of the identified mutations remain relevant for the newer variants.

"The authors have applied meta-prediction methods for variant analysis. If relevant data in terms of genetic clade, and region where the sample was collected from, can be added to this analysis

the results will be more relevant and can be verified using laboratory techniques. Hence, we suggest the authors try to incorporate these points and resolve a few issues mentioned below and resubmit the paper with additional tables and content."

RE: We included the information on primary lineages for the past and present VOCs associated with the top 10 nonsynonymous mutations and their mutation frequency in Table 3. We also included the information about the geographical distribution of the SARS-CoV-2 dataset with the date range summarized in a table as extended data. We are not performing extra experiments to verify the prediction data since our work is primarily focused on prediction analysis of the mutations. The lab work is out of the scope of this paper, and it is too time-consuming and labour-intensive to perform experimental works.

"Below are a few comments for the authors we as reviewers suggest:
The introduction seems too small. More details about the virus and work showing the effect of non-synonymous mutations affecting the viral efficacy need to be mentioned."

RE: The introduction section with the examples of the effect of non-synonymous mutations affecting the viral efficacy has been included in the last paragraph.

"In the Methods section

For the downloaded datasets, what was the threshold used for coverage? A table showing the number of genomes with date (range should do), coverage above threshold, genetic nomenclature (clade name), and geographical information will be helpful to understand the diversity within the dataset."

RE: For the downloaded dataset, high coverage filter has been applied. The high coverage is defined as only entries with <1% Ns and <0.05% unique amino acid mutations (not seen in other sequences in database) and no insertion/deletion unless verified by submitter, according to GISAID. We included the information on primary lineages for the past and present VOCs associated with the top 10 nonsynonymous mutations and their mutation frequency in Table 3. We have the information about the geographical distribution of the SARS-CoV-2 dataset with the date range summarized in a table as extended data. While we observe a diverse genome dataset coming from different regions, we don't know if there is a good correlation between the reported COVID case number and the number of SARS-CoV-2 genome data deposited to GISAID database. There may be some disparity in genomic surveillance in different countries due to these possible reasons, such as the quality of the sequencing data, the accessibility to the research funding resource, the socioeconomic status, the government policy. Therefore, it is less relevant for our study since we are not aimed to monitor the SARS-CoV-2 mutation profile in different regions.

"You have performed a Co-mutation analysis using the GESS database. Does it provide information about the mutation frequency, which genetic nomenclature it was observed? If you can provide that information, it will be useful. The concurrence table is good, but with knowledge of the above information it will become more relevant."

RE: The GESS database doesn't have the information about the mutation frequency of the mutations associated with the lineages or clades. However, we included the information obtain from COVID CG database on primary lineages for the past and present VOCs associated with the top 10 nonsynonymous mutations and their mutation frequency in Table 3. In addition, we expand the discussion part on co-mutation analysis based on the information of mutation frequency percentage by the lineages.

"What was the criteria used for "top 10 nonsynonymous mutations"?"

RE: The top 10 nonsynonymous mutations are identified based on the mutations with the highest frequency identified from this study.

"For prediction of mutation effect on protein function, where was the GTF file, as well as the VCF files, obtained from? There is no mention of whole genome SNP analysis. This part is a bit confusing. Clarification is required."

RE: Additional information on GTF and VCF files are added in the methods. The GTF and VCF files are deposited in Figshare and the related information are included in Data and software availability section. It is a whole genome SNP analysis, but the SIFT 4G results of ORF1b nsp12 P323L and A423V, S protein N501Y and N protein P151L mutations cannot be obtained due to missing data in the Ensembl database.

"Mutations with a value less than-2.5 were considered as deleterious". Can you provide a reference showing why -2.5 is used as a threshold? Same with the SIFT 4G score threshold of 0.05.

RE: Both references for the scoring method of SIFT4G and PROVEAN have been included.
In the Results section:

"Figure 2: Is it possible to add the amino acid changes (e.g., D614G) instead of just nucleotide mutations for better understanding?"

RE: Figure 2 has been revised with the additional information on amino acid changes.

"Is it possible to show the genetic nomenclature associated with the top 10 nonsynonymous mutations?"

RE: The information on the mutation frequency percentage in the primary lineages associated with the past and present variant of concern (VOC) have been updated in Table 3.

"Also, if possible, can you add figures showing the domain or the position on a 3D protein structure? This is optional as you have discussed the domain for few proteins in the discussion section."

RE: There are multiple SARS-CoV-2 proteins mentioned in the paper. We are not doing any work on protein structure modelling. The readers should refer to Protein Data Bank if they want to know more specific information on the protein domains.

"In the Discussion section:

You write “showed very high concurrence ratio with other mutations since they emerged in the early phase of the pandemic”. How did you come to this conclusion? With reference to our comments in the results and methods section, if you can add this information in a tabular format it will be more informative."

RE: Table 3 shows that S protein D614G and ORF1b nsp12 P323L mutations have the top 2 highest frequency. They are found in more than 90% of 30229 SARS-CoV-2 genome sequences, which are from the early batch of SARS-CoV-2 genome data. A similar finding has been reported by S. Ilmjärv et al., “Concurrent mutations in RNA-dependent RNA polymerase and spike protein emerged as the epidemiologically most successful SARS-CoV-2 variant,” Sci. Rep., vol. 11, no. 1, pp. 1–13, Jul. 2021.

“The combination of both mutations may enhance viral fitness based on epidemiological data”. There is no mention of the epidemiological data in the results section. If it's in the supplementary files, mention it.

RE: We are referring to the study published by S. Ilmjärv et al., “Concurrent mutations in RNA-dependent RNA polymerase and spike protein emerged as the epidemiologically most successful SARS-CoV-2 variant,” *Sci. Rep*., vol. 11, no. 1, pp. 1–13, Jul. 2021.

“P108S mutation diminished its activity, possibly leading to a reduction in disease severity.” Can you mention in which genetic clade it was observed?

RE: The genetic clade associated with ORF1a nsp5 P108S is 20B-T (lineage B.1.1.284). However, it is unknown if this mutation is associated with the past and current of variants of concern since the data is unavailable from COVID CG database.

Competing Interests: No competing interests were disclosed.