


RESEARCH

Open Access



# A multi-population-based genomic analysis uncovers unique haplotype variants and crucial mutant genes in SARS-CoV-2

Afzal Sheikh<sup>1\*</sup> , He Huang<sup>2</sup>, Sultana Parvin<sup>3</sup>, Mohammad Badruzzaman<sup>1</sup>, Tofayel Ahamed<sup>4</sup>, Ekhtear Hossain<sup>5\*</sup>, Iri Sato Baran<sup>6</sup> and Zahangir Alam Saud<sup>7</sup>

## Abstract

**Background:** COVID-19 is a disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Rigorous detection and treatment strategies against SARS-CoV-2 have become very challenging due to continuous evolutions to the viral genome. Therefore, careful genomic analysis is sorely needed to understand transmission, the cellular mechanism of pathogenicity, and the development of vaccines or drugs.

**Objective:** In this study, we intended to identify SARS-CoV-2 genome variants that may help understand the cellular and molecular foundation of coronavirus infections required to develop effective intervention strategies.

**Methods:** SARS-CoV-2 genome sequences were downloaded from an open-source public database, processed, and analyzed for variants in target detection sites and genes.

**Results:** We have identified six unique variants, G---AAC, T---AAC---T, AAC---T, AAC-----T, C-----T, and C-----C, at the nucleocapsid region and eleven major hotspot mutant genes: nsp3, surface glycoprotein, nucleocapsid phosphoprotein, ORF8, nsp6, nsp2, nsp4, helicase, membrane glycoprotein, 3'-5' exonuclease, and 2'-O-ribose methyltransferases. In addition, we have identified eleven major mutant genes that may have a crucial role in SARS-CoV-2 pathogenesis.

**Conclusion:** Studying haplotype variants and 11 major mutant genes to understand the mechanism of action of fatal pathogenicity and inter-individual variations in immune responses is inevitable for managing target patient groups with identified variants and developing effective anti-viral drugs and vaccines.

**Keywords:** SARS-CoV-2, Genome, Variants, Haplotype, Nucleocapsid

## Background

The new case of SARS-CoV-2 outside China was first announced by the Director-General of the WHO on February 26, 2020, and is now officially known as COVID-19 disease [1]. The human COVID-19 pandemic disease caused by the infections of severe acute

respiratory syndrome coronavirus-2 (SARS-CoV-2), which impacts the lower respiratory tract, has spread across the globe in diverse methods and speed [2–4]. The spectrum of symptoms ranges from developing mild to moderate respiratory illness that recovers without hospitalization to the lethal form of COVID-19 associated with severe pneumonia, difficulty in breathing or shortness of breath, chest pain, loss of speech or movement, and fatality [3–7].

Structurally, SARS-CoV-2 is an enveloped, 5'-capped, single-stranded polyadenylated positive-strand RNA virus of a non-segmented genome of ~29.7 kb long encoding 16 non-structural proteins (NSPs), which are

\*Correspondence: asheikh@bsmrau.edu.bd; hekhtear@gmail.com

<sup>1</sup> Department of Biochemistry and Molecular Biology, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Dhaka 1706, Bangladesh

<sup>5</sup> Department of Biological Sciences and Chemistry, Southern University and A&M College, 244 William James Hall, Baton Rouge, LA 70813, USA  
Full list of author information is available at the end of the article

required for virus replication and pathogenesis. Four structural proteins, including envelope (E), membrane (M), nucleocapsid (N), and spike glycoprotein (S), are essential for virus subtyping, structural rearrangement of the RNA genome, assembly, budding, viral replication, pathogenesis, response to vaccines, and viral entry to host. Moreover, nine others are accessory factors that facilitate the unwinding of dsRNA, viral RNA cap formation, exonuclease activity, membrane fusion, interaction with host cells, and immune response to the host [8–12]. Thus, mutations in these genes may interfere with changing protein structures, RNA dimerization, and alterations in the functions as mentioned earlier, including interaction with RNA and signaling events [13–15]. Moreover, some functional features of these genes are yet to be discovered.

To date, many drugs have been applied to manage COVID-19 patients, along with several vaccines. Unfortunately, there are no effective drugs so far, and if some of the drugs are functioning with some adverse side effects, individual patient groups are not responding to those drugs [16–21] (<https://www.who.int/news/item/07-04-2021-interim-statement-of-the-covid-19-subcommittee-of-the-who-global-advisory-committee-on-vaccine-safety>). In addition, scientific communities are aware of some repeatedly reported limitations of the already available vaccines, including recurrence infections after being vaccinated with multiple doses, adverse side effects, and fatality [22–26]. These limitations are reported in a particular group of patients while some other target groups effectively responded to those already available drugs or/and vaccines [16–25] (<https://www.who.int/news/item/07-04-2021-interim-statement-of-the-covid-19-subcommittee-of-the-who-global-advisory-committee-on-vaccine-safety>). Therefore, in order to develop new effective therapeutic strategies for these non-responsive patient groups and adverse drug effects, it is crucial to study the association of the target variants with pathogenicity, replication rate, recurrence infections, response to host immunity, and target drugs at the cellular and molecular level.

In the present study, we focused on characterizing the accumulation of mutations and a detailed understanding of the geographic distribution of genetic variants in 1,012,582 sequences, including 405,461 complete genome sequences from the NCBI database as of August 4, 2021. From the shreds of evidence, we are reporting for the first time the seven unique haplotype variants in the nucleocapsid region, four of which is in the target RT-PCR detection sites recommended by the central research institute CDC in the USA, China, Germany, and Japan's Center of Infectious Disease (NIID) testing protocol (<https://www.cdc.gov/coronavirus/2019-ncov/lab/>

[rt-pcr-panel-primer-probes.html](#)) [27, 28]. In addition, we have identified the major hotspot mutant genes, some of which have been reported before to be associated with RNA capping and viral replication, infection, and pathogenesis. Therefore, this study will be of great interest to scientists working in cellular and molecular biology, molecular pathogenicity, medicine, and researchers working in vaccine development, including the scientific community working on infectious diseases detection, diagnosis methods, and human health care.

## Methods

### SARS-CoV-2 sequence data analysis

We intended to analyze the major hotspot mutations at the nucleocapsid phosphoprotein and envelop region since these two regions are the major target for RT-PCR-based detection of COVID-19-positive cases by CDC in the USA, China, and Germany and NIID in Japan. In addition, major hotspot mutant sites were analyzed for the complete genome of COVID-19 sequence data of global samples from the open sources database. Therefore, we first downloaded 1,012,582 available SARS-CoV-2 global sequence data from the National Center for Biotechnology Information (NCBI) database. We then separately processed and analyzed the complete and partial sequence data.

### Data processing

After downloading the sequence, data were processed for variant analysis using a Linux terminal using the following command lines, python and muscle program for the target region, N and E (nucleocapsid phosphoprotein and envelope protein).

```
grep ">" covid_19.fasta | grep nucleocapsid > goi.txt
grep ">" covid_19.fasta | grep "nucleocapsid" | sed -e 's/>/g' | cut -d " " -f1 > nucleocapsid.list
Create_sorted fasta file (nucleocapsid gene):
python3 pull_fasta.py -f covid_19.fasta -l nucleocapsid.list > nucleocapsid.fasta
Collapse duplicate sequences into single sequence:
Vsearch --derep_fulllength nucleocapsid_covid-19.fasta --output uniquenucleocapsid_covid-19.fasta
```

### Sequence alignment and mutation analysis

The commands were performed repeatedly for each target gene. The obtained sequences were further analyzed using muscle for the alignment and to separate unique sequences against the reference COVID-19 genome, NC\_045512 from Wuhan in China. The command-line used is as follows:

```
muscle -in unigenvelop_Covid-19.fasta -out unigenvelopseq.fasta_alignedseq.
```

The unique aligned sequences are then used for the analysis of variation/mutation using jalview application.

In addition, we also analyzed the hotspot mutation sites towards the complete genome of SARS-CoV-2 based on our data filtering criteria stated above and using “View Mutations” in the SARS-CoV-2 SRA Data” link. ([https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType\\_s=Nucleotide&VirusLineage\\_ss=SARS-CoV-2,%20taxid:2697049](https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049)). The data obtained were then analyzed for each gene and mutations, including the protein change type (synonymous/ non-synonymous)

**Results**

**Unique SARS-CoV-2 clones were identified with mutations at the target detection sites in global samples**

Since we identified false-negative results in ~16% of the COVID-positive patients, which were confirmed using several primer sets, we intensely wanted to investigate variations in the SARS-CoV-2 genome, particularly the region where primer-probe sets are designed and recommended by CDC, USA; NIID, Japan; CDC, China; Germany, and others (<https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html>) [27, 28]. We have identified many global samples which have multiple mutations at the same primer-probe binding site. Some clones were identified to have mutations at both primer and probe binding sites. Some were identified to have mutations at multiple primers-probe binding sites, while some global samples were found to have mutations at either of the two primers or probe binding sites (Table 1).

We speculated in the previous study that mutation at the target detection site might significantly impact false-negative results, which we reported is ~16% in Japanese samples [28]. Our present results also indicate the essence of using multiple (at least three) primer sets to reduce the transmission of SARS-CoV-2 infections rate caused by false-negative results.

**SARS-CoV-2 clones with unique haplotype variants are present in the nucleocapsid region**

While analyzing the processed data, we identified six unique haplotype mutation patterns G----AAC, T---AAC----T; AAC---T; AAC-----T; C-----T; and C-----C present in the nucleocapsid region of the SARS-CoV-2 genome (Fig. 1a–d). No similar haplogroup pattern could be identified at other RT-PCR target detection sites. Furthermore, although not at the target detection (primer-probe binding site) sites, three different haplotype variants were also observed (Fig. 1e–f) in the nucleocapsid region that encodes for nucleocapsid phosphoproteins.

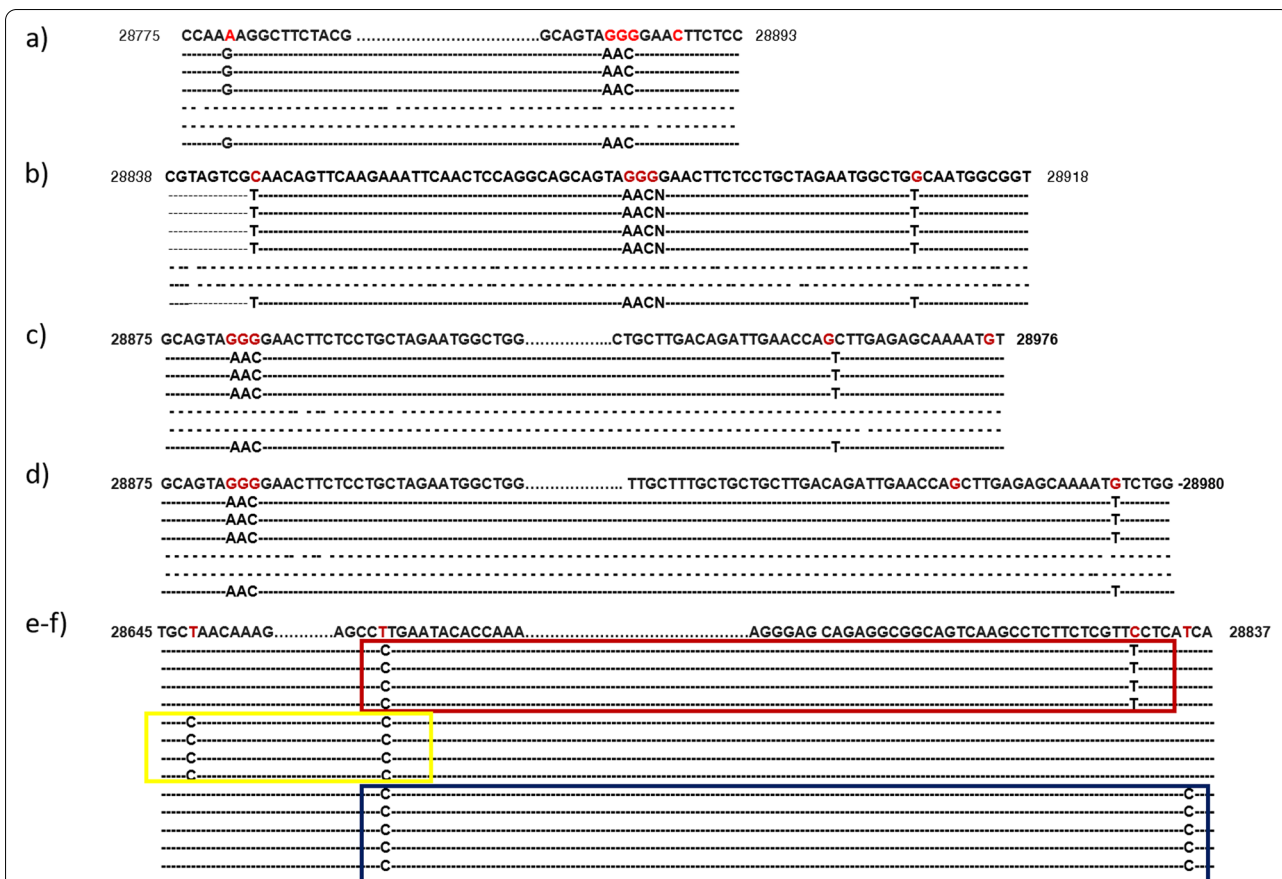
This protein is associated with the viral structural rearrangement of genomic RNA and serves several functions essential for viral replication and RNA dimerization [13–15]. Therefore, these unique haplotype variants may have the possibility to play a role in the variation of pathogenicity, infection rate, recurrence infection, and mortality rate, including the immune response. Therefore, it demands the molecular level study of those haplogroups for their possible association with the parameter mentioned earlier, including mortality rate! Moreover, recently developed vaccine functionality could be validated against those haplogroups from those who did not respond to the given vaccine.

**Major hotspot mutant genes and sites were identified in the global SARS-CoV-2 genome**

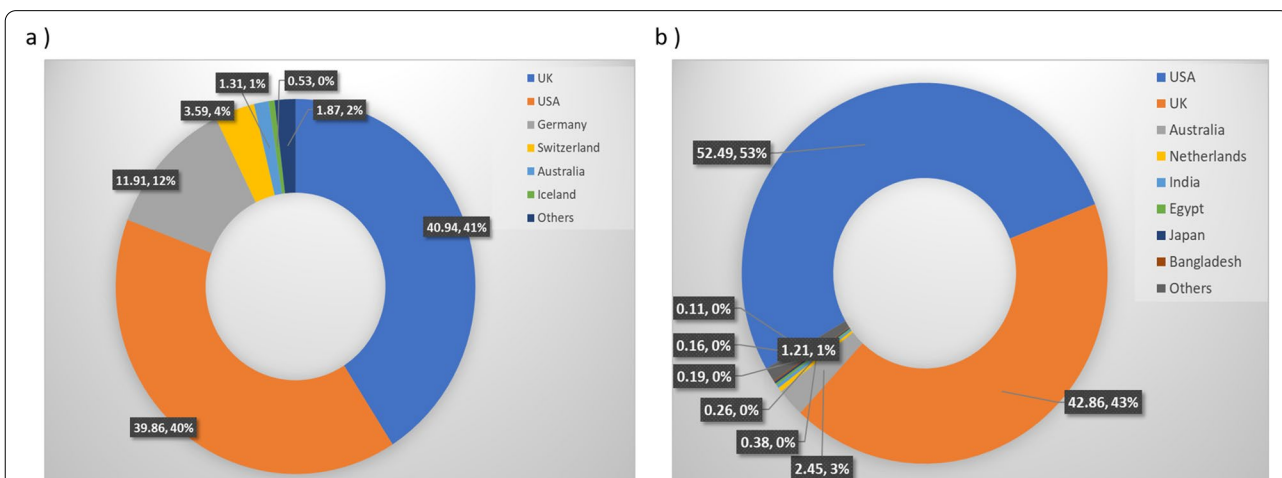
To identify major hotspot mutants, we analyzed 1,012,582 global SARS-CoV-2 sequence data available in the NCBI database as of August 4th, 2021. We analyzed the global distribution of these sequences for complete and partial genome sequences (Fig. 2) and identified major hotspot mutant genes (Table 2). The global mutation distribution data revealed that the top 11 major mutant genes had synonymous mutations observed in > 50,000 global samples (Table 3). Mutation at each of

**Table 1** The number of global samples identified with the mutation at the primer-probe binding sites for RT-PCR-based detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

SARS-CoV-2	PCR detection sites (primers and probes)						
	US_CDC_N2	JP_NIID_N2	China_CDC_N2	JP_NIID_2 and US_CDC_N2	China_CDC_N2 and JP_NIID_N2	China_CDC_N2 and US_CDC_N2	JP_NIID_2; China_CDC_N2 and US_CDC_N2
Number of clones with mutation at the detection sites	52	44	60	10	10	17	1



**Fig. 1** Representation of the haplotype variants was observed in the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome. **a–d** Represent the genomic location of each haplogroup identified in the nucleocapsid region which is a target of RT-PCR-based detection of severe acute respiratory syndrome coronavirus 2 (binding sites of primers recommended by the World Health Organization, the Centers for Disease Control and Prevention in USA and China, and Germany including Japan’s Center of Infectious Disease testing protocol). **e–f** haplotype variants were observed at nucleocapsid other than the RT-PCR target region



**Fig. 2** Global distribution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes by sample collection locations from 101 countries. **a** Distribution of 1,012,582 SARS-CoV-2 genomes (complete and partial genome sequence) by countries as of August 4th, 2021. **b** Global distribution of 405,461 complete SARS-CoV-2 genomes by region (from 88 countries)

**Table 2** Major hotspot mutant genes with synonymous mutation, codon, and protein changes with mutation sites of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome in 1,012,582 global samples

Protein	Protein change	Count (sample)	Genomic location	Codon change	Protein change type
nsp3	F106F	607711	3038	TTT > TTC	synonymous
nsp3	F1089F	250156	5987	TTT > TTC	synonymous
nsp3	T1189T	63447	6287	ACT > ACC	synonymous
nsp3	T1607T	12512	7541	ACC > ACT	synonymous
nsp3	P512P	10612	4256	CCA > CCG	synonymous
surface glycoprotein	I68I	31555	21767	ATC > ATA	synonymous
surface glycoprotein	A924A	15709	24335	GCT > GCC	synonymous
surface glycoprotein	Q613Q	12596	23402	CAA > CAG	synonymous
nucleocapsid phosphoprotein	R203R	322626	28883	AGA > AGG	synonymous
ORF8 protein	H17H	47834	27945	CAT > CAC	synonymous
ORF8 protein	F120F	10048	28254	TTT > TTC	synonymous
nsp6	V120V	113439	11333	GTG > GTA	synonymous
nsp2	S36S	249655	914	TCT > TCC	synonymous
nsp2	N435N	17370	2111	AAT > AAC	synonymous
nsp4	D144D	113780	8987	GAT > GAC	synonymous
helicase	N268N	35967	17041	AAC > AAT	synonymous
helicase	T137T	12905	16648	ACT > ACG	synonymous
membrane glycoprotein	L93L	62743	26802	CTG > CTC	synonymous
membrane glycoprotein	F53F	10873	26682	TTT > TTC	synonymous
3'-to-5' exonuclease	D172D	14808	18556	GAT > GAC	synonymous
2'-O-ribose methyltransferase	A199A	62438	21256	GCC > GCG	synonymous

the surface glycoprotein, nucleocapsid phosphoprotein, ORF8, and ORF3a protein-coding gene was observed in 24%, 19%, 7%, and 3% of the global samples, respectively, while top mutations at non-structural protein-coding genes nsp3 and at each of the nsp6, nsp4, and nsp2 were observed in 18% and 4% of the global samples, respectively. In addition, ORF3a protein and Helicase coding genes were observed to have a mutation at 3%, while both ORF7a and 3'-5' exonuclease were observed to have a mutation in 2% of the global samples (Fig. 3).

## Discussion

In this multi-population-based SARS-CoV-2 genome analysis study, we have identified six unique haplotype variants and 11 major hotspot mutant genes that might play a crucial role in inter-individual variation in COVID-19 pathogenicity, severity, immune response, and mortality rate. Studying the association of these variants and hotspot mutant genes at the cellular and molecular level

may help in understanding the mechanisms of pathogenicity, progression of the disease to more severity and mortality, response to drugs, and immune response to vaccines. Therefore, it will help manage individual SARS-CoV-2 patient groups with identified haplotype variants and major mutant genes by developing effective drugs and vaccines for the target subgroups. In addition, we have identified SARS-CoV-2 clones with mutations at primer and probe binding sites that might cause false-negative PCR detection results. Efficient diagnosis and treatment strategies have become a big challenge to the medical community and healthcare professionals due to the significantly high false-negative detection rate. We showed in our previous study that high false-negative results might be due to the genomic variations at the primer and probe binding sites. In the current study, we have identified global SARS-CoV-2 clones with mutations at target PCR-detections sites. The mutations were observed at either primer-probe binding sites or

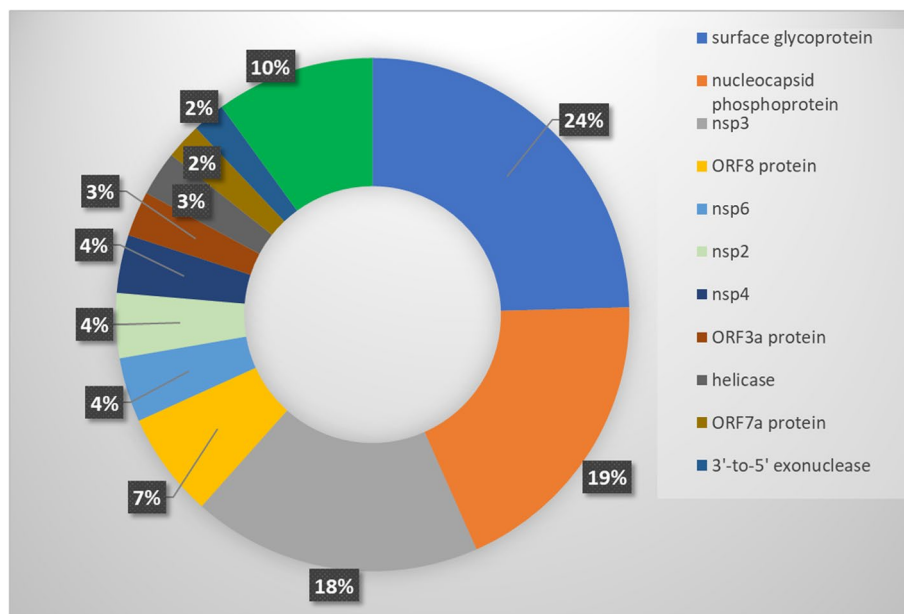
**Table 3** Global distribution of 11 major mutant genes with the number of synonymous and non-synonymous mutations of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome present at least in >10,000 global samples

protein	No. of synonymous Mut.	No. of non-synonymous mut.	Total Mut.
nsp3	5	9	14
surface glycoprotein	3	22	25
nucleocapsid phosphoprotein	1	15	16
ORF8 protein	2	7	9
nsp6	1	8	9
nsp2	2	3	5
nsp4	1	4	5
helicase	2	2	4
membrane glycoprotein	2	1	3
3'-to-5' exonuclease	1	2	3
2'-O-ribose methyltransferase	1	1	2

both sites (Table 1). We also observed that some clones showed mutations at the multiple primer-prob sites recommended by CDC, USA; China; and Japan's NIID (Supplementary Table 1) raises the concern. These concerns have emerged recently, notably regarding the sensitivity and accuracy of the RT-PCR-based detection of false-negative data even after frequent retesting procedures, and might play a significant role in transmitting the virus without traceability of the sources. Therefore, we recommend using at least three more alternative primer-probe sets for RT-PCR detections of SARS-CoV-2 along with the currently used primers and probes sets.

While we were analyzing the variants at the target detections sites, we identified six unique haplotype variants, at the nucleocapsid regions, N encoding nucleocapsid phosphoprotein of which three variants are present at or near the target detection sites; however, the other three haplotype variants are located at the distant upstream of the target detection sites

(Fig. 2). Nucleocapsid phosphoprotein (N), also known as the replication-transcription complexes (RTCs), has been reported to be associated with early and late viral replication, structural rearrangement of the genomic RNA, viral RNA dimerization and serves several functions essential for viral replication [29–31]. Therefore, it demands the molecular level studies if these haplotype variants present in target subgroups are associated with the alterations of nucleocapsid functions in SARS-CoV-2 pathogenicity and if they facilitate the functions of other structural or non-structural proteins. We also investigated other genes that have been reported to exhibit functions in viral pathogenesis and are the targets for anti-viral drug development [32–40]. We identified eleven major mutant genes with major hot-spot and synonymous mutations, each of which mutations were observed in at least 50,000 global samples (Table 2). The global distribution of these major mutant genes revealed that the highest mutations were present



**Fig. 3** Distribution of major hotspot mutant genes in 405,461 global samples with complete SARS-CoV-2 genome sequences. Complete genomes sequences from National Center for Biotechnology Information (NCBI) database were analyzed for top mutant genes as of August 4th, 2021, a total of 405,461 complete SARS-CoV-2 genomes from 88 countries were analyzed

in structural protein-coding gene surface glycoproteins, nucleocapsid phosphoprotein, and non-structural coding gene nsp3 with 25, 16, and 9 hotspots mutant sites, respectively, in the global samples. Mutations identified in surface glycoprotein could affect its function in the receptor recognition and cell membrane fusion process with host-receptors angiotensin-converting enzyme 2 (ACE2) [39–41].

Synonymous and non-synonymous mutation detected in the nucleocapsid phosphoprotein region residing at SARS-CoV-2 RNA synthesis sites might have a negative regulatory influence in viral genomic RNA packaging during virion assembly and suppression of host immune response through RNA-dependent phase separation [30, 42]. In addition, the C-terminal domain of nucleocapsid phosphoprotein has been reported to be associated with anchoring the viral Nsp3, also known as papain-like protease, a component of RTCs. nsp3 catalyzes the reaction that preferentially cleaves ubiquitin-like interferon-stimulated gene 15 (ISG15) protein from interferon factor 3 (IRF3) which weakens the type I interferon response, could exacerbate hyperinflammatory conditions and progression to severe COVID-19 [43, 44]. nsp2 and nsp3 are conserved sequences that have no homology with other Coronaviruses. Moreover, ORF8 and 3'-to-5' exonuclease (nsp14) has been reported to suppress immune response through disrupting IFN-I signaling,

down-regulating MHC-I, and inhibiting IFN $\gamma$ -induced anti-viral gene expression in human lung epithelial cells [45–48] while membrane glycoprotein (M) has been reported to acts as a negative regulator of innate immune response [29, 30, 35].

Therefore, mutation analysis of these genes may reveal potential mechanisms that distinguish COVID-19 from other viruses, as well as inter-individual differences in immune response and COVID-19 severity.

Helicase (nsp13), the most conserving site of SARS-CoV-2, contains two druggable pockets, nucleoside triphosphate hydrolase (NTPase) and helicase activities that hydrolyze and unwind RNA helices. In viral life cycles, nsp13 and nsp14 play the central role in RNA replication by unwinding the duplex RNA and its exoribonuclease (ExoN) N7-methyltransferase (N7-MTase) activities, respectively. In addition, Nsp13 facilitates the correct folding of the viral protein into 2ndary and tertiary structures to become functional. Therefore, studying the mutations in this gene could suggest possible interindividual variation in the drug response and pathogenicity [48–51].

To understand the COVID-19-related target drug-gene interactions and for the selection of effective drugs, molecular level studies will be needed for each of the proposed target variants. Any target drug or chemical compound should be molecularly docked for its binding

**Table 4** Comparison of analytics of the present study with the existing similar literature. A comparison of analytics was made representing the global sample volume, distribution, unique variants, major mutant genes, variant sites, and filtering criteria

Types of analytics	The one existing similar literature	In the current study
Number of genomes analyzed	3067	1, 012,582
Global distribution of host	55 countries	88 countries
Unique haplotype variants	No such studies	6 unique variants
Number of hotspot mutant genes	6	11
Total variant sites	782, of which 582 (65 %) had non-synonymous effects	21, 509, of which 12,910 (60%) had non-synonymous effects
Technical error filtering (exclusion of samples having a mutation in 10,000 samples, < 2%) for major mutant genes	No	Yes
Information on mutation sites on major mutant genes (synonymous or non-synonymous)	No	yes

affinity with the proteins of the host cells for example angiotensin-converting enzyme II (ACE II) as well as with the proteins expressed by the target genes of SARS-CoV-2 genome. In addition, studying the molecular network or signaling will be needed. Furthermore, investigating if these unique variants will have impacts on drug-gene interaction and signaling network, as well as impacts on pharmacokinetics using target chemical compounds that are used to treat COVID-19 patients, for example, Diosgenin, Syringaresinol-O-beta-D-glucoside, etc., are present in the traditional Chinese medicinal herb used to treat COVID-19 patients as an alternative could be a subject for future studies [52, 53]. We did analyze the number of mutations and sites of mutations of each of the crucial eleven mutant genes (Table 3). To avoid the biases of the sequence data, sequencing procedures including PCR-based sequencing and machines and analysis pipeline may cause errors we avoided genes that have been found mutated in < 10,000 global samples (Supplementary Table 2). For the first time, we are reporting the unique haplotype variants and other potential target variants in 11 major mutant genes by analyzing a large number of SARS-CoV-2 global samples ( $n=1,012,582$ ). A comparison of the analytics has been performed in the present study with the one existing similar investigation (Table 4) demonstrating the importance of the present study.

All these crucial mutant genes have been reported to be linked to SARS-CoV-2 pathogenicity, viral replication, virus-host interaction, transmission, and immune response to the host [30, 35, 39–51]. Therefore, any individual subgroups with these mutations may have shown variations in gene functions and mechanisms mediating the traits or phenotypes caused by mutations and may require special management procedures, treatment strategies, and effective vaccinations. Further molecular level studies are needed to investigate the effects of these mutations.

## Conclusion

Genome analysis data of our study may play a significant role in understating interindividual variations in drug response and immune response by vaccines and variations in the pathogenicity, recurrence of infection, and mortality among nations and subgroups.

## Abbreviations

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2; E: Envelope; M: Membrane; N: Nucleocapsid; S: Spike glycoprotein; CDC: Centers for Disease Control and Prevention; NIID: Japan's Center of Infectious Disease; NCBI: National Center for Biotechnology Information; ACE2: Angiotensin-converting enzyme 2; ISG15: Interferon-stimulated gene 15; IRF3: Interferon factor 3; IFN-I: Interferon type I; MHC-I: Major Histocompatibility Complex Class I.



## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s43141-022-00431-3>.

**Additional file 1: Supplementary Figure 1.** Structural representation of SARS-CoV-2 and primer/probe sites. a) global target detection (primer/probe binding) sites and b) representation envelop and nucleocapsid region. The diversity sites were sourced from Hadfield et al. (2018).

**Additional file 2: Supplementary Table 1.** List of clones with mutation observed at the detection (primer/probe binding) sites. Positive sign indicates the presence of a single or multiple mutation near the 3' or 5'-end of the RT-PCR primer-probes.

**Additional file 3: Supplementary Table 2.** Represents the top variants at the 11 major mutant genes observed at least at > 100,000 global SARS-CoV-2 genomes. The top 11 major mutant genes with mutation distribution in >100,000 global samples were filtered out and analyzed for top mutant sites.

**Additional file 4.**

**Additional file 5.**

### Acknowledgements

We acknowledge the physicians from the originating medical facilities responsible for obtaining the specimen from patients and the authors and originating and submitting laboratories of the sequences from the National Center for Biotechnology Information (NCBI) database.

### Authors' contributions

AS and SP: Conceived the presented idea, carrying out analysis for tables, figures and drafting the manuscript. AS carried out processing the data, analysis, interpretation, and editing the manuscript; HH: Analyzed the data, and prepared figures; MB, EH, IS, TA, and ZAS: Contributed in critical revisions of the article. All authors read and approved the final version of this manuscript.

### Funding

The authors declare that this study did not receive any funding from any financial institutes. All the authors contributed to this study from social responsibility in response to SARS-CoV-2 pandemic situation.

### Availability of data and materials

All data generated or analyzed during this study are included in this published article ([supplementary information files](#)).

### Declarations

#### Ethics approval and consent to participate

Not applicable

#### Consent for publication

Not applicable

#### Competing interests

The authors declare that there is no competing interests.

#### Author details

<sup>1</sup>Department of Biochemistry and Molecular Biology, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Dhaka 1706, Bangladesh.

<sup>2</sup>Research and Development Department, Bioengineering Lab. Co., Ltd, 657 Nagatake Midori-ku, Sagami-hara-shi, Kanagawa-ken 252-0154, Japan. <sup>3</sup>Department of Biology, Faculty of Science and Engineering, Saitama University, Saitama, Japan. <sup>4</sup>Department of Agroforestry and Environment, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Dhaka 1706, Bangladesh. <sup>5</sup>Department of Biological Sciences and Chemistry, Southern University and A&M College, 244 William James Hall, Baton Rouge, LA 70813, USA. <sup>6</sup>Genesis Institute of Genetic Research, Genesis Healthcare Corporation, Yebisu Garden Place Tower 15F/26F 4-20-3 Ebisu, Shibuya-ku, Tokyo, Japan.

<sup>7</sup>Department of Biochemistry and Molecular Biology, University of Rajshahi, Rajshahi 6205, Bangladesh.

Received: 25 October 2021 Accepted: 12 October 2022

Published online: 01 November 2022

### References

- World Health Organization (2020) Coronavirus disease (COVID-19) Situation Report– 102, 01 Mai 2020. Data as received by WHO from national authorities by 10:00 CEST, 1 May 2020, World Health Organization Available from: [https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200501-covid-19-sitrep.pdf?sfvrsn=742f4a18\\_4](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200501-covid-19-sitrep.pdf?sfvrsn=742f4a18_4)
- Ksiazek TG, Erdman D, Goldsmith CS et al (2003) A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* 348:1953–1966
- Zhu N, Zhang D, Wang W et al (2020) A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 382:727–733. <https://doi.org/10.1056/NEJMoa2001017>
- Andersen KG, Rambaut A, Lipkin WI et al (2020) The proximal origin of SARS-CoV-2. *Nat Med* 26:450–452. <https://doi.org/10.1038/s41591-020-0820-9>
- Fan W, Zhao S, Bin Y et al (2020) A new coronavirus associated with human respiratory disease in China. *Nature* 579:265–269. <https://doi.org/10.1038/s41586-020-2008-3>
- Heymann DL, Shindo N (2020) WHO Scientific and Technical Advisory Group for Infectious Hazards. COVID-19: what is next for public health? *Lancet* 395:542–545. [https://doi.org/10.1016/S0140-6736\(20\)30374-3](https://doi.org/10.1016/S0140-6736(20)30374-3)
- Zhou P, Yang XL, Wang XG et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579(7798):270–273. <https://doi.org/10.1038/s41586-020-2012-7>
- Chen Y, Liu Q, Guo D (2020) Emerging coronaviruses: genome structure, replication, and pathogenesis. *J Med Virol* 92:418–423. <https://doi.org/10.1002/jmv.25681>
- Holmes KV, Enjuanes L (2003) The SARS coronavirus: a postgenomic era. *Science* 300:1377–1378. <https://doi.org/10.1126/science.1086418>
- Lai MMC (2003) SARS virus: the beginning of the unraveling of a new coronavirus. *J Biomed Sci* 10:664–675. <https://doi.org/10.1159/000074077>
- Marra MA, Jones SJ, Astell CR et al (2003) The genome sequence of the SARS-associated coronavirus. *Science* 300:1399–1404. <https://doi.org/10.1126/science.1085953>
- Nicholls JM, Poon LL, Lee KC et al (2003) Lung pathology of fatal severe acute respiratory syndrome. *Lancet* 361:1773–1778. [https://doi.org/10.1016/S0140-6736\(03\)13413-7](https://doi.org/10.1016/S0140-6736(03)13413-7)
- Azad GK (2021) Identification and molecular characterization of mutations in nucleocapsid phosphoprotein of SARS-CoV-2. *PeerJ* 9:e10666. <https://doi.org/10.7717/peerj.10666>
- Harvey WT, Carabelli AM, Jackson B et al (2021) SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 19:409–424. <https://doi.org/10.1038/s41579-021-00573-0>
- Haagmans BL, Osterhaus AD (2006) Coronaviruses and their therapy. *Antivir Res* 71(2-3):397–403. <https://doi.org/10.1016/j.antiviral.2006.05.019>
- Jin Z, Du X, Xu Y et al (2020) Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582(7811):289–293. <https://doi.org/10.1038/s41586-020-2223-y>
- Xia S, Duan K, Zhang Y et al (2020) Effect of an Inactivated Vaccine Against SARS-CoV-2 on Safety and Immunogenicity Outcomes: Interim Analysis of 2 Randomized Clinical Trials. *JAMA* 324(10):951–960. <https://doi.org/10.1001/jama.2020.15543>
- Liu J, Liu Y, Xia H et al (2021) BNT162b2-elicited neutralization of B.1.617 and other SARS-CoV-2 variants. *Nature* 596(7871):273–275. <https://doi.org/10.1038/s41586-021-03693-y>
- Xia S, Zhang Y, Wang Y et al (2021) Safety and immunogenicity of an inactivated SARS-CoV-2 vaccine, BBIBP-CorV: a randomised, double-blind, placebo-controlled, phase 1/2 trial. *Lancet Infect Dis* 21(1):39–51. [https://doi.org/10.1016/S1473-3099\(20\)30831-8](https://doi.org/10.1016/S1473-3099(20)30831-8)
- Yang H, Xie W, Xue X et al (2005) Design of wide-spectrum inhibitors targeting coronavirus main proteases. *PLoS Biol* 3:e324. <https://doi.org/10.1371/journal.pbio.0030324>

21. Tripathi N, Tripathi N, Goshisht MK (2022) COVID-19: inflammatory responses, structure-based drug design and potential therapeutics. *Mol Divers* 26(1):629–645. <https://doi.org/10.1007/s11030-020-10176-1>
22. Singh AK, Singh A, Singh R, Misra A (2020) Remdesivir in COVID-19: A critical review of pharmacology, pre-clinical and clinical studies. *Diabetes Metab Syndr* 14(4):641–648. <https://doi.org/10.1016/j.dsx.2020.05.018>
23. Orsini A, Corsi M, Santangelo A et al (2020) Challenges and management of neurological and psychiatric manifestations in SARS-CoV-2 (COVID-19) patients. *Neurol Sci* 41(9):2353–2366. <https://doi.org/10.1007/s10072-020-04544-w>
24. Keehner J, Horton LE, Pfeffer MA et al (2021) SARS-CoV-2 Infection after Vaccination in Health Care Workers in California. *N Engl J Med* 384(18):1774–1775. <https://doi.org/10.1056/NEJMc2101927>
25. Edler C, Klein A, Schröder AS, Sperhake JP (2021) Ondruschka B (2021) Deaths associated with newly launched SARS-CoV-2 vaccination (Comirnaty®). *Leg Med (Tokyo)* 51:101895. <https://doi.org/10.1016/j.legalmed.2021.101895>
26. Jacobson KB, Pinsky BA, Montez Rath ME et al (2021) Post-vaccination SARS-CoV-2 infections and incidence of the B.1.427/B.1.429 variant among healthcare personnel at a northern California academic medical center. medRxiv preprint. <https://doi.org/10.1101/2021.04.14.21255431>
27. Vogels CBF, Brito AF, Wyllie AL et al (2020) Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT-qPCR primer-probe sets. *Nat Microbiol* 5(10):1299–1305. <https://doi.org/10.1038/s41564-020-0761-6>
28. Tsutae W, Chaochaisit W, Aoshima H, Ida C, Miyakawa S, et al (2021) Detecting and Isolating False Negatives of SARS-Cov-2 Primers and Probe Sets among the Japanese Population: A Laboratory Testing Methodology and Study. *J Infect Dis Ther* 51:004. <https://doi.org/10.4172/2165-7386.s1.10004>
29. Siu YL, Teoh KT, Lo J et al (2008) The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. *J Virol* 82(22):11318–11330. <https://doi.org/10.1128/JVI.01052-08>
30. Lu S, Ye Q, Singh D et al (2021) The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat Commun* 12(1):502. <https://doi.org/10.1038/s41467-020-20768-y>
31. Dinesh DC, Chalupska D, Silhan J et al (2020) Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathog* 16(12):e1009100. <https://doi.org/10.1371/journal.ppat.1009100>
32. Choppin PW, Scheid A (1980) The Role of Viral Glycoproteins in Adsorption, Penetration, and Pathogenicity of Viruses. *Rev Infect Dis* 2(1):40–61. <https://doi.org/10.1093/clinids/2.1.40>
33. Ou X, Liu Y, Lei X, Purnell W et al (2021) Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV-2. *Nat Commun* 12(1):2144. <https://doi.org/10.1038/s41467-021-22614-1>
34. da Silva SJR, Alves da Silva CT, Mendes RPG, Pena L (2020) Role of nonstructural proteins in the pathogenesis of SARS-CoV-2. *J Med Virol* 92(9):1427–1429. <https://doi.org/10.1002/jmv.25858>
35. Fu YZ, Wang SY, Zheng ZQ et al (2021) SARS-CoV-2 membrane glycoprotein M antagonizes the MAVS-mediated innate antiviral response. *Cell Mol Immunol* 18(3):613–620. <https://doi.org/10.1038/s41423-020-00571-x>
36. Chen J, Malone B, Llewellyn E et al (2020) Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-Transcription Complex. *Cell*. 182(6):1560–1573.e13. <https://doi.org/10.1016/j.cell.2020.07.033>
37. Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R (2020) A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proof-reading and Final Capping. *Cells* 9(5):1267. <https://doi.org/10.3390/cells9051267>
38. Yuen CK, Lam JY, Wong WM et al (2020) SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists. *Emerg Microbes Infect* 9(1):1418–1428. <https://doi.org/10.1080/22221751.2020.1780953>
39. Huang Y, Yang C, Xu XF, Xu W, Shu-wen Liu S (2020) Structural and functional properties of SARS-CoV-2 spike protein: potential antiviral drug development for COVID-19. *Acta Pharmacol Sin* 41:1141–1149. <https://doi.org/10.1038/s41401-020-0485-4>
40. Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M (2020) Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* 369(6501):330–333. <https://doi.org/10.1126/science.abb9983>
41. Ou X, Liu Y, Lei X et al (2021) Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun* 11:1620. <https://doi.org/10.1038/s41467-020-15562-9>
42. Khan MT, Zeb MT, Ahsan H et al (2021) SARS-CoV-2 nucleocapsid and Nsp3 binding: an in silico study. *Arch Microbiol* 203(1):59–66. <https://doi.org/10.1007/s00203-020-01998-6>
43. Barretto N, Jukneliene D, Ratia K, Chen Z, Mesecar AD, Baker SC (2005) The papain-like protease of severe acute respiratory syndrome coronavirus has deubiquitinating activity. *J Virol* 79(24):15189–15198. <https://doi.org/10.1128/JVI.79.24.15189-15198.2005>
44. Lee JS, Shin EC (2020) The type I interferon response in COVID-19: implications for treatment. *Nat Rev Immunol* 20(10):585–586. <https://doi.org/10.1038/s41577-020-00429-3>
45. Li JY, Liao CH, Wang Q, Tan YJ, Luo R, Qiu Y, Ge XY (2020) The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus Res* 286:198074. <https://doi.org/10.1016/j.virusres.2020.198074>
46. Zhang Y, Zhang J, Chen Y et al (2020) The ORF8 protein of SARS-CoV-2 mediates immune evasion through potentially downregulating MHC-I. *bioRxiv*. <https://doi.org/10.1101/2020.05.24.111823>
47. Geng H, Subramanian S, Wu L et al (2021) SARS-CoV-2 ORF8 Forms Intracellular Aggregates and Inhibits IFN $\gamma$ -Induced Antiviral Gene Expression in Human Lung Epithelial Cells. *Front Immunol* 12:679482. <https://doi.org/10.3389/fimmu.2021.679482> eCollection 2021
48. Hsu JC, Laurent-Rolle M, Pawlak JB, Wilen CB, Cresswell P (2021) Translational shutdown and evasion of the innate immune response by SARS-CoV-2 NSP14 protein. *Proc Natl Acad Sci* 118(24):e2101161118. <https://doi.org/10.1073/pnas.2101161118>
49. Jang KJ, Jeong S, Kang DY (2020) A high ATP concentration enhances the cooperative translocation of the SARS coronavirus helicase nsP13 in the unwinding of duplex RNA. *Sci Rep* 10:4481. <https://doi.org/10.1038/s41598-020-61432-1>
50. Shu T, Huang M, Wu D et al (2020) SARS-Coronavirus-2 Nsp13 Possesses NTPase and RNA Helicase Activities That Can Be Inhibited by Bismuth Salts. *Virol Sin* 35:321–329. <https://doi.org/10.1007/s12250-020-00242-1>
51. Newman JA, Douangamath A, Yazdani S et al (2021) Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nat Commun* 12:4848. <https://doi.org/10.1038/s41467-021-25166-6>
52. Mu C, Sheng Y, Wang Q, Amin A, Li X, Xie Y (2021) Potential compound from herbal food of *Rhizoma Polygonati* for treatment of COVID-19 analyzed by network pharmacology: Viral and cancer signaling mechanisms. *J Funct Foods*. <https://doi.org/10.1016/j.jff.2020.104149>
53. Mu C, Sheng Y, Wang Q, Amin A, Li X, Xie Y (2020) Dataset of potential *Rhizoma Polygonati* compound-druggable targets and partial pharmacokinetics for treatment of COVID-19. *Data Brief* 33:106475. <https://doi.org/10.1016/j.dib.2020.106475>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.