Saudi Journal of Biological Sciences 29 (2022) 3494–3501

Contents lists available at ScienceDirect

Saudi Journal of Biological Sciences

journal homepage: www.sciencedirect.com

Original article

In-silico genomic landscape characterization and evolution of SARS-CoV-2 variants isolated in India shows significant drift with high frequency of mutations

Eltayib H. Ahmed-Abakur [a,b], Mohammad Fahad Ullah [a,b], Elmutuz H. Elssaig [a,b], Tarig M.S. Alnour [a,b,c,*]

[a] Department of Medical Laboratory Technology (FAMS), University of Tabuk, P.O. Box 741, Tabuk 71411, Saudi Arabia
[b] Prince Fahad Research Chair, University of Tabuk, P.O. Box 741, Tabuk 71411, Saudi Arabia
[c] Faculty of Medical Laboratory Science, Department of Microbiology and Immunology, Alzaiem Alazhari University, Khartoum North 11111, Sudan

ARTICLE INFO

Article history:
Received 2 January 2022
Revised 27 January 2022
Accepted 20 February 2022
Available online 25 February 2022

Keywords:
In-silico
SARS - CoV- 2
Mutations, synonymous
Non-synonymous
FAB1a/b
S gene

ABSTRACT

In-silico studies on SARS-CoV-2 genome are considered important to identify the significant pattern of variations and its possible effects on the structural and functional characteristics of the virus. The current study determined such genetic variations and their possible impact among SARS-CoV-2 variants isolated in India. A total of 546 SARS-CoV-2 genomic sequences (India) were retrieved from the gene bank (NCBI) and subjected to alignment against the Wuhan variant (NC_045512.2), the corresponding amino acid changes were analyzed using NCBI Protein-BLAST. These 546 variants revealed 841 mutations; most of these were non-synonymous 464/841 (55.1%), there was no identical variant compared to the original strain. All genes; coding and non-coding showed nucleotide changes, most of the structural genes showed frequent nonsynonymous mutations. The most affected genes were ORF1a/b followed by the S gene which showed 515/841 (61.2%) and 120/841 (14.3%) mutations, respectively. The most frequent non-synonymous mutation 486/546 (89.01%) occurred in the S gene (structural gene) at position 23,403 where A changed to G leading to the replacement of aspartic acid by glycine in position (D614G). Interestingly, four variants also showed deletion. The variants MT800923 and MT800925 showed 12 consecutive nucleotide deletion in position 21982–21993 resulting in 4 consecutive amino acid deletions that were leucine, glycine, valine, and tyrosine in positions 141, 142, 143, and 144 respectively. The present study exhibited a higher mutations rate per variant compared to other studies carried out in India.

1. Introduction

Since the last recorded exceptionally virulent "Spanish Flu" pandemic (1918–20) which infected approximately five hundred million people resulting in fifty million deaths across the world (Kilbourne, 2006), the recent "COVID-19" pandemic of SARS-CoV-2 has appeared as an enormous burden and a global health challenge. In an unprecedented scale in the recorded history, the pandemic has left billions of people seeking to healthcare facilities or economically and socially affected, with more than 2.5 million human lives lost worldwide within a year (Ashour et al., 2020; Seoane 2021). The causative agents responsible for these pandemics and several other epidemics are coronaviruses, which are structurally enveloped with a single-stranded ~ 30 kb RNA genome (positive-sense), that predominantly accommodates six ORFs (Open Reading Frames); and present influenza-like symptoms and show crown-like appearance with the presence of glycoprotein spikes on its envelope (Fehr and Perlman 2015, Di Gennaro et al 2020). The species comprises a large group of viruses that has the ability to infect a broad range of hosts, including humans, animals, and avians (Masters 2006). These include four genera- alpha, beta, gamma, and delta coronaviruses (CoVs) and as reported in epidemiological and etiological shreds of evidence, only alpha-CoVs and beta-CoVs possess the potential to infect humans (Masters 2006). Most of the highly pathogenic human CoVs such as SARS-CoV, MERS-CoV, and the current SARS-CoV-2 are Beta-CoVs, which are highly transmissible and associated with high

* Corresponding author at: Department of Medical Laboratory Technology (FAMS), University of Tabuk, P.O. Box 741, Tabuk 71411, Saudi Arabia.
E-mail address: telnour@ut.edu.sa (T.M.S. Alnour).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.sjbs.2022.02.030
1319-562X/

mortality rates (Ashour et al., 2020, Fehr and Perlman 2015). It is understood that due to the natural selection within the host, these viruses are transmitted from one animal to another and often undergo transmission from animal to humans "zoonotic transmission". Moreover, it could end up with human-to-human transmission due to further natural selection expressed through high rates of genomic instability in the viral RNA. As the novel variants emerge, these compete for dominance, thus resulting in multiple waves of infections (Kirby 2021). SARS-CoV-2 spread estimation from the published literature shows that the basic reproduction number (R0) extends from 2.2 to 5.7 (population-dependent) which is higher than seasonal influenza, and indicates a substantial potential for human-to-human transmission within a community (Sanche et al. 2020). The clinical features of COVID-19 and risk factors are highly variable, thus making the severity of the disease differ from asymptomatic infections to mild symptoms, extending to severe lethal forms, which are associated with severe acute pneumonia, accompanied by respiratory distress, leading to death (Uddin et al. 2020).

The studies on genome analysis show that SARS-CoV-2 shares about 79.5% identity with SARS-CoV and nearly 96% with bat SARS-like CoVs (Zhu et al. 2019). The SARS-CoV-2 viral genome size is 29.8 kb − 29.9 kb and the genome landscape expresses specific genes that are universally present in all known coronaviruses (Wang et al. 2020). The genomic landscape has more than two-thirds of the length at the 5′ end that houses 5′UTR (265 nucleotides) and orf1ab (21290 nucleotides) which encode polyproteins that are processed into 16 non-structural proteins. Whereas the 3′ end which is one-third of the genome, shelters 3′UTR (229 nucleotides), and genes that encode important structural proteins like surface (S) gene (3822 nucleotides), envelope (E) gene (228 nucleotides), membrane (M) gene (669 nucleotides), and nucleocapsid (N) gene (908 nucleotides). There are also six accessory proteins, encoded by the SARS-CoV-2 genome, including ORF3a (828 nucleotides), ORF6 (186 nucleotides), ORF7a (366 nucleotides), ORF7b (132 nucleotides), ORF8 (193 nucleotides), and ORF10 (117 nucleotides) genes (Khailany et al. 2020).

Among all the organisms, the viruses exhibit one of the highest mutation rates and RNA viruses are known to mutate more frequently than DNA viruses due to the error-prone RNA polymerase (Sanjuan and Domingo-Calap 2016). The genomic instability has given rise to genetic variations that have been shown to demonstrate a strong association between time, place, rate of infection, and mortality with the accumulation of genetic diversity (Wang et al 2020). The emergence of new variants possesses a high risk to the success of containment strategies, and thus studies on the genomic variation of SARS-CoV-2 are significant for identifying alterations that are critical for pathogenesis, disease course, prevention, and treatment. In the present study, we intended to characterize genetic variability of the viral population with the analysis of 546 genomic sequences of SARS-CoV-2 strains from India (March 2020-September 2020) which were retrieved from the National Center for Biotechnology Information (USA). The study assumes significance as it covers all the variants during the first wave till the peak was observed in September-2020 and has the potential to highlight the continuous evolution of the heterogeneity in the viral genome which might be a contributory factor to a more hostile second wave of the viral pandemic in India.

2. Material and methods

This research is an in silico, a cross-sectional study covering the NCBI published SARS COVID-19 complete genome sequences during the 1st peak of infection in India within March-September 2020. The study is part of a project based on the objective to analyze the genetic variations of SARS-CoV-2 and the heterogeneity of the evolving viruses according to the geographical location (province) and isolation timeline. All the SARS-CoV-2 sequences published from India during the study period were retrieved from NCBI Covid-19 Resource Repository (https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/). A total of 554 sequences were enrolled in the study; 8 were excluded due to incomplete genome and/or fragmented sequence. The rest of the sequences were aligned with the first characterized isolate, the Wuhan strain from China (Reference sequence Accession number NC_045512.2) (Lu et al. 2020).

Nucleotides mutation and gene variation from Wuhan standard reference were reported after studying the pairwise alignment against the reference sequence using the Basic Local Alignment Search Tool (BLASTn)-NCBI Nucleotide-BLAST, and the corresponding amino acid changes in protein were analyzed using Basic Local Alignment Search Tool (BLASTp)-NCBI Protein-BLAST.

Evolutionary relationships: The phylogenetic tree was obtained using the Neighbor-Joining method to understand the evolutionary relationship between the reference strain and the evolving variants of SARS-CoV-2 (Saitou and Nei, 1987). The clustered taxa are closely associated in the bootstrap test (500 replicates), and these have been reported as a percentage of replicate trees displayed next to the branches (Felsenstein 1985). The evolutionary distances were estimated by the Maximum Composite Likelihood method (Tamura et al. 2004) and show the number of base substitutions/site units. This analysis involved a set of SARS-CoV-2 variants reported in India's early months of the first pandemic wave. All the enigmatic positions were disregarded with a pairwise deletion option for each sequence pair. A total of 29,903 positions were present in the final dataset. MEGA X software for Evolutionary analyses was utilized to achieve the phylogenetic relationship. Two phylogenetic trees were built; timeline (collection date) phylogenetic trees where one SARS-CoV-2 isolate per month was randomly selected (no isolate reported for February and October 2020 in NCBI whereas April /May 2021 reported partial genomic sequence) and provincial (Geo location) phylogenetic trees, one SARS-CoV-2 isolate per province was randomly selected (Most genomic sequences were submitted from Gujarat whereas other provinces have either no or least representation in the NCBI).

3. Results

A total of 570 SARS - CO V- 2 sequences isolated in India were retrieved from the NCBI gene bank. Twenty-Four sequences were excluded from the study due to incompletion or fragmentation of the sequences. The remaining 546 were subjected to alignment against the Wuhan variant (NC_045512.2), the Indian variants showed homology of 99.90 to 99.99 %, there was no identical variant compared to the original strain.

The 546 sequences of SARS- CoV- 2 variants revealed 841 mutations; most of them 464/841 (55.1%) were non-synonymous (missense point mutation), the most affected genes were ORF1a/b followed by the S gene which demonstrated 515/841 (61.2%) and 120/841(14.3%) mutations respectively, whereas the ORF7b was the least affected gene as it displayed only 4/841(0.48%) mutations (Table 1). Non-synonymous mutations occurred more frequently in ORF1a/b- 294/515 (57.1%); the NSP3, NSP14, NSP10, and NSP13 were most affected regions in which they showed 175(24.07%), 147(20.22%), 107(14.72%), 88(12.1%) mutations respectively while NSP 7, NSP9 and NSP5 had fewer mutations (Table 2).

All genes (coding and non-coding) showed nucleotide changes, the notable observation is that most of the structural genes showed frequent nonsynonymous mutations i.e. nonsynonymous mutation, except the E gene which showed only 5 mutations with

Table 1
Type and frequency of mutations among SARS- CoV- 2 variants isolated in India.

	5′UTR	ORF 1a/b	S gene	ORF3a	E gene	M gene	ORF6	ORF7a	ORF7b	ORF8	N gene	ORF 10	3′ UTR	None coding region	Total
synonymous		220	48	14	3	13	4	7	2	4	22	1			338
nonsynonymous mutation		294	70	32	2	7	2	3	2	8	43	1			464
Frameshift mutation		1	2												3
Silent in the 5′UTR	12														12
Silent in the 3′ UTR													20		20
None coding region														4	4
Total	12	515	120	46	5	20	6	10	4	12	65	2	20	4	841

Table 2
Non-synonymous mutations among ORF1a/b gene of SARS- CoV- 2 variants isolated in India.

Gene name	Gene location	Mutation	
		Frequency	Percentage %
NSP1	266–805	12	1.66
NSP2	806–2719	52	7.15
NSP3	2720–8554	175	24.07
NSP4	8555–10054	39	5.36
Nsp5	10055–10972	11	1.51
NSP6	10973–11842	22	3.03
NSP7	11843–12091	1	0.14
NSP8	12092–12685	12	1.65
NSP9	12686–13024	5	0.69
NSP10	13025–16236	107	14.72
NSP11	13442–13480	–	–
stem-loop 1	13476–13503	–	–
stem-loop 2	13488.0.13542	–	–
NSP13	16237–18039	88	12.1
NSP14	18040–19620	147	20.22
NSP15	19621–20658	26	3.58
NSP16	20695–21552	30	4.13

2 of them as nonsynonymous mutations. The most affected gene was the S gene, exhibiting 70/120 (58.3%) nonsynonymous mutations followed by the N gene, which exhibited 65/841 (7.7%) of total mutations, with the majority of these 43/65 (66.2%) being nonsynonymous mutations. Additionally, 12/841 (1.43%) and 4/841 (0.48%) of the mutations occurred in 5′UTR and 3′UTR respectively as fewer mutations were detected in non-coding regions (Table 1).

The most common synonymous mutation 493/546 (90.3%) occurred in the 5′UTR region at the position C241T, followed by several common mutations in ORF1 a/b at the positions C3037T, C17788T, and C2836T which displayed the following frequencies- 456/546 (83.5%), 297/546 (54.4%), and 216/546 (39.6%) respectively. Some other common synonymous mutations were also observed such as in the M gene at the position C26735T and the S gene at the position C22444T occurring as 288/546 (52.7%) and 233/546 (42.7%), respectively) Table 3). Fig. (1a & 1b) show the corresponding amino acid changes for non-synonymous mutations among Indian variants of SARS - CO V- 2 (only the mutations which occurred in more than one variant have been listed here; for a complete list refer to the supplementary file). The most frequent non-synonymous mutation 486/546 (89.01%) occurred in the S gene at position 23,403 where A changed to G leading to the replacement of aspartic acid by glycine in position (D614G)) Table 3 and Fig. 1b). The second most dominant non-synonymous mutation 465/546 (85.2%) occurred in ORF1 a/b at position 14,408 where C changed to T which changed the amino acid in RNA-dependent RNA polymerase (RdRp/ P323L)) Table 3 and Fig. 1a). The third common mutation- 299/546 (54.8%) occurred in ORF3a at position 25,563 where G changed to T leading to- Q75H, followed by a mutation in N gene at position 28,854 where C changed to T leading to S194L) Table 3 and Fig. 1b). The mutations C241T, A23403G, C14408T, and C3037T appeared together in 83 % of Indian SARC-CoV-2 variants.

The common nucleotides changes were Cytosine to Thymine (C-T) followed by Guanine to Thymine (G-T), which appeared in 330/841 (39.2%) and 181/841 (21.5%) of the variants respectively, while the least observed nucleotide change was Cytosine to Guanine (C-G) 5/841 (0.59%) Fig. 2.

Four variants showed frameshift mutation (deletion); variant MT451876 which was isolated in Ahmedabad showed nucleotide deletion in ORF1 a/b at the position 669–671 which lead to deletion of Tyrosine 136. The two variants (MT800923 and MT800925) which were isolated in Daskroi showed 12 consecutive nucleotide deletions in position 21982–21993 resulting in 4 consecutive amino acid deletions that were leucine, glycine, valine, and tyrosine in positions 141, 142, 143, and 144 respectively. The fourth variant (MT012098) isolated in Kerala State displayed deletion at nucleotides number 21991–21993 leading to deletion of single amino acid tyrosine at position 144 (Table 2 and Fig. 1b).

Fig. 3a and 3b show the phylogenetic trees based on timeline and provincial criteria. The phylogenetic trees showed that the variants MT012098.1 and MT050491.1 were isolated in Kerala state and represented the first isolated SARS-CoV-2 in India (both variants isolated on January 31, 2020) clustered closely to the original Wuhan strain. However, there was a subsequent drift in the genomic landscape of the virus locally, leading to several genetic variants.

4. Discussion

The ongoing spread of the disease and rapid transmission of SARS-CoV-2 raised serious inquiries concerning the adaptation and evolution of the viral population which are effectively

Table 3
Common mutation among SARS- CoV- 2 variants isolated in India.

	N gene	M gene	ORF3a	S gene	ORF1 a/b	5′UTR	Gene			
Mutation site	*C28854T*	*C26735T*	*G25563T-*	*A23403G*	*C22444T*	*C18877T*	*C14408T*	*C3037T*	*C2836T*	*C241T*
Mutation type	S194L	Synonymous	Q57H	D614G	Synonymous	Synonymous	P4715L	Synonymous	Synonymous	5′UTR
Frequency	220/546	288/546	299/546	486/546	233/546	297/546	465/546	456/546	216/546	493/546
Percent	40.30%	52.70%	54.80%	89.01%	42.70%	54.40%	85.20%	83.50%	39.60%	90.30%

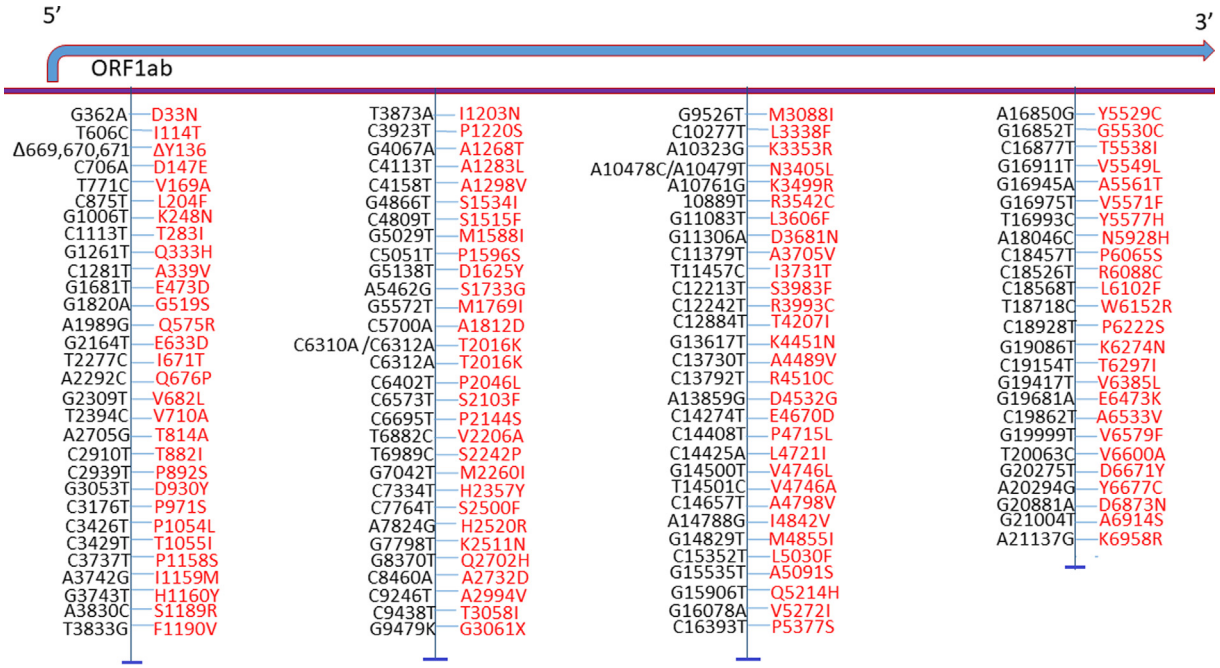

Fig. 1. a. The genomic landscape of ORF1 a/b, of Indian SARS - CO V- 2 variants representing non-synonymous mutations and deletions along with the corresponding amino acid changes. b. The genomic landscape of S gene, ORF3a, M gene, orf7a, orf8 and N gene of Indian SARS-COV-2 variants representing non-synonymous mutations and deletions along with the corresponding amino acid changes.

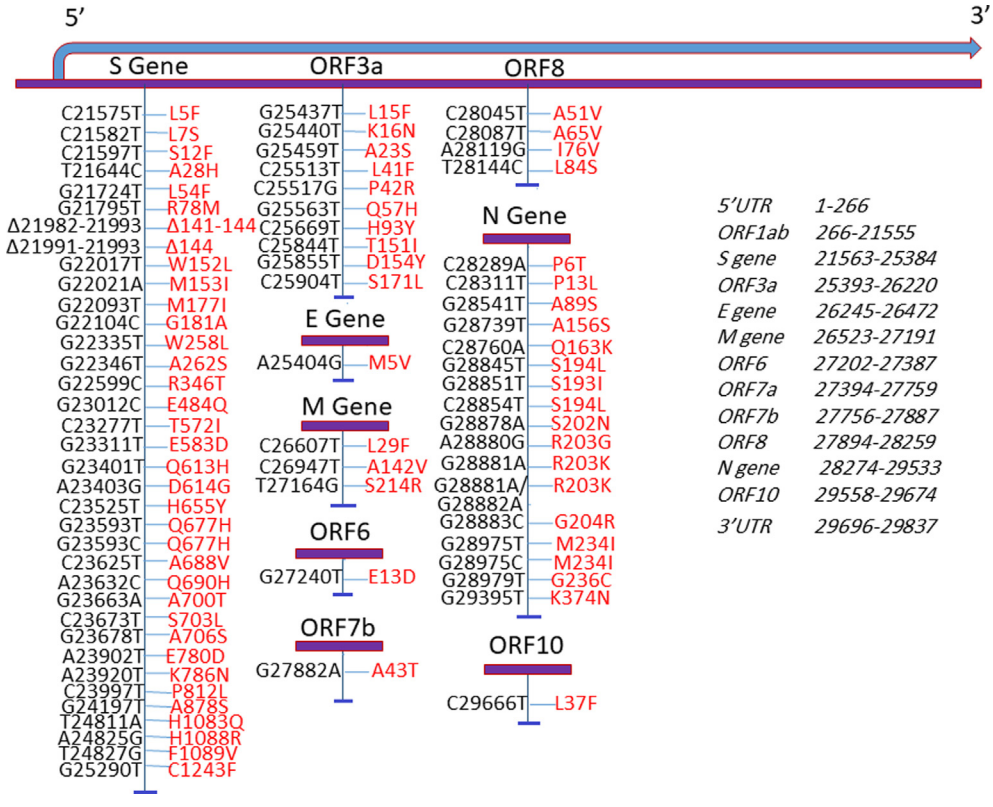

Fig. 1 (*continued*)

governed by mutations, recombination, and/or deletions (Islam et al. 2020), *In-silico* study on SARS-CoV-2 genomic variations is of utmost importance to identify the significant pattern and its possible effect on the structural and functional characteristics of the virus. Since India is one of the most affected countries by the COVID 19 infection (Das et al. 2021), the present study determined

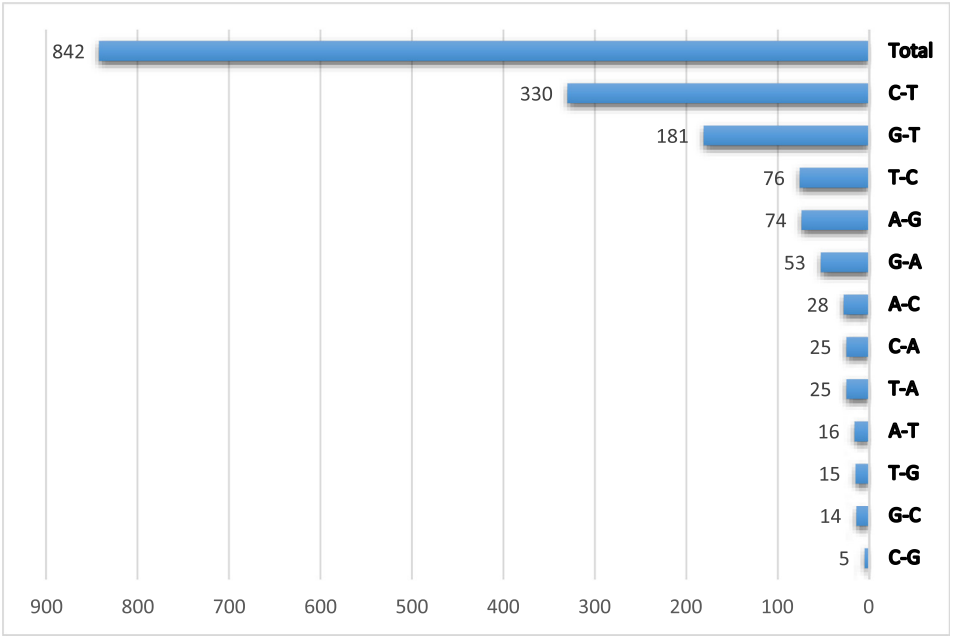

Fig. 2. The frequency of nucleotide changes among Indian SARS-CoV-2 variants.

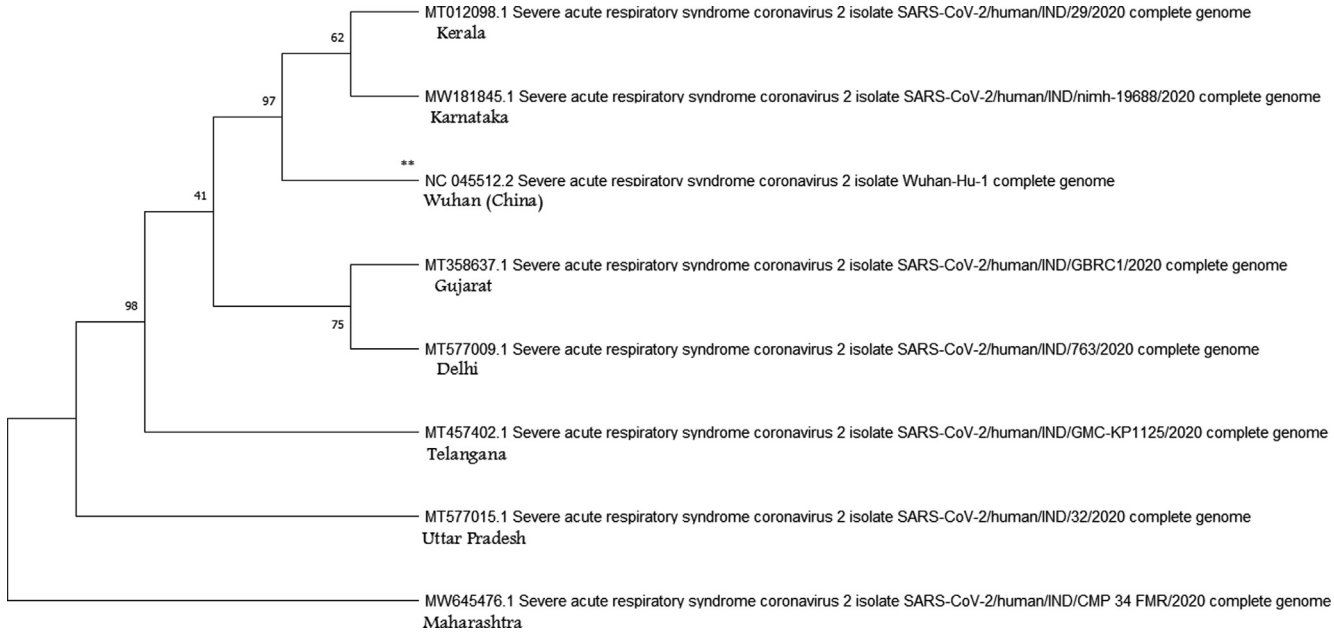

Fig. 3. a. Phylogenetic affiliation of Indian SARS-CoV-2 variants compared to Wuhan strain based on provincial location. b. Phylogenetic affiliation of Indian SARS-CoV-2 variants compared to Wuhan strain based on timeline.

the genetic variations and their possible impact among SARS-CoV-2 variants extracted in India.

Our study showed that the Indian SARS- Co– V-2 isolates had a homology of 99.90 to 99.99 %, with no identical variant compared to the original Wuhan strain. The earlier publications showed high homology and identical variants (Wang et al. 2020, Khailany et al. 2020, Lu et al. 2020, Ahmed-Abakur and Alnour 2020). The disappearance of identical variants indicates that the original strain is no longer in the pathogenicity race. The emergence of new variants with potential pathogenicity such as delta variant (CDC 2021) has emphasized that the SARS- Co– V-2 may develop critical mutations. The present study exhibited a higher mutations rate per variant 841/546 compared to other studies carried out in India. Raghav et al., 2020 analyzed 202 SARS-CoV-2 isolates and they have reported 247 single-nucleotide variants, Das et al., 2021 analyzed 463 genomes and stated 536 mutated positions within the coding regions, and Khailany et al., 2020 showed 116 mutations from the analysis of 95 genomes of SARS-CoV-2; these variations may be attributed to sampling size and time of collection.

Our results showed that both kinds of mutations (synonymous and non-synonymous) appeared in all the genes, the most affected gene was ORF1a/b followed by S genes while the ORF7b was the

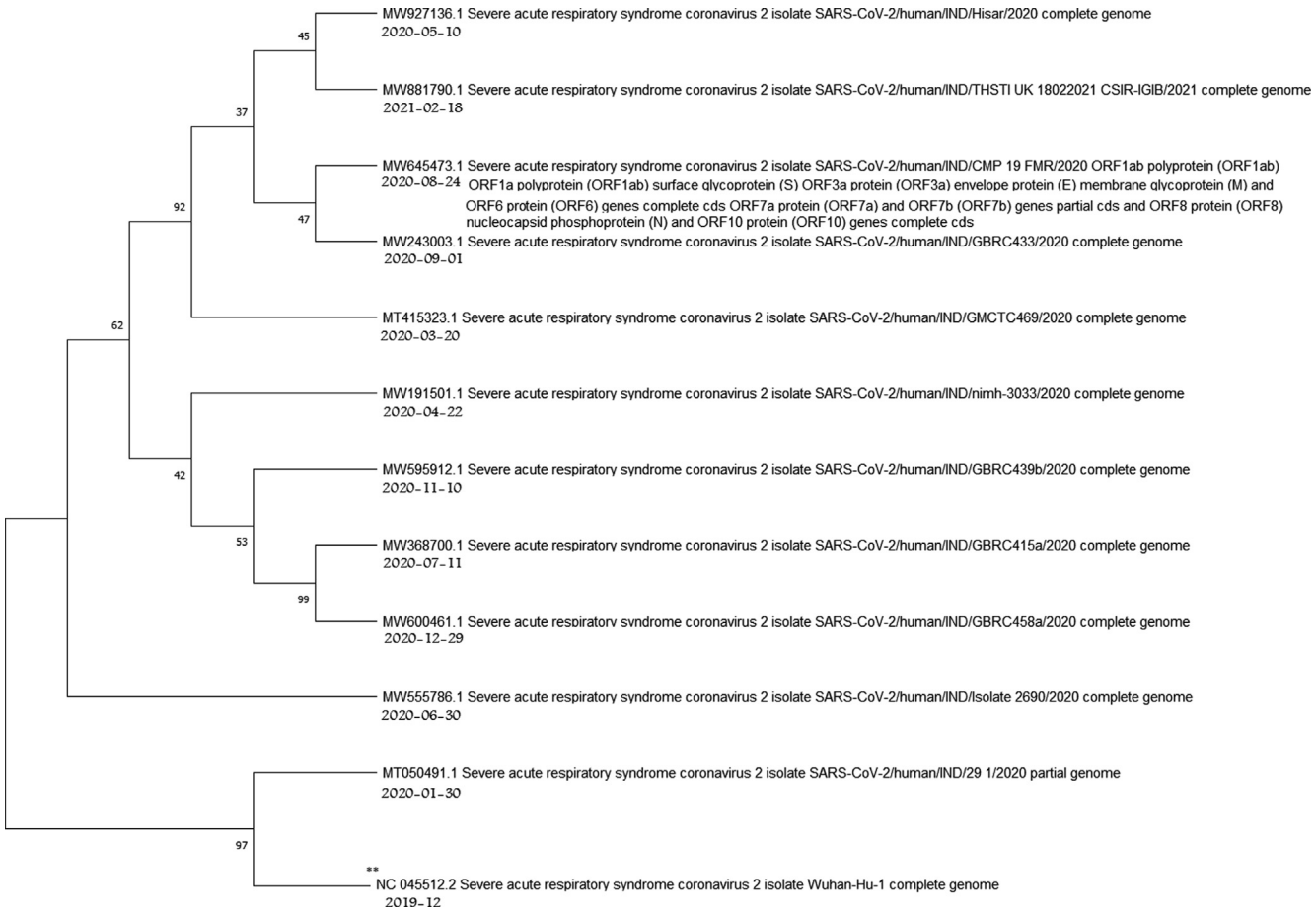

Fig. 3 (*continued*)

least affected gene. These findings are in alignment with some recent studies including (Shishir et al., 2021) who showed that ORF1ab was occupied with more than 60% of mutations and relatively agreed with (Tiwari and Mishra 2020), who found that the most prevalent mutation occurred in the spike region followed by ORF1a/b (Nsp2, Nsp3) and the N gene, while contradictory to (Das et al., 2021), who stated that mutations in ORF7b and E proteins were 100% non-synonymous and ORF3a gene was most affected gene followed by the N gene.

Our data displayed that the non-synonymous mutations occurred more frequently in ORF1a/b; the NSP3, NSP10, NSP13, and NSP14 regions were the most affected regions while NSP 7 and NSP9 had very few mutations. Interestingly, these results disagreed with (Almubaid and Al-Mubaid, 2021) who analyzed and compared 1200 genomes of the SARS-CoV-2 and reported that the ORF1a region contained mostly synonymous mutations, and the NSP13 gene is highly conserved, they proposed utilizing the NSP13 to produce treatments and inhibitors for SARS-CoV-2. The discrepancy in the mutations rate, type, and site may indicate the unique evolution of SARS-Co V-2 variants isolated in Indian. Such genetic discrepancy in SARS-CoV-2 sequences may be owing to location and host viral evolution in the infected patients due to varying immune responses (Al-Qaaneh et al 2021; Islam et al 2020). Generally, RNA viruses have a great mutation rate that might reach up to a million times within the hosts due to the lack of appropriate proofreading activities during viral replication (Al-Qaaneh et al. 2021; Wang et al. 2020). Specifically, 10 thousand single nucleotide polymorphisms were identified among the SARS-CoV-2, with anticipated rates of mutation between (0.0001–0.01) changes per nucleotide site per cell infection (Al-Qaaneh et al 2021).

The present study showed that the most affected structural gene was S followed by N and M genes, with many nonsynonymous mutations observed among these structural genes except in the E gene. These findings matched with (Islam et al. 2020) who expect possible adaptation at the nucleotides and amino acids, providing a window of heterogeneity structure in the viral proteins, specifically in the S protein, and disagreed with (Shishir et al., 2021) who reported that ORF7b and E encoded envelope protein did not carry any mutation. However, the most liable proteins were ORF1ab, Spike (S), Nucleocapsid (N), ORF3a, ORF8a, and ORF7a whereas fewer numbers of mutations occurred in ORF10, ORF7b, ORF6, and E proteins (Das et al. 2021). The highly frequent non-synonymous mutations in the S gene may be one of the factors that influenced the epidemic of the second wave in India. The spike protein plays vital roles in attachment, fusion, entry of the virus into the host cell and persistent mutation in this region might show that the virus is developed to be more competent in human-to-human transmission (Almubaid and Al-Mubaid, 2021). Pascarella et al., 2021 studied the SARS-CoV-2B.1.617 Indian variants and reported a noticeable change of the surface electrostatic potential of the receptor-binding domain of S protein, and proposed that these changes can aid in the interaction between the SARS-CoV-2 and the negatively charged ACE2, thus increasing the transmission of the disease. Since the S protein is the core targeted protein in covid-19 vaccines, mutations in this protein may reduce

the efficiency of the vaccine (Dai and Gao, 2021). Nevertheless, N and M proteins are responsible for inhibition of IFN production in hosts and suppression of interferon regulatory protein 3 stimulation, respectively (Alhusseini et al. 2021).

Our study showed that the most common synonymous mutation occurred in the non-coding region at the position C241T, followed by several synonymous mutations in ORF1 a/b, at the positions C3037T and C17788T. These mutations do not affect protein function or structure but may be prevalent because they affect human-to-human transmission or they might help the virus to cloak itself in the host body. The most frequent nonsynonymous mutation in our study occurred in the S gene at the position A23403G where A changed to G leading to the replacement of aspartic acid by glycine in position (D614G). This mutation occurs on the S2 domain that is important for the split of S1 by TMPRSS2 enzyme to aid the fusion of the viral spike protein with the host cell membrane and thus may enhance the spread and infectivity of the virus (Raghav et al., 2020). Alterations in the receptor-binding site of the viral spike protein propose that variants were implausible to decrease binding affinity with ACE2 (Koyama et al. 2020). The second most dominant nonsynonymous mutation occurred in ORF1 a/b at position C14408T which changed the amino acid in RNA-dependent RNA polymerase and may influence the replication rate of the viruses. This site is the target of anti-viral medications favipiravir and remdesivir; the susceptibility for changes indicates that drug-resistant variants may develop rapidly (Koyama et al. 2020), as the mutations enable the virus to overcome the immune system and promote drug resistance (Al-Qaaneh et al 2021). The third common mutation occurred in ORF3a at position G 25563 T (Q75H) followed by a mutation in the N gene at position C28854T (S194L). The N gene could be a potential candidate after the S gene for determining the role of mutation in viral pathogenesis and assembly (Khan et al., 2020). Similar results concerning the abundant nonsynonymous mutations at positions C241T and C3037T have been reported by numerous authors (Ahmed-Abakur and Alnour 2020, Almubaid and Al-Mubaid, 2021, Koyama et al. 2020). In alignment with our findings Koyama et al., 2020 studied 10,022 SARS CoV-2 sequences from different countries. They found that C3037T was the most common synonymous mutation and C14408 > T mutation is the second common mutation, Saha et al., 2020 observed A23403G (D614G) in 60% of Indian isolate, Das et al, 2021 observed it in 93% out of 77 Indian variants. A23403G (D614G) was observed in West European states in the early stage of the disease (Korber et al. 2020) and this finding might point towards the source of infection from which it was introduced in India. Accordingly, Raghav et al., 2020 reported Southeast Asia and Europe as the main routes for the introduction of the Covid 19 in India. Moreover, the present study showed that A23403G (D614G) co-occurred with three other mutations that were C241T, C14408T, and C3037T, a similar pattern of co-existence was reported by (Korber et al., 2020), while (Shishir et al., 2021) observed the co-existent of C14408T and A23403G. The nonsynonymous mutation G25563T (Q75H) was reported in Bangladesh and USA (Shishir et al., 2021). Our results disagreed with Khailany et al., 2020 who studied the mutations among SARS-CoV-2 collected from various locations and reported C29095T in the N gene, T28144C in the ORF8 gene, and C8782T in the ORF1ab gene as the most frequent mutations. This variation might be attributed to the difference in the sample size, time of collection, and location.

Our study showed three deletions, one in ORF1 a/b at the positions 669–671 which led to delete Tyrosine 136, and two in the S gene wherein 12 consecutive nucleotide deletions occurred at positions 21982–21993 resulting in 4 consecutive amino acid deletion of leucine, glycine, valine, and tyrosine at the positions 141, 142, 143 and 144, respectively. The second deletion was detected at nucleotides number 21991–21993 leading to deletion of single amino acid tyrosine at position 144. Numerous reports figure out deletions in SARS-CoV-2 throughout the viral genome, often the deletion occurred in accessory proteins and the non-structural genes that may have a direct influence upon viral infectivity (Islam et al. 2020). Shishir et al., 2021 reported 19 isolates that had lost a significant part of their genome and stated that the deletions were related to non-structural proteins and probably affected definite viral properties. Islam et al., 2020 analyzed 2,492 genomes of SARS-CoV-2 strains and concluded that viral genome deletions are a normal phenomenon, mainly inevitably associated with the weakening of the virus, and sometimes linked to the severity of infection. Phan, 2020 reported three deletions in the genomes of the SARS-CoV-2 from Australia, USA, and Japan, one deletion in the 3′ end of the genome, and two deletions were in the ORF1a/b polyprotein. Holland et al., 2020 found 81-nucleotide deletions in the SARS-CoV-2 AZ-ASU2923 genome that occurred in the ORF7a gene, resulting in a 27 amino acid deletion. There is a lack of research linking the deletions that have occurred in the entire genome of SARS-CoV-2 worldwide. Studying and linking the entire deletions may contribute to recognizing the pathogenic dynamics of the virus over time (Islam et al., 2020). Concerning the base changes our study showed that C > T mutation was the most prevalent mutation flowed by G > T, this result is in alignment with most of the earlier reports (Khailany et al., 2020, Almubaid and Al-Mubaid, 2021, Koyama et al., 2020).

Provincial and timeline phylogenetic trees showed that the variants MT012098.1 and MT050491.1 represented the first isolated SARS-CoV-2 in India (both were isolated in Kerala state on January 31, 2020) were clustered close to the original Wuhan strain and subsequently diversified. This finding pointed to multiple sources of infection in India as another study published previously considered Gujarat state as one of the main routes of disease transmission during the first peak of disease, with Southeast Asia and Europe as two major routes for introduction of the disease in India, followed by local transmission (Raghav et al., 2020). However, it needs to be mentioned that one of the earlier phylogenetic studies from different geographical locations showed that all the variants were clustered together in a single clad as compared to the original SARS-CoV strain (Khan et al. 2020).

Funding

The authors declare that this research was NOT funded partially or completely by any company or agency.

Research involving Human Participants and/or Animals

As our study depends on the COVID – 19 sequences uploaded in the NCBI, there was NO involvement of humans or animals in this study.

7. Informed consent

Not applicable.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sjbs.2022.02.030.

References

Ahmed-Abakur, E.H., Alnour, T.M.S., 2020. Genetic variations among SARS-CoV-2 strains isolated in China. Gene Rep. 21, 100925. https://doi.org/10.1016/j.genrep.2020.100925.

Alhusseini, N.K., Sajid, M.R., Alsheikh, H.A., Sriwi T.H., Odeh, N.B., Elshaer, R.E., Altamimi, R.E., Cahusac, P.M.B., 2021. Evaluation of COVID-19 myths in Saudi Arabia. Saudi Med. J. 42, 4, 377–383. http://doi.org/10.15537/smj.2021.42.4.20200706.

Almubaid, Z., Al-Mubaid, H., 2021. Analysis and comparison of genetic variants and mutations of the novel coronavirus SARS-CoV-2. Gene Reports 23, 101064. https://doi.org/10.1016/j.genrep.2021.101064.

Al-Qaaneh, A.M., Alshammari, T., Aldahhan, R., Aldossary, H., Alkhalifah, Z.A., Borgio, J.F., 2021. Genome composition and genetic characterization of SARS-CoV-2. Saudi J. Biol. Sci. 28 (3), 1978–1989. https://doi.org/10.1016/j.sjbs.2020.12.053.

Ashour, H.M., Elkhatib, W.F., Rahman, M.M., Elshabrawy, H.A., 2020. Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks. Pathogens 9 (3), 186. https://doi.org/10.3390/pathogens9030186.

Centers for disease control and preventation (CDC). Delta Variant: What We Know About the Science. https://www.cdc.gov/coronavirus/2019-ncov/variants/delta-variant.html. Accessed on 26.08.2021.

Dai, L., GAO, G.F., 2021. Viral targets for vaccines against COVID-19. Nat Rev Immunol., 2, 73–82. http://doi.org/10.1038/s41577-020-00480-0.

Das, J.K., Sengupta, A., Choudhury, P.P., Roy, S., 2021. Characterizing genomic variants and mutations in SARS-CoV-2 proteins from Indian isolates. Gene Rep. 25, 101044. https://doi.org/10.1016/j.genrep.2021.101044.

Di Gennaro, F., Pizzol, D., Marotta, C., Antunes, M., Racalbuto, V., Veronese, N., Smith, L., 2020. Coronavirus Diseases (COVID-19) Current Status and Future Perspectives: A Narrative Review. Int. J. Environ. Res. Publ. Health 17 (8), 2690.

Fehr, A.R., Perlman, S., 2015. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol. Biol. 1282, 1–23.

Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39 (4), 783–791. https://doi.org/10.1111/j.1558-5646.1985.tb00420.x.

Holland, L.A., Kaelin, E.A., Maqsood, R., Estifanos, B., Wu, L.I., Varsani, A., Halden, R. U., Hogue, B.G., Scotch, M., Lim, E.S., 2020. An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). J. Virol. 94 (14), e00711–e720. https://doi.org/10.1128/JVI.00711-20.

Islam, M.R., Hoque, M.N., Rahman, M.S., Alam, A.S.M.R.U., Akther, M., Puspo, J.A., Akter, S., Sultana, M., Crandall, K.A., Hossain, M.A., 2020. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci. Rep. 10 (1), 14004. https://doi.org/10.1038/s41598-020-70812-6.

Khailany, R.A., Safdar, M., Ozaslan, M., 2020. Genomic characterization of a novel SARS-CoV-2. Gene Rep. 19, 100682. https://doi.org/10.1016/j.genrep.2020.100682.

Khan, M.I., Khan, Z.A., Baig, M.H., Ahmad, I., Farouk, A.-E., Song, Y.G., Dong, J.-J., Ashraf, G.M., 2020. Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An in silico insight. PLoS One 15 (9), e0238344. https://doi.org/10.1371/journal.pone.0238344.

Kilbourne, E.D., 2006. Influenza pandemics of the 20th century. Emerg. Infect. Dis. 12 (1), 9–14.

Kirby, T., 2021. New variant of SARS-CoV-2 in UK causes surge of COVID-19. Lancet Respir. Med. 9 (2), e20–e21.

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., Hastie, KM., Parker, M.D., Partridge, D.G., Evans, C.M., Freeman, T.M., de Silva, T.I., Sheffield COVID-19 Genomics Group, McDanal C,, Perez, L.G., Tang, H., Moon-Walker, A., Whelan, S. P., LaBranche, C.C., Saphire, E.O., Montefiori, D.C., Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020, 182, 4, 812–827.e19. http://doi.org/10.1016/j.cell.2020.06.043.

Koyama, T., Platt, D., Parida, L., 2020. Variant analysis of SARS-CoV-2 genomes. Bull. World Health Organ. 98 (7), 495–504. https://doi.org/10.2471/BLT.20.253591.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W.J., Wang, D., Xu, W., Holmes, E.C., Gao, G.F., Wu, G., Chen, W., Shi, W., Tan, W., 2020. Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet, 22, 395,10224, 565–57410. http://doi.org/1016/S0140-6736(20)30251-8.

Masters, P.S., 2006. The molecular biology of coronaviruses. Adv. Virus Res. 66, 193–292.

Pascarella, S., Ciccozzi, M., Zella, D., Bianchi, M., Benedetti, F., Benvenuto, D., Broccolo, F., Cauda, R., Caruso, A., Angeletti, S., Giovanetti, M., Cassone, A., 2021. SARS-CoV-2 B.1.617 Indian variants: Are electrostatic potential changes responsible for a higher transmission rate? J. Med. Virol. 93 (12), 6551–6556.

Phan, T., 2020. Genetic diversity and evolution of SARS-CoV-2. Infect. Genet. Evol. 81, 104260.

Raghav, S., Ghosh, A., Turuk, J., Kumar, S., Jha, A., Madhulika, S., et al., 2020. Odisha COVID-19 Study Group; ILS COVID-19 Team, Pati, S., Parida, A., Analysis of Indian SARS-CoV-2 Genomes Reveals Prevalence of D614G Mutation in Spike Protein Predicting an Increase in Interaction With TMPRSS2 and Virus Infectivity. Front, Microbiol., 11, 594928. http://doi.org/10.3389/fmicb.2020.594928.

Saha, I., Ghosh, N., Maity, D., Sharma, N., Sarkar, J.P., Mitra, K., 2020. Genome-wide analysis of indian sars-cov-2 genomes for the identification of genetic mutation and snp. Infection. Genet. Evol. 85, 104457.

Saitou, N., Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

Sanche, S., Lin, Y., Xu, C., Romero-Severson, E., Hengartner, N., Ke, R., 2020. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg. Infect. Dis. 26 (7), 1470–1477.

Sanjuan, R., Domingo-Calap, P., 2016. Mechanisms of viral mutation. Cell Mol. Life Sci. 73 (23), 4433–4448.

Seoane, B., 2021. A scaling approach to estimate the age-dependent COVID-19 infection fatality ratio from incomplete data. PLoS ONE. 16 (2), e0246831.

Shishir, T.A., Naser, I.B., Faruque, S.M., 2021. In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. PLoS One 16 (1), e0245584. https://doi.org/10.1371/journal.pone.0245584.

Tamura, K., Nei, M., Kumar, S., 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA 101 (30), 11030–11035. https://doi.org/10.1073/pnas.0404206101.

Tiwari, M., Mishra, D., 2020. Investigating the genomic landscape of novel coronavirus (2019-nCoV) to identify non-synonymous mutations for use in diagnosis and drug design. J. Clin. Virol. 128, 104441. https://doi.org/10.1016/j.jcv.2020.104441.

Uddin, M., Mustafa, F., Rizvi, T.A., Loney, T., Suwaidi, H.A., Al-Marzouqi, A.H.H., Eldin, A.K., Alsabeeha, N., Adrian, T.E., Stefanini, C., Nowotny, N., Alsheikh-Ali, A., Senok, A.C., 2020. SARS-CoV-2/COVID-19: Viral Genomics, Epidemiology, Vaccines, and Therapeutic Interventions. Viruses 10. 12(5):526.

Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., Zhang, Z., 2020. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J. Med. Virol. 92 (6), 667–674. https://doi.org/10.1002/jmv.25762.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Ta,n W., 2020. China Novel Coronavirus Investigating and Research Team. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N. Engl. J. Med., 382, 8, 727–733. http://doi.org/10.1056/NEJMoa2001017.