Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Research paper

Clade GR and clade GH isolates of SARS-CoV-2 in Asia show highest amount of SNPs

Antara Sengupta a, Sk. Sarif Hassan b, Pabitra Pal Choudhury c

a Department of Computer Science and Engineering, University of Calcutta, West Bengal, India
b Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram, Paschim Medinipur, West Bengal, India
c Applied Statistics Unit, Indian Statistical Institute, Kolkata, West Bengal, India

ARTICLE INFO

Keywords: COVID-19
Clades
SARS-CoV-2
Deleterious mutations
Asia

ABSTRACT

Clades are monophyletic groups composed of a common ancestor and all its lineal descendants. As the propensity of virulence of a disease depends upon the type of clade the virus belongs to and it causes different fatality rates of disease in different countries, so the clade-wise analysis of SARS-CoV-2 isolates collected from different countries can illuminate the actual evolutionary relationships between them. In this study, 1566 SARS-CoV-2 genome sequences across ten Asian countries are collected, clustered, and characterized based on the clade they belong to. The isolates are compared to the Wuhan reference sequence hCoV-19/Wuhan/WIV04/19' to identify the mutations that occurred at different protein regions. Structural changes in amino acids due to mutations lead to functional instability of the proteins. Detailed clade-wise functional assessments are carried out to quantify the stability and vulnerability of the mutations occurring in SARS-CoV-2 genomes which can shed light on personalized prevention and treatment of the disease and encourage towards the invention of clade-specific vaccines.

1. Introduction

Viruses have a remarkable capacity to adapt to new hosts and environments (Sanjuán & Domingo-Calap, 2016). Mutations may lead to different phenotypic changes in them, which may lead to occur biodiversity. Phylogenies are frameworks for analysing biodiversity. Phylogenetic analysis based on sequence similarity is one of the very efficient ways to do so (Das et al., 2020). However, it will be worth noting that due to the recent outbreak of pandemic COVID-19, people around the world are trying by every means to reach the origin, to get some ways of prevention and therapeutic pathways. Biodiversity is characterized by a continual replacement of branches in the tree of life, i.e. clade (Silvestro et al., 2015). Evolutionary pressure on host immunodeficiency leads to different clades of viruses (Tyor et al., 2013). A clade is a group of highly related sequences that share a common ancestor. They can provide hypotheses about the actual evolutionary history of that group of sequences. Some clinical studies suggest that the proclivity of virulence of a disease depends upon the type of clade the virus belongs to (Tyor et al., 2013). Clade differences can result in varying degrees of pathology. Millions of gene regulatory elements are there which contribute heavily to the variation in gene expression of complex traits and diseases (Kumar et al., 2020a). Determining mutation types influence a lot in gene regulation and is important for studying the role of regulatory variation in evolution. Genomic evolution helps a virus to escape host immunity (Li et al., 2020a; DeDiego et al., 2008). The clade-wise analysis of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) isolates collected from different countries can shed a light on the actual evolutionary history of the region or continent. In order to confirm the hypothesis in Coronavirus disease 2019 (COVID-19) pathogenesis, it is highly recommended to make a thorough study of mutations occurring in SARS-CoV-2 isolates collected from different demographic areas and characterizing them based on the clades they come from (Andersen et al., 2020). A plethora of papers already have been published, where researchers have tried to study the virus isolates of SARS-CoV-2, which is solely responsible for the disease to occur in human (Zeng et al., 2004; Walls et al., 2020; Maitra et al., 2020; Biswas & Majumder, 2020; Kumar et al., 2020b). Huge numbers of investigations are reported in order to find evolutionary relationships between SARS-CoV-2 and other coronaviruses and to determine the origin and molecular characteristics of SARS-CoV-2 (DiMaio & Nathans, 1982; Banerjee et al., 2020; Foy et al., 2003). Several works are also done on characterization and comparative analysis of structured and non-structured proteins of SARS-CoV-2.

https://doi.org/10.1016/j.meegid.2021.104724
Received 4 December 2020; Received in revised form 4 January 2021; Accepted 11 January 2021
Available online 19 January 2021
1567-1348/© 2021 Elsevier B.V. All rights reserved.
Infection, Genetics and Evolution 89 (2021) 104724

2. Methods and materials

2.1. Collection of gene sequences of SARS-COV-2

One of the primary features of the investigation and analysis of the COVID-19 is availability of real-time data in global databases. To carry out the experiment We have collected DNA sequence of 1566 isolates of SARS-CoV-2 from ten different Asian countries from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/coronavirus) on October 20, 2020. The information about collected dataset are presented as Supplemental Materials in Table S1 and summarized in Table 1. After collecting all the DNA sequences, CoVsurver mutations App of GISAID (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app) is used to compare the mutated sequences with reference sequence to find out the mutations occurred in each isolates in different protein regions along with clades they belong to. In this said web site the reference sequence is hCoV-19/Wuhan/WIV04/19. However, information in detail about the dataset Collected sequences are then gone through preliminary screening for excluding noisy sequences. Here noise includes no mutations and the amino acid changes due to mutations specified by ‘X’. Thus finally 1371 isolates are taken for further investigations. It is worth to mention that like GISAID CoVsurver mutation app, CovidGC (https://covidgc.org/) is also an open resource from which we can track SARS-CoV-2 SNVs (single-nucleotide variations) (Chen et al., 2020). Starting from tracking evolution to immune interactions, identifying the clades and lineages present in a region at a specific period of time, evaluation of community outcomes after particular vaccines, or therapeutics are applied, searching for transient mutations to illuminate common mechanism of resistance to grow immunity. But it works based on the data collected from GISAID only. Whereas, CoVsurver mutation app of GISAID works on influenza data as a whole and accepts protein or nucleotide sequences in a FASTA file from other resources like NCBI GeneBank.

2.2. Methods

The present work aims to make a clade-wise classification and analysis of SARS-CoV-2 isolates of ten Asian countries. The isolates of each country are then compared with the reference sequence to find out the mutations that occurred. Clade-wise clustering of the given dataset is taken place. The observed mutations are then gone through different online software tools to investigate different biological functionalities that may change and affect the variants due to mutations. Here it is to be noted that we have used two web-based software tools (PROVEAN (Choi & Chan, 2015) and I-mutant (Caprioti et al., 2005)) for the aforesaid functional assessments.1-Mutant is a suite developed based on Support Vector Machine(SVM). (ΔΔG > −0.5 Kcal/mol) indicates that the mutation can largely destabilize the protein, ΔΔG >0.5 Kcal/mol indicates about the strong stability and −0.5 ≥ΔΔG ≥ 0.5 Kcal/mol tells about weak effect of mutations. Isolates with a score equal to or below −2.5 are

Table 1

| Country   | # Associated | #Isolate |
|-----------|--------------|----------|
| INDIA     | 570          | 565      |
| THAILAND  | 227          | 104      |
| BANGLADESH| 231          | 231      |
| IRAN      | 172          | 106      |
| CHINA     | 189          | 189      |
| JAPAN     | 96           | 96       |
| SAUDI ARABIA | 58      | 58       |
| PAKISTAN  | 10           | 9        |
| MALAYSIA  | 9            | 9        |
| SRILANKA  | 4            | 4        |
3. Results

3.1. Clade-wise clustering of SARS-CoV-2 strains taken as dataset from different countries

After excluding the noisy sequences finally 1371 isolates are found. Each strain belongs to a particular clade, so the isolates are clustered according to the clade from which they belong to. It has been observed that as a whole isolates of five clades (G, GH, GR, L, S, O, and V) are participated in those countries of the Asian continent. According to (Fig. 1) the order of the clade-wise participation of isolates is GH > GR > O > G > S > L > V. It is to be noted here that among the entire dataset taken Indian isolates hold a big amount of data. According to the country-wise view shown in Table 2 SARS-CoV-2 isolates of clade ‘O’ are present in the dataset of all countries and isolates of clade ‘V’ have been circulated only at China and Thailand. The country-wise analysis has a mixed result. In Sri Lanka, Thailand, China, Malaysia, and Iran the isolates are majorly from clade ‘O’ or ‘Other’. India and Saudi Arabia have a prevalence of clade ‘GH’. Pakistan, Bangladesh, and Japan have the prevalence of clade ‘GR’. It indicates viral diversity regarding infection as the infection is transmitting from one country to another. Remarkable viral diversities are also present even in different regions within a country too.

3.2. Investigating trend of mutations in various clades

In this subsection firstly the positions of mutations are identified in each isolate and then it is aimed to calculate clade-wise percentage of mutations occurred in each country as shown in (Fig. 2). Secondly, a microscopic view has been given on clade-wise clustering of total mutations found in the whole dataset and calculating the protein-wise percentage of the mutations occurred according to the clades they belong to (Fig. 3).

According to (Fig. 2), India and Saudi Arabia have a prevalence of clade ‘GH’ (55.76%, and 53.69% respectively). Whereas, in Pakistan, Bangladesh, and Japan strains have a prevalence of clade ‘GR’ (60.66%, 92.33% and 88.06% respectively). Mutations have occurred in strains from Clade ‘S’ at China and Thailand (51.63% and 45% respectively). SARS-CoV-2 isolates of clade ‘O’ have significant participation in Iran, Sri Lanka, and Malaysia (59.31%, 46.67% and 94.29% respectively). SARS-CoV-2 isolates of clades G, GH, GR, S, L, and O are circulating in India and Japan. Whereas, clades V, O, S, L, GR, and G are circulating at different regions of China and clades O, S, GR, GH, and G are circulated in Saudi Arabia. In Malaysia (clades O and GH) and Sri Lanka (clades O, GR, and G) the SARS-CoV-2 isolates do not have the viral diversities a lot.

Mutations refer to the virus to undergo certain changes which can lead to develop some new isolates after replications. Non-synonymous substitutions play a very significant role as this type of mutation makes change in amino acid. Alteration in amino acid causes structural change. With the aim of understanding the trend of non-synonymous mutations in different clades in the context of disease severity, a detailed protein-wise comparative analysis has been taken place. Mutations identified at different protein regions in all the isolates are shown at Table S2 in Supplementary file. To do so we have considered the total dataset as a whole. Clade-wise percentages of non-synonymous mutations at different protein regions are calculated. The clade-wise characterization of mutations of different proteins are shown in (Fig. 3).

According to the dataset taken, we have got 6665 numbers of non-synonymous mutations. We can observe at Table 3 that the chronological order of clades at per number of mutations taken place in whole dataset is GR > GH > G > O > S > L > V. It can be observed in (Fig. 3) that mutations are majorly taken place at isolates of clades GH and GR which are 31.33% and 31.93% of respectively. Samples of clade V have been affected rarely (0.29%). Clade-wise distribution of mutations in each protein does not have a very similar trend(s). Although the majority of proteins mutated are either of clade GR (N, NS7a, NSP2, NSP6, NSP13, and NSP15) or GH (M, NS3, NSP1, NSP12, NSP14, and NSP16) but clade G also has large numbers of mutations in some proteins (S, NS6, and NSP11). Isolates of Clade L and clade G here have got mutations maximum only in protein E and NSP4 respectively. In proteins NS7b, NS8, and NSP7 of the isolates from clade S have maximum distributions of non-synonymous mutations. In the isolates of clade G, GR, and O number of mutations at NSP10 are equal. Mutations are also equally distributed in protein NSP5 of clades GH and GR.

Microscopic view towards mutations occurred gives a list of mutations which are found in number of isolates, shown at Table S2 in supplementary file. Mutation D614G in Spike protein is found in huge numbers of isolates throughout all the 9 countries except China and mutation P323L in protein NSP12 is affected lots of isolates in all 8 countries except China and Iran. Few other largely occurred mutations are (G204R, R203K, S194L) in N protein, Q57H in NS3, L84S in NS8, (L54F, A262T, A829T) in Spike protein.

3.3. Quantitative assessment of functional changes occur due to mutations

Structural changes in amino acids due to mutations lead to create functional instability of the isolates themselves, cause vulnerable

---

| Table 2 Clade-wise counting of isolates reported in 10 countries. |
|--------------------------|----------|--------|--------|-------|--------|--------|--------|
| COUNTRY                  | G        | GH     | GR     | L     | S      | O      | V      |
| THAILAND                 | 12       | 0      | 0      | 1     | 39     | 51     | 1      |
| BANGLADESH               | 6        | 13     | 208    | 1     | 1      | 2      | 0      |
| SRILANKA                 | 1        | 0      | 1      | 0     | 0      | 2      | 0      |
| CHINA                    | 2        | 0      | 1      | 64    | 48     | 71     | 3      |
| JAPAN                    | 2        | 5      | 71     | 13    | 2      | 3      | 0      |
| INDIA                    | 191      | 303    | 18     | 7     | 19     | 27     | 0      |
| IRAN                     | 32       | 1      | 0      | 4     | 0      | 69     | 0      |
| MALAYSIA                 | 0        | 1      | 0      | 0     | 8      | 0      | 0      |
| PAKISTAN                 | 0        | 1      | 5      | 2     | 0      | 1      | 0      |
| SAUDI ARABIA             | 2        | 32     | 22     | 0     | 1      | 1      | 0      |
diseases and even increase the magnitude of virulence. In this subsection, we have tried to find the impact of single point mutations on the biological function of proteins of each isolates through the light of PROVEAN (Protein Variation Effect Analyzer) score, which may be deleterious or neutral (Choi et al., 2012). We have also calculated the change in Gibbs free energy ($\Delta \Delta G$) occur due to single point mutations as the difference in folding free energy change between wild type and mutant protein ($\Delta \Delta G$) is considered as an impact factor of protein stability changes (Capriotti et al., 2008). The motivation here is to understand the effect of those mutations on protein stability. The quantitative analysis will give an insight into the probable mutations that occur in a particular clade and the magnitude of virulence of them. It is to be noted

Fig. 2. Calculating clade-wise percentage of mutations occurred in different countries.

Fig. 3. Quantitative analysis of clade-wise non-synonymous mutation. (a) Percentage of non-synonymous mutations found in different clades; (b) Clade-wise percentage of non-synonymous mutations occurred in different protein regions.
that here we have excluded the mutations which are occurred only once. The deleterious mutations are shown in Table 4. It is observed at Table 4 that if we consider the dataset as a whole, then among structural proteins the mutations occurred in spike protein(s) are more deleterious than others. Among the accessory proteins, NS3 is affected the most. NSP2, NSP5, and NSP12 are the non-structural proteins that have most of the deleterious mutations that occurred. Furthermore, we have calculated clade-wise percentage of deleterious mutations that occurred in different protein regions. To do so, we have segregated each deleterious mutation occurred in ten different countries along with their clades Table 5. The (Fig. 4) shows the protein regions that are mostly affected by the deleterious mutations. Maximum deleterious mutations occurred in structural and accessory proteins belong to clade GH. Most of the deleterious mutations in non-structural protein regions are occurred in the isolates of both clades GH and GR. The isolates of clade V are rare and only found in the isolates of China and Thailand, but interestingly it is observed that few of deleterious mutations are also enlisted there. (Fig. 4) depicts the fact that most of the deleterious mutations take place in amino acid sequences of clade GH.

Table 6 gives us a microscopic view of the severity of the mutations that occurred in the dataset taken. In 82% of deleterious mutations protein stability has been decreased due to single point mutation. It is already observed that maximum mutations have occurred in the isolates which belong to clade GH. Out of 18 deleterious mutations happened in isolates from clade GH in 15 isolates (S194L, D936Y, P863H, W161L, F366V, E309A, A1914D, G309C, L75F, H64N, Q57H, N257D, E95K, P45L, V104F) stability have been decreased due to mutations. It can be observed at Table 7 that due to mutations majorly amino acid Glutamine and Serine are affected. Glutamine(Q) has been changed to Histidine(H), and Serine(S) changed to Leucine(L).

### 4. Discussion

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is continuously changing its characteristics and degree of infectivity.

#### Table 3
Clade-wise segregation of mutated data.

| PROTEIN | G | GH | GR | L | S | O | V |
|---------|---|----|----|---|---|---|---|
| E       | 2 | 1  | 4  | 9 | 2 | 2 | 0 |
| M       | 6 | 15 | 13 | 0 | 2 | 3 | 0 |
| N       | 60 | 264 | 709 | 13 | 31 | 52 | 0 |
| NS3     | 30 | 376 | 709 | 12 | 10 | 24 | 3 |
| NS6     | 1  | 1  | 3  | 0 | 1 | 47 | 0 |
| NS7a    | 4  | 3  | 1  | 1 | 12 | 0 | 2 |
| NS7b    | 0  | 1  | 0  | 6 | 0 | 0 | 0 |
| NS8     | 2  | 18 | 18 | 1 | 69 | 4 | 0 |
| NSP1    | 4  | 11 | 11 | 9 | 2 | 9 | 0 |
| NSP10   | 4  | 3  | 4  | 0 | 1 | 4 | 0 |
| NSP11   | 2  | 1  | 1  | 0 | 13| 1 | 0 |
| NSP12   | 241 | 383 | 366 | 9 | 18 | 45 | 3 |
| NSP13   | 24 | 21 | 26 | 7 | 3 | 12 | 0 |
| NSP14   | 49 | 109 | 20 | 4 | 5 | 10 | 0 |
| NSP15   | 22 | 10 | 31 | 5 | 5 | 11 | 0 |
| NSP16   | 20 | 23 | 9  | 11| 1 | 11 | 0 |
| NSP2    | 52 | 82 | 245 | 9 | 11| 20 | 4 |
| NSP3    | 109 | 114 | 156 | 27 | 21 | 82 | 0 |
| NSP4    | 50 | 8  | 10 | 16| 9 | 11 | 0 |
| NSP5    | 16 | 24 | 21 | 4 | 3 | 0 | 0 |
| NSP6    | 3  | 25 | 32 | 4 | 4 | 26 | 7 |
| NSP7    | 0  | 1  | 0  | 3 | 1 | 0 | 0 |
| NSP8    | 2  | 7  | 8  | 1 | 14| 7 | 0 |
| NSP9    | 4  | 4  | 1  | 0 | 1 | 0 | 0 |
| S       | 217 | 583 | 284 | 18| 26| 599 | 1 |
| TOTAL MUTATION | 1024 | 2088 | 2128 | 150 | 265 | 991 | 19 |

#### Table 4
Investigate the deleterious mutations in total dataset.

| PROTEIN | DELETERIOUS MUTATION | PROTEIN | DELETERIOUS MUTATION |
|---------|----------------------|---------|----------------------|
| S       | D936Y                | NSP2    | TB51                 |
| S       | L752R                | NSP2    | F368V                |
| S       | A280P                | NSP2    | G339S                |
| S       | CS15W                | NSP2    | P129S                |
| S       | D820N                | NSP2    | E309A                |
| S       | E725K                | NSP3    | A1914D               |
| S       | G646R                | NSP3    | P1558L               |
| S       | G669R                | NSP4    | G309C                |
| S       | L533K                | NSP5    | P108S                |
| S       | L916F                | NSP5    | I106S                |
| S       | P863H                | NSP5    | L75F                 |
| S       | R905S                | NSP5    | N142L                |
| S       | S758I                | NSP5    | R279C                |
| S       | S875F                | NSP5    | G277S                |
| S       | T716P                | NSP6    | H64N                 |
| S       | T874P                | NSP6    | A1914D               |
| S       | V534G                | NSP12   | D309G                |
| N       | S193I                | NSP12   | W161L                |
| N       | S194L                | NSP13   | R392C                |
| N       | R191L                | NSP14   | T113I                |
| N       | S180L                | NSP15   | M3307                |
| NS3     | G573H                | NSP16   | R2071                |
| NS3     | G251V                | NSP16   | P125R                |
| NS3     | D1555                | NSP12   | T8031                |
| NS3     | G172C                | NSP12   | M1241                |
| NS3     | G172V                | NSP12   | P2275                |
| NS3     | N257D                | NSP12   | S607I                |
| NS3     | G251V                | NSP12   | A979V                |
| NS3     | W45R                 | NSP16   | S607I                |
| NS7a    | E95K                 | NSP16   | S607I                |
| NS7a    | P45L                 | NSP16   | A979V                |
| NS7a    | V104F                | NSP16   | A979V                |
| NS3     | P38R                 | NSP16   | A979V                |
Mutations can strengthen its severity and infectivity. It is an important observation that although both the mutations P323L in protein region NSP12 and D614G in Spike protein are mostly occurred mutations in the whole dataset taken and is globally dominant. Although the mutations change the structural stability of proteins by making changes in free energy but they are not deleterious. It reveals that the mutations are not deleterious but have great impact on the biological functions of the proteins respectively. According to the papers reported, G614 variant has higher transmission ability than the D614 variant, associated with low RT-PCR cycle thresholds, but does not increase disease severity (Korber et al., 2020). In other hand the P323L mutation is located in the interface domain of the RNA-dependent RNA polymerase (RdRp) and it changes the intramolecular interactions in the protein as its stability changes. Among the remaining most frequently observed non-synonymous mutations, S194L in N protein and Q57H in NS3 are deleterious. Clade-wise analysis of mutations performed here in this present study indicates that in Asian countries SARS-CoV-2 isolates responsible for COVID-19 majorly belong to the clades GR and GH. Among them mutations that occurred in isolates of clade GH are deleterious in nature, so have an impact on the biological function of proteins. The mutations also change the structural stability of proteins by making changes in free energy ($\Delta G$). It is reported that the human genome may carry large numbers of deleterious mutations which as a whole make a significant contribution to fatal diseases. Identification and analysis of deleterious mutations can shed light on personalized treatment and medicine (Chun & Fay, 2009). Hence, the identification of these kinds of Table 5
Clade-wise clustering of deleterious mutations in total dataset.

| MUTATION | G | GH | GR | L | S | O | V |
|----------|---|----|----|---|---|---|---|
| S193I    | 1 | 7  | 0  | 0 | 0 | 0 | 0 |
| S194L    | 21| 190| 0  | 0 | 2 | 0 | 0 |
| R191L    | 2 | 0  | 0  | 0 | 0 | 0 | 0 |
| S180L    | 1 | 1  | 0  | 0 | 0 | 0 | 0 |
| D936Y    | 1 | 1  | 0  | 0 | 0 | 0 | 0 |
| L732R    | 0 | 0  | 0  | 0 | 2 | 0 | 0 |
| A288P    | 0 | 0  | 0  | 0 | 3 | 0 | 0 |
| C851W    | 1 | 0  | 0  | 0 | 1 | 0 | 0 |
| D620N    | 0 | 0  | 0  | 0 | 2 | 0 | 0 |
| E725K    | 0 | 0  | 0  | 0 | 2 | 0 | 0 |
| G648R    | 1 | 0  | 0  | 0 | 3 | 0 | 0 |
| G669R    | 0 | 0  | 0  | 0 | 4 | 0 | 0 |
| L533K    | 0 | 0  | 0  | 0 | 2 | 0 | 0 |
| L916F    | 1 | 0  | 0  | 0 | 1 | 0 | 0 |
| P863H    | 4 | 15 | 0  | 0 | 0 | 0 | 0 |
| R905S    | 2 | 0  | 0  | 0 | 0 | 0 | 0 |
| S758I    | 2 | 0  | 0  | 0 | 0 | 0 | 0 |
| S875F    | 1 | 0  | 0  | 0 | 2 | 0 | 0 |
| T716F    | 0 | 0  | 0  | 0 | 2 | 0 | 0 |
| T874P    | 0 | 0  | 0  | 0 | 3 | 0 | 0 |
| V334G    | 0 | 0  | 0  | 0 | 2 | 0 | 0 |
| Q57H     | 12| 342| 7  | 0 | 7 | 0 | 0 |
| G251V    | 0 | 0  | 0  | 0 | 5 | 4 | 0 |
| D155Y    | 0 | 2  | 5  | 0 | 0 | 0 | 0 |
| G172C    | 0 | 0  | 3  | 0 | 0 | 0 | 0 |
| G172V    | 0 | 0  | 3  | 0 | 0 | 0 | 0 |
| N275D    | 1 | 1  | 0  | 0 | 0 | 0 | 0 |
| W45R     | 0 | 0  | 7  | 0 | 0 | 0 | 0 |
| E95K     | 1 | 1  | 0  | 0 | 0 | 0 | 0 |
| S45L     | 0 | 2  | 0  | 0 | 0 | 0 | 0 |
| A97V     | 0 | 0  | 3  | 0 | 22| 0 |

Fig. 4. Quantification of clade-wise deleterious non-synonymous mutation. (a) Clade-wise percentage of deleterious non-synonymous mutations in different protein regions; (b) Clade-wise calculations of deleterious non-synonymous mutations of total data set.
isolates of clade GR, and 31.33% of the mutations from GH. Hence, it indicates the diversity of the infection indeed. O, GH, and GR are the most prevalent clades. When clades G, GH, and GR traversed almost in all countries specified here, the isolates of clade V are affected rarely. The most frequently mutated amino acids are Glutamine and Serine. In most of the cases glutamine is changed into Histidine and serine is changed to glutamine. It is to be noted that both the mutations are deleterious and the isolates of clade GH carry the major deleterious mutation load (44.19% of the total dataset). The majority of mutations taken place in the isolates of clade GH are deleterious in nature. 82% of deleterious mutations are unstable and so their biological functions are affected. As a whole in this present work, the investigation provides us clade-wise characteristics of the SARS-CoV-2 isolates of the Asian continent. When reported research papers shed the light on development of clade-specific vaccines (Yi et al., 2020), our analysis can encourage drug designers for development of customized drugs or vaccines for Asian continent in order to combat COVID-19.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2021.104724.

Credit authorship contribution statement

Antara Sengupta: Investigation, Conceptualization, Data curation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. Sk. Sarif Hassan: Conceptualization, Visualization, Validation, Writing - review & editing. Pabitra Pal Choudhury: Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 6
Investigating Stability of deleterious mutations in total dataset.

| Mutation | Stability | ΔΔG |
|----------|-----------|-----|
| S194L    | Deleterious Decreased | −0.47 |
| D936Y    | Deleterious Decreased | −0.35 |
| L752R    | Deleterious Increased  | 0.05 |
| G851W    | Deleterious Increased  | 0.31 |
| D850N    | Deleterious Decreased | −1.25 |
| E725K    | Deleterious Increased  | 0.35 |
| G648R    | Deleterious Increased  | 0.06 |
| G669R    | Deleterious Decreased | −0.14 |
| L533K    | Deleterious Increased  | 0.81 |
| L916F    | Deleterious Increased  | 0.12 |
| P863H    | Deleterious Increased  | 1.44 |
| R905S    | Deleterious Increased  | 1.18 |
| T716F    | Deleterious Decreased | −0.8 |
| T874P    | Deleterious Increased  | 0.22 |
| V534G    | Deleterious Decreased | −2.029 |
| Q57H     | Deleterious Decreased | −0.9 |
| G251V    | Deleterious Decreased | −0.54 |
| G172C    | Deleterious Decreased | −0.83 |
| G172V    | Deleterious Decreased | −0.41 |
| N257I    | Deleterious Increased  | 0.21 |
| W45R     | Deleterious Increased  | 1.05 |
| E95K     | Deleterious Increased  | 0.58 |
| P45L     | Deleterious Increased  | 0.71 |
| V161F    | Deleterious Decreased | −1.47 |
| P38R     | Deleterious Decreased | −0.9 |
| W161L    | Deleterious Increased  | 0.51 |
| R392C    | Deleterious Increased  | 1.29 |
| T113I    | Deleterious Increased  | 0.43 |
| M330T    | Deleterious Decreased | −1.0 |
| R287I    | Deleterious Decreased | −0.66 |
| P12S     | Deleterious Decreased | −1.29 |

Table 7
Amino Acid changed due to deleterious mutations taken place.

| MUTATION | COUNT | MUTATION | COUNT |
|----------|-------|----------|-------|
| Q > H    | 368   | M > I    | 3     |
| S > L    | 215   | P > R    | 3     |
| A > V    | 25    | R > I    | 3     |
| P > H    | 19    | S > F    | 3     |
| P > S    | 16    | C > W    | 2     |
| G > R    | 12    | D > G    | 2     |
| G > V    | 12    | D > N    | 2     |
| S > I    | 12    | E > A    | 2     |
| P > L    | 9     | E > L    | 2     |
| T > I    | 9     | F > V    | 2     |
| D > Y    | 9     | H > N    | 2     |
| W > R    | 7     | I > S    | 2     |
| G > C    | 6     | L > K    | 2     |
| A > D    | 5     | L > R    | 2     |
| L > F    | 5     | L > S    | 2     |
| N > L    | 5     | M > T    | 2     |
| T > P    | 5     | N > D    | 2     |
| E > K    | 4     | R > L    | 2     |
| R > C    | 4     | R > S    | 2     |
| A > P    | 3     | V > F    | 2     |
| G > S    | 3     | V > G    | 2     |

number of mutations are really high in the isolates belong to both the clades. When clades G, GH, and GR traversed almost in all countries specified here, the isolates of clade V are affected rarely. The most frequently mutated amino acids are Glutamine and Serine. In most of the cases glutamine is changed into Histidine and serine is changed to glutamine. It is to be noted that both the mutations are deleterious and the isolates of clade GH carry the major deleterious mutation load (44.19% of the total dataset). The majority of mutations taken place in the isolates of clade GH are deleterious in nature. 82% of deleterious mutations are unstable and so their biological functions are affected. As a whole in this present work, the investigation provides us clade-wise characteristics of the SARS-CoV-2 isolates of the Asian continent. When reported research papers shed the light on development of clade-specific vaccines (Yi et al., 2020), our analysis can encourage drug designers for development of customized drugs or vaccines for Asian continent in order to combat COVID-19.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2021.104724.

5. Conclusions

The in silico analysis performed in this study states that the isolates in ten Asian countries are from clades G, GH, GR, L, S, O, and V. It indicates the diversity of the infection indeed. O, GH, and GR are the most widely affected ancestors of isolates among them. But when there is a talk about mutations, 31.93% of total mutations have taken place in the isolates of clade GR, and 31.33% of the mutations from GH. Hence, mutations in SARS-CoV-2 isolates and their impacts on the host body seek attention of virologists.
