Current Research in Microbial Sciences 2 (2021) 100065

Contents lists available at ScienceDirect

Current Research in Microbial Sciences

journal homepage: www.sciencedirect.com/journal/current-research-in-microbial-sciences

A next generation sequencing (NGS) analysis to reveal genomic and proteomic mutation landscapes of SARS-CoV-2 in South Asia

Tousif Bin Mahmood, MethodologyFormal analysisWriting – original draft [a], Ayan Saha, ValidationWriting – original draftWriting – review & editing [b,c], Mohammad Imran Hossan, Formal analysisWriting – original draft [a], Shagufta Mizan, Writing – original draftWriting – review & editing [d], S M Abu Sufian Arman, Formal analysis [a], Afrin Sultana Chowdhury, ConceptualizationWriting – review & editingSupervision [a,*]

[a] Department of Biotechnology and Genetic Engineering, Noakhali Science and Technology University, Noakhali 3814, Bangladesh
[b] Department of Genetic Engineering and Biotechnology, East West University, Dhaka 1212, Bangladesh
[c] Faculty of Medicine, Children's Cancer Institute, University of New South Wales, Australia
[d] Department of Genetic Engineering and Biotechnology, University of Chittagong, Chattogram 4331, Bangladesh

ARTICLE INFO

Keywords:
SARS-CoV-2
South Asian country
Single nucleotide polymorphisms
NSP2 protein
Transmission linkage

ABSTRACT

Counts for SARS-CoV-2 associated infections and fatalities are on the rise globally even in regions which contained the spread momentarily. The pattern of infections has been found to be controlled by the distinctive selection pressures exerted by fluctuating environmental nature and hosts. A total of 410 whole-genome sequences submitted by the South Asian countries were retrieved from the GISAID database and analyzed to assess the impact and pattern of mutations in this region. Most common and frequent mutations in the South Asian countries are 241C > *T*, 3037C > *T*, 14408C > *T*, and 23403A > *G* and about 85% SNPs are localized in ORF1ab, spike protein, and nucleocapsid. Among the identified mutations, the proportion of missense type (54.17%) was highest, followed by the synonymous (41.66%) and the non-coding types (4.17%). While analyzing transmission source in terms of geolocation, the largest clustered group from the South Asian countries was based on the G-clade (D614G) (81.7%; 335/410 samples), tracing the inception and transmission of SARS-CoV-2 infections in the South Asian countries from European regions. Phylogenetic analysis also revealed that the South Asian strains are highly related to the South American and European strains. We found that G-clade mutations are more prevalent (96.19%) in the samples of Bangladesh which were also prevalent in the European isolates. Surprisingly, one missense mutation (1163A > *T*) in ORF1ab gene became dominant only in Bangladesh (78.8%), which led to debates regarding effects on the pathogenicity and transmissibility of the virus. Overall, the findings of this study highlight the frequently mutated SARS-CoV-2 variants among the COVID-19 patients in the South Asian countries which might ease containment of the disease in this region through investigating the virulence reducing factors as the identified mutations are strongly correlated with low infection and mortality rate.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 otherwise known as the SARS-CoV-2 is a virus that is known to cause mild to severe lower respiratory tract infections in humans which has the capability of progressing deplorably in almost all age cohorts, regardless of comorbidities and establishing abrupt clinical outcomes. Since its emergence in December 2019 SARS-CoV-2 has caused over 46 million casualties and 1.2 million deaths globally, the count of which is still escalating. The reason for this ever increasing numbers in terms of fatalities and infections is the virus's rapid spread through human to human contact after the first set of infections occurred (Gorbalenya et al., 2020; Guan et al., 2020; Mahmood et al., 2021). Recent studies have highlighted that over a short span of time, the SARS-CoV-2 virus genome has mutated several times (Wang et al., 2020). Since analysis of these mutations as well as the overall genome sequence of the virus can address major

* Corresponding author.
E-mail address: afrin.bge@nstu.edu.bd (A.S. Chowdhury).

https://doi.org/10.1016/j.crmicr.2021.100065
Received 2 June 2021; Received in revised form 18 August 2021; Accepted 22 August 2021
Available online 24 August 2021

epidemiological parameters such as identification of the infection source, transmission routes and doubling time of the outbreak, the whole genome sequencing of SARS-CoV-2 and its scrutinization in terms of mutation and other additive features has become a key aspect to COVID-19 management (Rambaut et al., 2008; Grenfell et al., 2004). In terms of understanding the comprehensive pathophysiology of the virus, its interactions with the human immune system, analyzing and identifying regions of the mutations is crucial since the ability to frequently mutate and evolve can facilitate SARS-CoV-2 to delude the immune system of the host (Pachetti et al., 2020; Dediego et al., 2008). Single-nucleotide polymorphisms (SNPs) at both coding and non-coding regions have a high capacity of drastically affecting protein structure and functions (Khalid and Sezerman, 2020). Even minimal, local structural differentiation to the closest viral protein may lead to the inheritance or loss of pathogenic properties by the SARS-CoV-2 (Angeletti et al., 2020). Additionally, the changes in virulence, pathogenesis and immunogenic potential can also be brought about by single point mutations in virus proteins (André et al., 2019). When it comes to identifying conditions and factors that could amplify the transmission of SARS-CoV-2, humidity and temperature have been found to be significant mediators of the virus's incidence (Huang et al., 2020; McClymont and Hu, 2021). While multiple experiments are being conducted daily in order to develop easy diagnostic protocols against SARS-CoV-2, to provide more integrity and transparency in these methods, analyzing mutations in SARS-CoV-2 with reference to temperature variations is essential for predicting possible mutations in SARS-CoV-2.

Now frequent and spontaneous mutations in the SARS-CoV-2 resulting in the creation of different viral variants does not only influence its own pathogenesis and virulence but also impacts the overall infection rates and fatalities in different geolocations as well. In spite of the fragile healthcare facilities and a robust population, infection and mortality rate of COVID-19 in Bangladesh, Nepal and Pakistan have been comparatively lower than multiple regions having better medical facilities such as USA, France, Russia etc. (Gupta et al., 2020). Since epidemiological manifestations of COVID-19 have a deviant image in the South Asian region as compared to other geolocations, this study was designed to identify and analyze the pattern of SARS-CoV-2 mutations in South Asia as per the genome sequences obtained from samples isolated from the countries of this region. Through this study, we aimed to provide new insights and shed more light upon the virus's transmission mode, evolution and pathophysiology in this region.

2. Materials and methods

2.1. Data retrieval

A total of 822 whole-genome sequences in FASTA format were retrieved from the GISAID database (www.gisaid.org) submitted by the following South Asian countries: Bangladesh (184), India (215), Pakistan (06), Sri Lanka (04) and Nepal (01); South American country- Brazil (103); North American country- USA (103); Europe - Italy (103) and Germany (103) from 27th January to 07th July, 2020 using the complete genome and high coverage filters of GISAID database (Hadfield et al., 2018). Low-quality sequences containing NNNs were excluded and the accession id, sample location, submitting lab, sample gender, sample age, sequence length, and GC content were also recorded. The genome sequence of the Wuhan isolate of SARS-CoV-2 (Accession ID: EPI_ISL_402124) was also retrieved to be used as a reference genome for subsequent analysis. No sequence data from the other South Asian countries (Bhutan, Afghanistan, and Maldives) were found in the GISAID database for the aforementioned time.

2.2. Data processing

2.2.1. Data pre-processing

The retrieved FASTA sequences were converted to FastQ format using FASTA-to-Tabular and Tabular-to FASTQ tools (Galaxy Version 1.1.0) with default parameters (Blankenberg et al., 2010). The FASTA-to-tabular is an easy to use tool that converts FASTA formatted sequence into tab delimited format and Tabular-to-FASTQ changes tabular file having sequencing-read information into a FASTQ formatted file containing both sequence information and corresponding quality scores (Blankenberg et al., 2010). This conversion was applied to achieve better feasibility for the targeted study as FASTQ formatted sequences are required to perform Multiple Sequence Alignment in the Galaxy platform.

2.2.2. Sequence alignment and data post-processing

The FASTQ formatted sequences were then mapped against the reference genome (Wuhan isolate of SARS-CoV-2) using Map with BWA-MEM (Galaxy version 0.7.17.1) (Li, 2013). Map with BWA-MEM uses Burrows-Wheeler Aligner's Smith-Waterman Alignment (BWA-SW) algorithm which is an efficient, speedy algorithm having higher sensitivity and specificity for aligning longer sequence reads (>100 nucleotide) against query sequence. This tool takes FASTQ formatted file as a input file and returns the aligned file in a Sequence Alignment/Map (SAM) format (Li, 2013). Mapped-reads were then filtered using Filter SAM or BAM, output SAM or BAM (Galaxy Version 1.8+galaxy1) (Li et al., 2009), and potential PCR duplicates were removed by RmDup (Galaxy Version 2.0.1) (Li, 2011) to avoid false variant calling due to misalignment (Li, 2011). The Filter SAM or BAM, output SAM or BAM tool can be used to filter SAM or BAM (Binary Alignment/Map) formatted alignment file based on mapping quality (MAPQ), read group, library or region (Li et al., 2009) while RmDup works only in paired end data and retain the pairs having highest mapping quality (Li, 2011). Both of the tools were run in default settings.

2.2.3. Whole genome variant calling and genomic diversity prediction

For the detection and annotation of SARS-CoV-2 variants from our aligned sequences, FreeBayes (Galaxy Version 1.3.1) (Garrison and Marth, 2012) and SnpEff eff: annotate variants for SARS-CoV-2 (Galaxy Version 4.5covid19) (Cingolani et al., 2012a) tools were used without changing default parameters. FreeBayes is a haplotype-based Bayesian genetic variant detector explicitly designed to detect single nucleotide polymorphisms (SNPs), indels (insertions and deletions), multi-nucleotide polymorphisms (MNPs) and complex events (composite insertion and substitution events) depending on literal sequences of reads aligned to a particular target (Garrison and Marth, 2012). SnpEff eff is a fast, flexible open source tool specially designed for annotating SARS-CoV-2 genetic variants and for the prediction of effects of changes such as synonymous or non-synonymous replacement, frameshift etc. on genes and proteins (Cingolani et al., 2012a) SnpSift Extract Fields (Galaxy Version 4.3+t.galaxy0) (Cingolani et al., 2012b) was used to convert the variant calling file (VCF) formatted output generated by SnpEff tool into tabular form for better representation and interpretation of the results. SnpSift Extract Fields is a freely available tool for manipulating and filtering annotated VCF formatted output file from the SnpEff tool based on chromosome, position, ID, annotation impact etc. (Cingolani et al., 2012a).

2.3. Prediction of transmission lineages of the targeted sequences

Lineage classification is a dynamic system to evaluate the genomic epidemiology and transmission pattern of the rapidly evolving virus. Therefore, we performed a nomenclature analysis to label SARS-CoV-2 lineages for each of the selected sequences of South Asia using Pangolin (Phylogenetic Assignment of Named Global Outbreak LINeages) COVID-19 lineage Assigner (https://pangolin.cog-uk.io/). Pangolin is a robust and rational approach for naming the phylogenetic diversity of SARS-CoV-2 sequences based on set of criteria to aid in the surveillance and understanding the evolution and spread of the virus to new locations (Rambaut et al., 2020).

2.4. Phylogenetic tree and evolutionary analysis

Phylogenetic tree is a simple branched tree or diagram depicting the origin and evolutionary relationship between group of species (Zhang et al., 2020). To probe out the relatedness between our analyzed 822 SARS-CoV-2 genome sequences, we generated a phylogenetic tree by MEGA X (Molecular Evolutionary Genetics Analysis) software (Kumar et al., 2016). MEGA is a high-throughput bioinformatics tool for analyzing the evolutionary patterns and diversity on earth using large datasets (Kumar et al., 2016). Firstly, the sequences were aligned using multiple sequence alignment algorithms and then a phylogenetic tree was constructed using the Maximum Likelihood (ML) method with default parameters. Finally, the phylogenetic tree file was annotated using the iTOL (Interactive Tree of Life) v4 for better interpretation and visualization (Letunic and Bork, 2019).

2.5. Protein structure prediction and homology modeling

To generate the tertiary structure of the mutant NSP2 protein, at first, the tertiary structure of the wild type NSP2 (NCBI Reference Sequence: YP_009725298.1) was retrieved from I-TASSER database (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) (Huang et al., 2020; Zhang et al., 2020; Roy et al., 2010). The iterative threading assembly refinement (I-TASSER) is a unified platform for automated prediction of protein structure and function from given amino acid sequence (Roy et al., 2010). Then, homology modeling of the mutated NSP2 (I120F) was carried out. Overall, the structure of NSP2 protein was predicted on the basis of different protein model quality parameters such as C-score, RMSD, TM-score etc. Finally, the generated tertiary structure of the mutant NSP2 was visualized with Discovery studio.

2.6. Mutational effect analysis on protein structure

Mutations have wide range of impact on the protein structure and function which in turn can increase the infectivity and transmissibility of a virus by altering native structure and function. To determine the effect of mutation on the structure of the NSP2 and predicting the impact of mutations on conformation, stability and flexibility of the protein, the protein sequence of the wild-type NSP2 was uploaded on DynaMut software (http://biosig.unimelb.edu.au/dynamut/) (Rodrigues et al., 2018) in default settings. DynaMut is a user friendly web-based predictor based on normal mode approaches for assessing the impact of mutation on protein conformation and stability resulting from changes in Gibbs free energy, vibrational entropy (Rodrigues et al., 2018). The difference in vibrational entropy ($\Delta\Delta S_{Vib}$) and Gibbs free energy ($\Delta\Delta G$) between wild-type and mutated proteins along with the atomic fluctuations and deformation energies were calculated. These calculations were performed over the first 10 non-trivial modes of the analysis.

2.7. Statistical analysis

Differential and significant distribution of the identified SNPs among gender and age-specific subgroups was analyzed by using *Cochran-Armitage test*. The statistical outcomes were further visualized through stacked bar plots that clustered the set mutations corresponding to distinct subgroups. All of the statistical analysis was conducted in R statistical environment version 4.0.4 and Graphpad Prism version (Kaya et al., 2019; Rahman et al., 2020).

3. Results

3.1. Analysis of the genomic diversity of SARS-CoV-2 in South Asia

As the intention of this study was to investigate the nature of genetic variations in SARS-CoV-2 genome sequences that were obtained from the samples of South Asian region, as a first step, we selected a total of 410 complete genome sequences of SARS-CoV-2 sampled from five different countries of South Asia- India (215), Bangladesh (184), Pakistan (6), and Sri Lanka (4) (Supplementary Tables 7–10).

From the retrieved sequences, a total of 48 SNPs were identified by aligning selected sequences with the reference genome NC_045512.2 (the initial isolate of Wuhan). Among all of these alterations, 4 SNPs (241C > *T*, 3037C > *T*, 14408C > *T*, and 23403A > *G*) were found to have the most occurrences (more than 300 times) (Fig. 1b). However, the sequence obtained from Nepal was in complete homology with the reference genome obtained from Wuhan. Among the other four countries, prominent mutations were observed in India, indicated by 36/48 SNPs at a frequency ranging between 1.40% and 72.56% (Supplementary Table 1, Fig. 1c). The mutational landscape analysis of SARS-CoV-2 in Bangladesh demonstrated diverse heterogeneity of the samples from where 18/48 SNPs were identified. The frequency of these SNPs ranged from 2.72% to 96.74%. Interestingly, 50% SNPs (9/18 SNPs) in the samples of Bangladesh were found unique in the South Asian region. Remarkably, among these unique SNPs, 1163A > *T* was found at a frequency >78% (Supplementary Table 1, Fig. 1c). On the counterparts, the least mutation frequencies were found in the genomes sequenced from Pakistan (10/48 SNPs) and Sri Lanka (4/48 SNPs). However, 3 SNPs (2416C > *T*, 8371G > *T*, and 22477 C > *T*) in the samples of Pakistan were found to be unique in the South Asian region (Supplementary Table 1). Overall, the diverse landscape of country-specific genetic alterations in SARS-CoV-2 indicated a multi-variant transmission pattern of the virus rather than only from its origin.

3.2. Classification of the SARS-CoV-2 variants found in the South Asian region

To observe the in-depth diversity of SARS-CoV-2 in terms of mutation, we aimed to classify the obtained 48 SNPs based on their types and gene-specific localizations. Considering the localizations in the genome, we observed that more than 85% SNPs found in the South Asian countries are clustered in *ORF1ab*, spike protein, and nucleocapsid of the virus (Fig. 2a). Previous studies have identified the SARS-CoV-2 spike protein as a mutation hotspot by which the interaction between receptor binding domain (RBD) of the spike protein and host Angiotensin Converting Enzyme 2 (ACE2) receptor becomes more stable (Teng et al., 2021). This specific localization of mutations is an indicative of the amenable impact of these three genes to frequent genetic alterations in the SARS-CoV-2 genome. Similar outcomes were also found for the country-based classification of the SNPs (Fig. 2b–e). On the contrary, the membrane (*M*) and envelope (*E*) proteins showed the least level of susceptibility towards genetic alterations (alteration frequency <3%) (Fig. 2a).

Following classification, we characterized the identified SNPs based on their functional attributes. We grouped the SNPs into three categories of heterogeneity- Synonymous, Missense, and Non-coding SNPs. Overall, the prevalence of missense type SNPs (54.17%) was observed at a higher rate compared to the synonymous (41.66%) and the non-coding ones (4.17%) (Fig. 2f). In country- specific classification, the most frequent level of missense type SNPs (61.11%) was found in the samples of Bangladesh (Fig. 2h). This type of mutation, where the amino acids were susceptible to change were found to be dominant and these alterations are predicted to have played their part in evaluating the genomic plasticity of the viral particle that geo-locates the genomic regions which are prone to mutations (Karamitros et al., 2020). Despite the missense type mutations, the frequency of synonymous mutations was also found at a higher rate (41.66%) compared to the non-coding mutations in the South Asian region. Among the country-specific classifications, the highest level of synonymous mutations (38.88%) was found in the samples of India (Fig. 2g). On the other hand, the least amount of SNPs (2 SNPs; 4.17%) were identified as non-coding in the overall analysis of the South Asian SARS-CoV-2 samples (Fig. 2f).

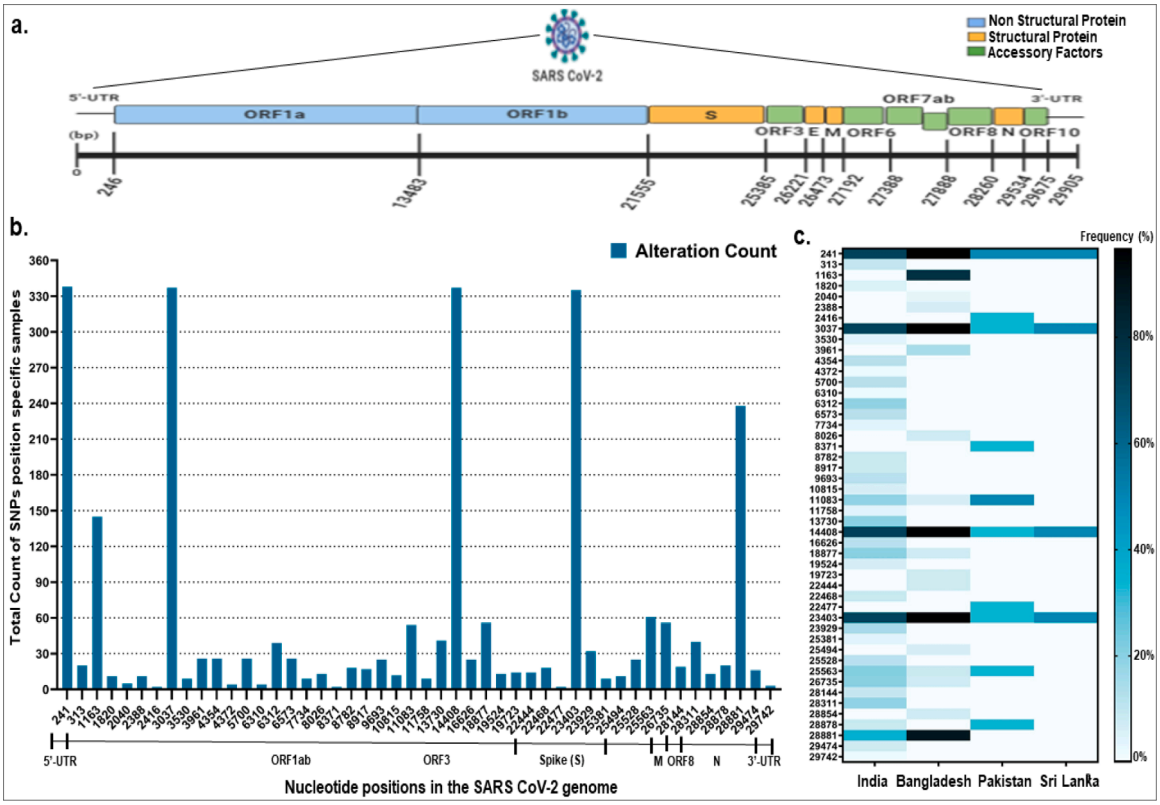

Fig. 1. Genomic diversity prediction of the SARS-CoV-2 genome isolated from South Asia. **(a)** The genomic map of the SARS-CoV-2 virus. The 29905bp long SARS CoV-2 virus consists of two non-structural protein region (*ORF1a, ORF1b*), four structural protein region (*S, E, M, N*) and five accessory factors region (*ORF3, ORF6, ORF7ab, ORF8, ORF10*) **(b)** The bar plots represented the total count of SNPs-specific samples per genomic position. Each of the bar was specific for each of the identified 48 SNPs **(c)** The heatmap illustrated the country-specific frequency of the SNPs in individual genomic positions of SARS-CoV-2 virus. The color bar represents mutation frequency- higher the frequency, higher the intensity of the color.
Abbreviations: SNPs- Single Nucleotide Polymorphisms; *5′-UTR-* 5'- Untranslated region; *ORF1ab-* Open reading frame 1ab; *S-* Spike protein; *M:* Membrane protein; *N-* Nucleocapsid; *3′-UTR-* 3′-Untranslated region.

3.3. Comparative study of the alteration frequency in SARS CoV-2 genome based on gender and different age groups

The physiological fitness of the host and underlying comorbidities have been found to be a notable determinant in analyzing the clinical manifestations and prognosis of SARS-CoV-2 (Mahmood et al., 2021; Sanyaolu et al., 2020). With that in mind, we aimed to observe the genomic diversity of SARS-CoV-2 based on country-specific age and gender cohorts. Although multiple sequence occurrences lacked information relevant to the gender and age of the person from whose sample it was sequenced, there were still more than 400 sequences available to relate gender and age with the distribution of mutation (Supplementary Tables 7–10). In terms of age, we classified sequences into four age groups and found most of the mutations exhibited a fluctuant trend with age. We also found that mutations in *5′-UTR, ORF1ab*, and *S* genes were common in all age groups of different geolocations. In India, two unique mutations in *ORF7a* and *ORF6* gene were found in the 1–20 and 41–60 years age group (Fig. 3a), the highest numbers of mutations were found in the 1–20 years age group as well. In Bangladesh, the highest number of mutations were observed in the age group 41–60 years (Fig. 3b) and mutation in the *ORF8* gene (28144T > *C*) was found only in this age group. Mutations in the *ORF3a* gene were common to most of the age groups of Bangladesh and India.

In case of gender, most of the mutations in different countries were similar between males and females. Further analysis revealed that mutations in the *5′-UTR, ORF1ab, ORF3a, M, N* and *S* gene were common in both groups of Bangladesh and India. *ORF8* (both groups) and *3′-UTR* (only in females) mutations were specific to the Indian population (Fig. 3c,d). The statistical significance of the distributed SNPs in different subgroups was also evaluated (Supplementary Table 2).

3.4. Major clades-specific clustering of the SNPs found in the South Asian countries

To predict and trace the source of transmission in multiple variants in terms of geolocation, six major clades of mutations were considered as the hotspots of widespread transmission of the new serotypes of SARS-CoV-2 (Takahiko et al., 2020). We clustered the analyzed SARS-CoV-2 samples according to these major clades of mutation for locating the transmission geolocation based on the variants found in the South Asian region (Fig. 4, Table 1). The largest clustered group of the samples was found based on the G-clade (D614G), containing 81.7% (335/410 samples) of the total sample count. G-clade mutations had the highest prevalence (96.19%) for the samples of Bangladesh. Considering the F-clade (L3606F), 54 samples (13.17%) were clustered and majority of these samples (40) were listed from India. However, the frequency of F-clade mutation was found higher (50%) in the samples of Pakistan compared to the other South Asian countries. The first observation of this clade containing viral strain was recorded on 18 January 2020 in Chongqing, China (Takahiko et al., 2020). Following this background, it can be predicted that many South Asian samples are transmitted as a mutated version of the viral strains directly from its origin. According to the S-clade (L84S) mutations, only the samples of India were found to be clustered at a frequency of 11.63%. This clade of mutation is one of the pre-pandemic staged mutations firstly observed in Wuhan, China on 30 December 2019 (Takahiko et al., 2020). Considering the transmission

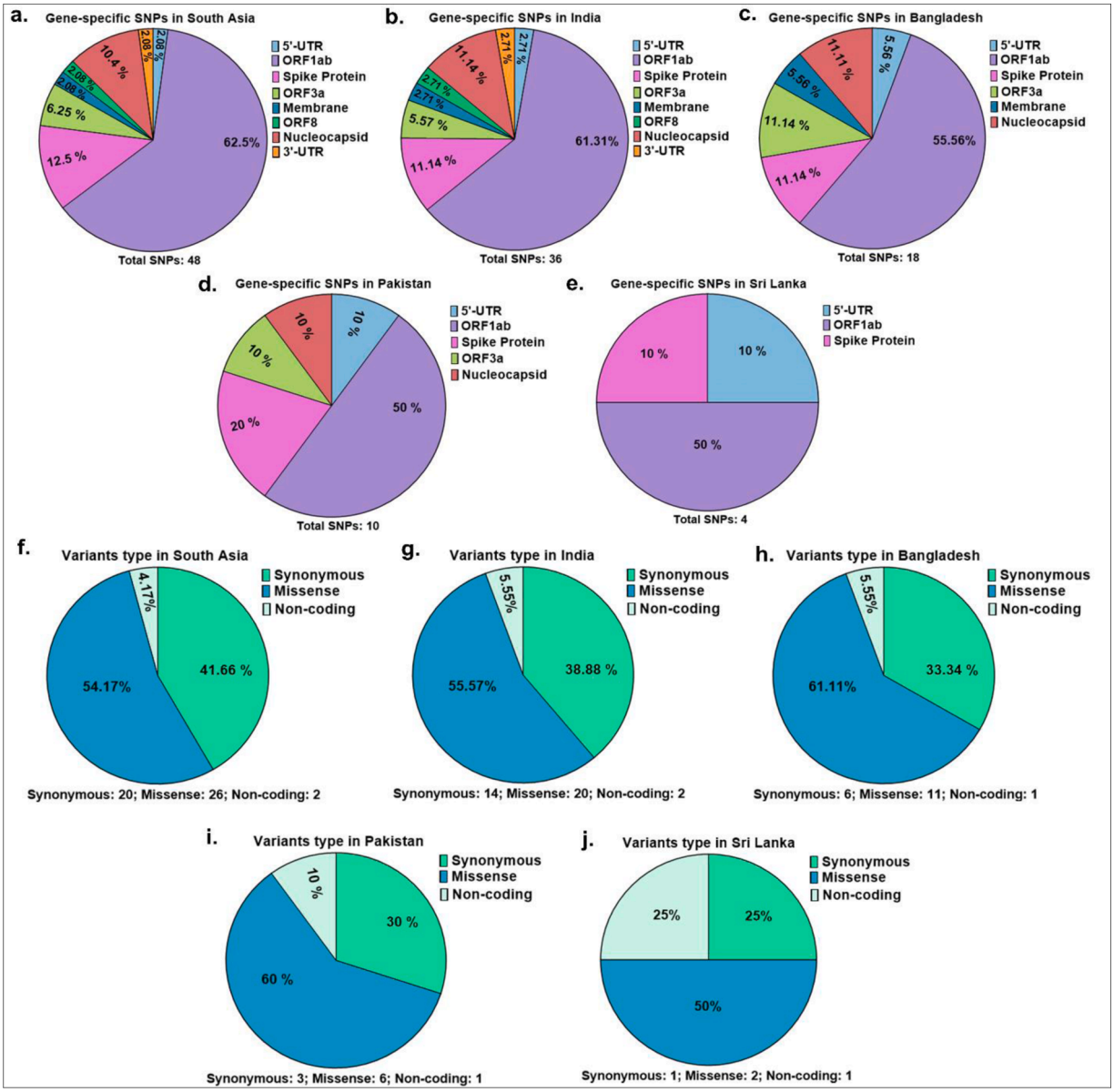

Fig. 2. Classification of the SNPs found in different regions of South Asia. **(a–e)** The pie charts represented gene-specific frequency of the SNPs. The identified SNPs were localized in eight different genomic regions (*5'-UTR, ORF1ab, Spike, ORF3a, Membrane, ORF8, Nucleocapsid and 3'-UTR*). Each of the genomic regions were color coded uniformly. **(f–j)** These pie charts classified the frequency of variants of SNPs in South Asian regions. Three types of SNPs (Synonymous, Missense and Non-coding) were identified. Each type was marked through uniform color code.

footstep of this clade, it can be predicted that some of the samples of India are holding a closer transmission network with the parental viral strains of China rather than the samples of other South Asian countries. Following the V-clade (G251V) mutations, we could not find any samples to be clustered. This clade of mutation was firstly reported in Australia, on 25 January 2020. Later on, the widespread transmission of this clade was reported in North America (USA), South America (Brazil) and Australia. According to the H-clades (Q57H), 71 South Asian samples (13.71%) were clustered where the samples of India (44) were found as the most dominant one. The first viral strain bearing this clade of mutation was reported in France on 26 February 2020 (Takahiko et al., 2020). The higher frequency of this clade in the South Asian region suggested the widespread transmission of the European viral strains into the South Asian countries. Following the KR-clade (RG203KR), we have clustered 238 South Asian samples (58.05%) where most of the samples (163) were reported from Bangladesh (Table 1). The earliest sequence consisting of this clade of mutation was notified in Germany, on 25 February 2020 (Takahiko et al., 2020). The frequent assembly of this clade in South Asia strongly remarked the possible transmission linkage of South Asian SARS-CoV-2 samples with the European ones. Overall, these clades-based clustering boldly recommended the robust transmission network of SARS-CoV-2 serotypes from China and European regions to South Asia.

3.5. *Comparative analysis of the SARS-CoV-2 genomic alterations found in North America, South America, Europe and South Asia*

In total 27 SNPs were identified by aligning the selected sequences (isolated from the USA, Brazil, Italy and Germany) with the reference genome NC_045512.2 (the initial isolate of Wuhan). Following the country-specific analysis, the highest number of alterations was observed in the USA (16) followed by Brazil (12), Germany (11) and Italy (07).

In case of Italy, four mutations (241C > *T*, 3037C > *T*, 14408C > *T*

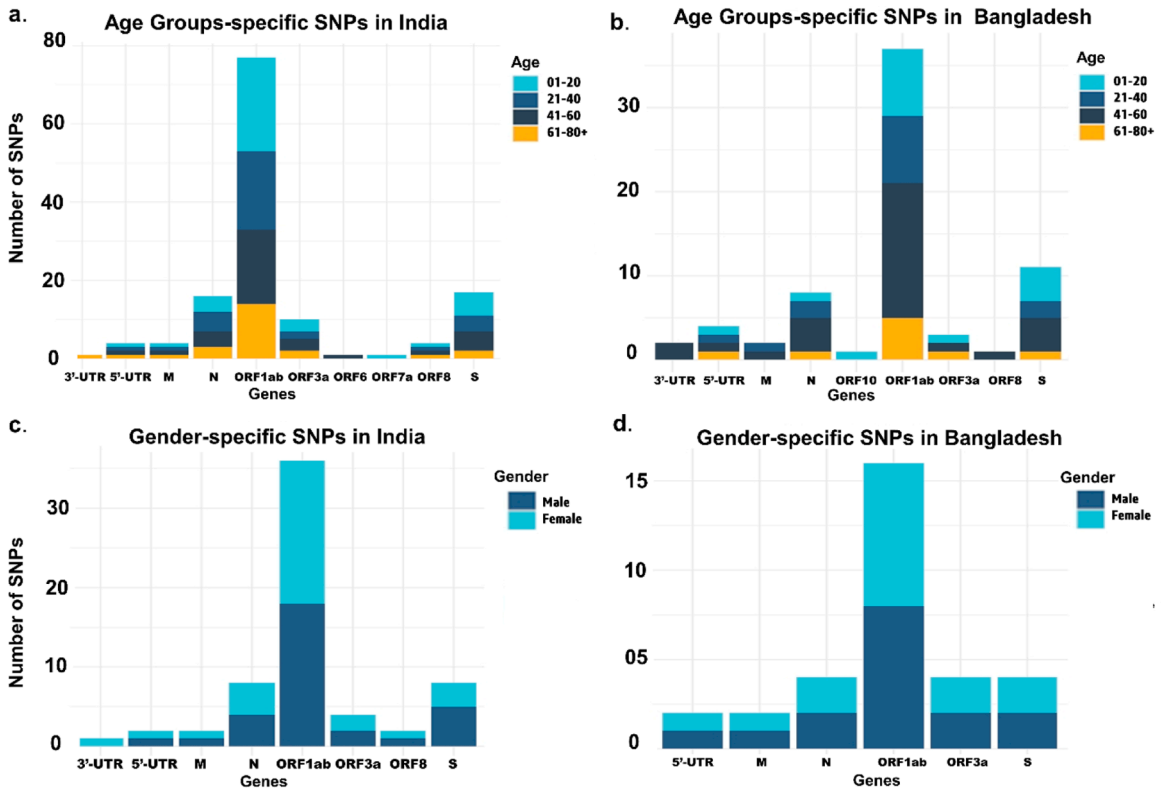

Fig. 3. Comparative study of the SNPs frequency according to variant age groups and gender. **(a,b)** The bar plots represented total count of SNPs according to four different age groups (01–20 years, 20–40 years, 41–60 years, 61–80+ years). Each of the age groups were marked through uniform color code **(c,d)** These bar plots represented gender-specific count of SNPs in South Asian countries.

and 23403A > *G*) were found at a frequency of 96.12% whereas 20268A > *G* and 26530A > *G* (unique mutation found in only Italian samples) were found at a very low frequency (≤11%) (Supplementary Table 3). A total of 12 mutations were found in Germany in which four mutations 241C > *T*, 3037C > *T*, 14408C > *T* and 23403A > *G* were found common and existed at a frequency of 87.38%, 89.32%, 77.62% and 88.35%, respectively. However, four other unique mutations (1059C > *T*, 6446G > *A*, 25550T > *A* and 27046C > *T*) were identified compared to the samples of South Asia though their frequency was quite low (≤14%) (Supplementary Table 4).

In the USA, out of 16 mutations, four mutations (241C > *T*, 3037C > *T*, 14408C > *T* and 23403A > *G*) were found at a frequency of 99.03%, meanwhile, nine unique mutations were identified but at a very low frequency (≤13%). Among those unique mutations, 1059C > *T* mutation was found at a considerable frequency of 46.60% (Supplementary Table 5). After analyzing the samples of Brazil, a total of 12 mutations were found in which four of them (241C > *T*, 3037C > *T*, 14408C > *T* and 23403A > *G*) were identified at a frequency of 100%. However, seven unique mutations were found from the isolates of Brazil those were not present in the samples of South Asia. Among these unique SNPs, 27299T > *C* and 29148T > *C* were recognized at a considerably high frequency of 66.02% and 65.05%, respectively (Supplementary Table 6).

Among all of these alterations originated from North America, South America and European regions, five SNPs (241C > *T*, 3037C > *T*, 14408C > *T*, 23403A > *G* and 28881GGG > AAC) were the most common and frequent which were also found in the samples of South Asia.

3.6. Transmission lineages and evolutionary analysis

To evaluate the evolutionary layout of the SNPs transmitted towards South Asia from diverse regions of the SARS-CoV-2 pandemic, we examined different types of transmission lineage. Overall, B.1 (12) and B.6 (15) lineages were higher in amount than other lineages during April in South Asia. B.1 and B.1.1.25 lineages were increased significantly in May, 2020 beside this B.1.1.306, B.6, B.1.1.8 and A lineages were also found in significant amount during this month. The amount of B.1.1.25 lineage was further increased in June, 2020 and decreased in July, 2020 while B.1.1.316, B.1.1.70 and B.1.1.80 linages were found in the highest amount during June, 2020 (Fig. 5a). In Bangladesh, B.1.1.25 lineage was predominant during April, 2020 (4) followed by gradual increase in May, 2020 (23) and June, 2020 (41) (Fig. 5a). B.1.1.316 lineage appeared in May, 2020 (7) and found to be increased in June, 2020 (17) while the highest number of B.1.1.80 lineage was found in June, 2020 (18) (Fig. 5b). For India, B.1 (11) and B.6 (17) lineages were found in the highest amount April, 2020 showing subsequent increase in amount during May, 2020 and gradual decrease during June, 2020 and July, 2020. Three other linages namely, B.1.1.8, B.1.1.306 and B.1.458 were also dominant in May, 2020 and June, 2020 (Fig. 5c).

Following the phylogenetic analysis, isolates from Wuhan, USA, Italy, Brazil, and Germany were aligned with the analyzed South Asian isolates to find out the evolutionary homology of the South Asian strains with the South American, North American and European strains. The phylogenetic tree represents a high correspondence of the South Asian strains with the South American and the European strains. Isolates of Bangladesh and Sri Lanka are more closely related to the isolates of Italy and the USA, whereas isolates of India and Nepal showed a closer relation with the isolates of Brazil and Germany (Fig. 5e). On the other hand, Pakistani isolates are related to the USA, Brazil and Germany isolates. This phylogenetic relationship may validate the fact that the initiation of SARS-CoV-2 infections in the South Asian regions were triggered by immigrants arriving from these highly infected countries.

3.7. Homology modeling and mutation effect analysis

Upon comprehensive analysis of the genomic diversity in SARS-CoV-

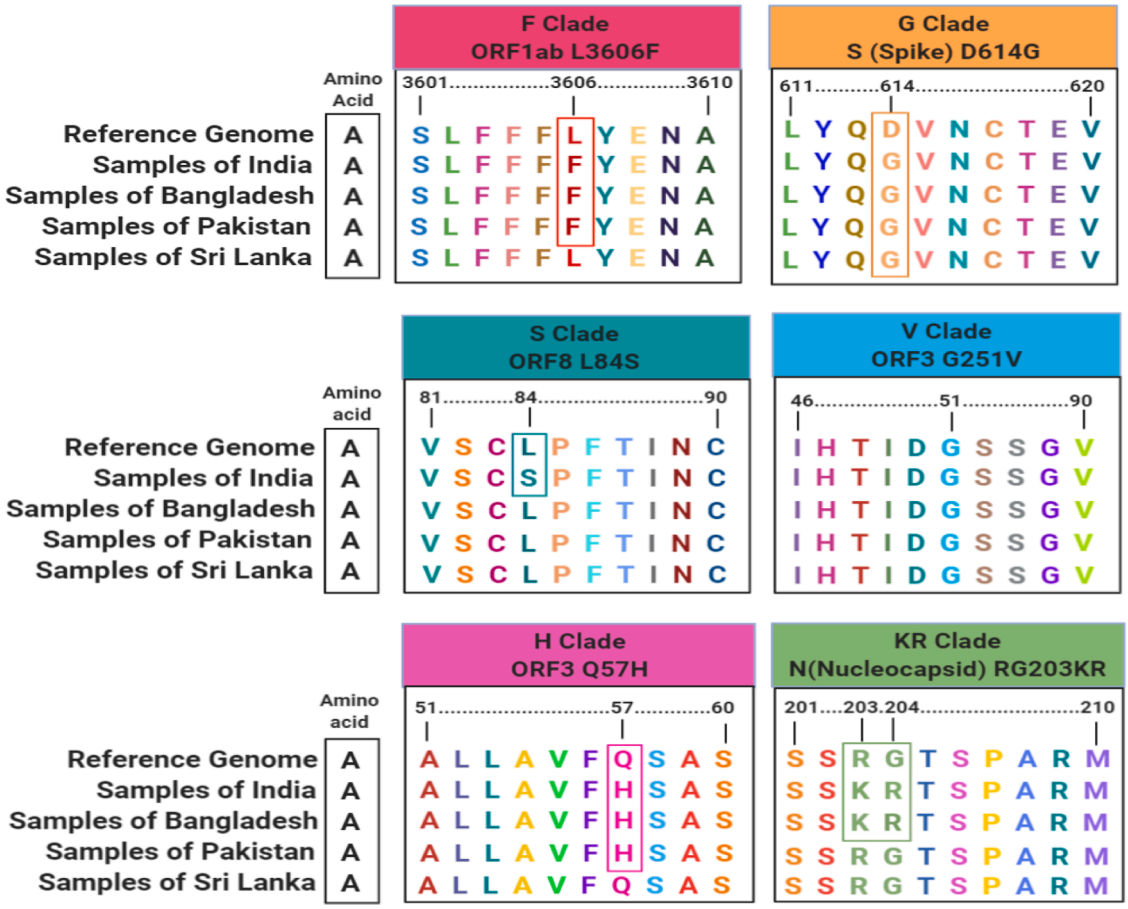

Fig. 4. Clustering of the South Asian SARS CoV-2 samples according to six major clades of mutation (F Clade, G Clade, S Clade, V Clade, H Clade and KR Clade). The color coded boxes of each clade shows wild type and mutant amino acid in a specific genomic position. In each of these boxes, samples isolated from four South Asian countries (India, Bangladesh, Pakistan and Sri Lanka) were compared with the reference sequence (Isolate of Wuhan) and altered regions were marked in a sub-box.

Table 1
List of the clades-specific clustered samples.

Clades	Earliest observation of the strain			No. of samples count/Frequency (%)			
	Date	Accession ID	Location	India	Bangladesh	Pakistan	Sri Lanka
D614G	24 January 2020	EPI_ISL_422425	China	154/71.63	177/96.19	2/33.33	2/50
L3606F	18 January 2020	EPI_ISL_408481	China	40/18.6	11/5.98	3/50	0/0
L84S	30 December 2019	MT291826	China	19/11.63	0/0	0/0	0/0
G251V	25 January 2020	EPI_ISL_408977	Australia	0/0	0/0	0/0	0/0
Q57H	26 February 2020	EPI_ISL_418219	France	44/20.47	15/8.15	2/33.33	0/0
RG203KR	25 February 2020	EPI_ISL_412912	Germany	75/34.88	163/88.59	0/0	0/0

2, we found four common SNPs (241C > *T*, 3037C > *T*, 14408C > *T* and 23403 C > *T*) in the samples isolated from countries of South Asia, Europe, North America and South America which encountered frequent infections. All of these four SNPs were found at a high frequency (≥71.6%). Additionally, a unique SNP (1163A > *T*) was found only in the isolates of Bangladesh at a considerably higher frequency (78.80%). Among the countries which were labeled to have frequent occurrences of COVID-19 cases, statistical records established that the mortality rate was the least in Bangladesh (Table 2). The ORF1ab polyprotein region, which comprises NSP1-3 (Wan et al., 2020) is a crucial factor for coronaviruses. The ORF1ab region has been identified as a potential mutational hotspot. Comparing these records with the frequency of 1163A > *T* mutation obtained from our analysis, it can be predicted that the 1163A > *T* mutation, which was local to the structural protein NSP2, may be responsible for restraining the severity of SARS-CoV-2 serotypes found in Bangladesh. To align this prediction, we analyzed the impact of 1163A > *T* mutation on the stability of NSP2 protein through homology modeling. In this case, we used one of the most common and well-studied D614G mutation found in the spike protein as a control agent of this analysis.

To map the structural variability between wild type and mutant NSP2, homology modeling was performed and the tertiary structure of 638aa long wild type and mutant type NSP2 proteins were generated (Fig. 6a,b). For viruses, mutations can affect and alter the functions of different viral proteins in either a deleterious way or an additive way and even lead to phenotypic changes which may be harmful to the virus itself. Hence viral pathogenesis is crucially dependent on the results of the mutation a virus undergoes. To assess the effect of the 1163A > *T* (I120F) mutation on NSP2 protein, different parameters such as Gibbs free energy, the difference in vibrational entropy between wild type and mutant NSP2, fluctuation and deformation energies were calculated by DynaMut tool. Changes in the ΔΔG due to the mutations ultimately correlated with changes of protein structure such as changes in cavity volume, accessible surface area and packing density. Hence, ΔΔG

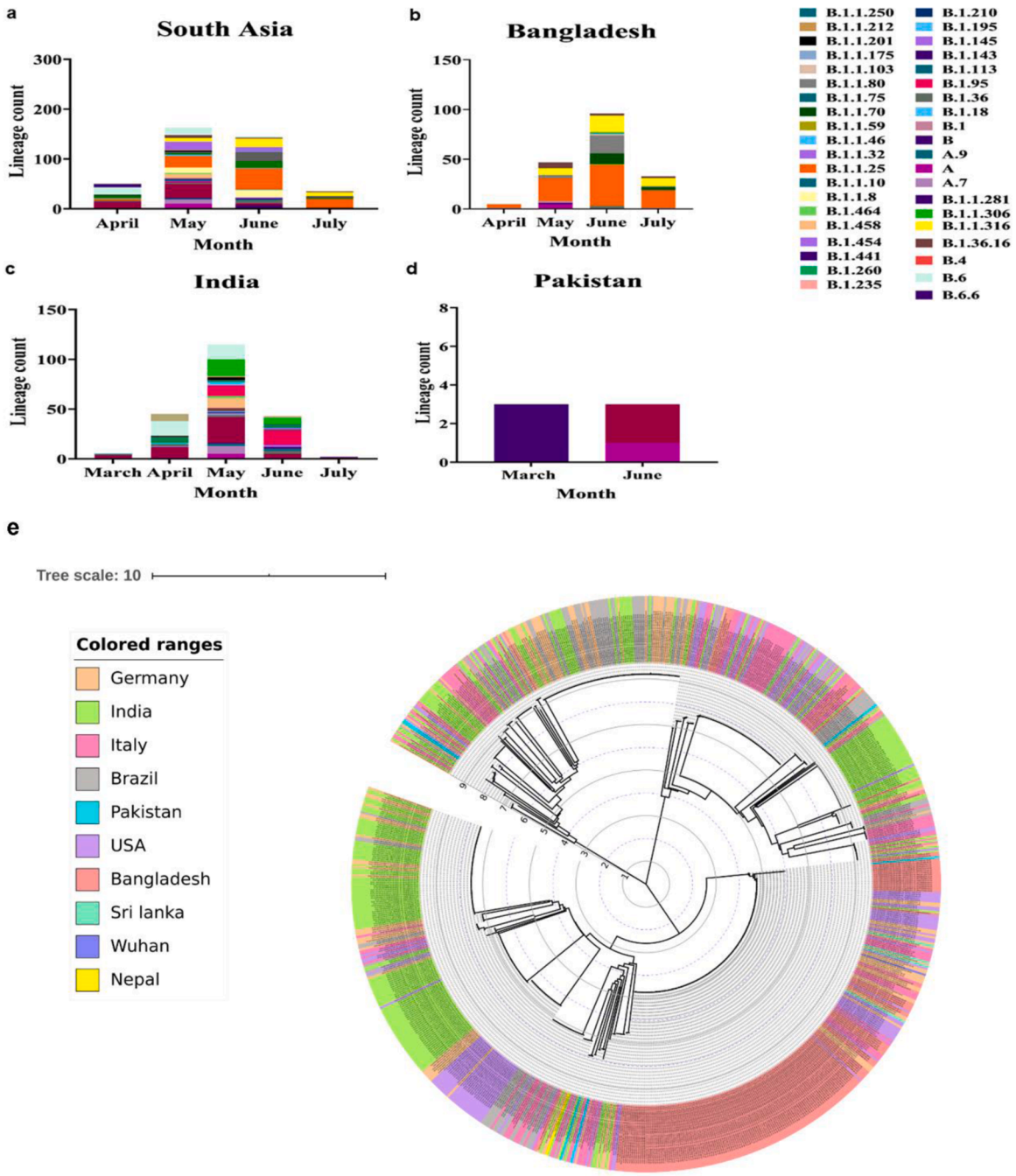

Fig. 5. Mapping of the transmission lineages of mutated SARS CoV-2 originated from Europe, North America, and South America to the South Asian countries. **(a–d)** The stacked bar histograms showed 40 different types of lineages for each of the sequences of South Asia. Each of the lineages were marked by uniform color-code. **(e)** The phylogenetic tree represents the evolutionary relationship between South Asian SARS-CoV-2 strains and the strains isolated from China, the USA, and two European countries (Germany and Italy). Each of the strains were specified through uniform color code.

summarizes the impact of mutation on protein stability (Eriksson et al., 1992). Generally, a positive ΔΔG value implies an increase in protein stability and on the other hand, a negative ΔΔG value means decrease in protein stability (Chand et al., 2020; Chaudhuri et al., 2021). From our structural analyses, data revealed that the ΔΔG value for wild and mutant NSP2 was 1.602 kcal mol^{-1} depicting a stabilizing mutation. The interatomic interaction between the amino acid at 120th position and surrounding residues of wild type and mutant NSP2 was visualized to analyze the core structural changes due to the mutation (Fig. 6c,d). Vibrational entropy is the measure of average configurational entropies in the single minima landscape of a protein (Goethe et al., 2015) and differences in $\Delta\Delta S_{Vib}$ gives a clear idea about the impact of mutation on overall protein flexibility (Teruel et al., 2021). In general, a positive $\Delta\Delta S_{Vib}$ value represents flexibility of a protein structure while a negative $\Delta\Delta S_{Vib}$ describes the rigidity of the protein structure (Teruel et al., 2021). The $\Delta\Delta S_{Vib}$ value between wild type and mutant NSP2 was found -1.230 kcal $mol^{-1}k^{-1}$ in out analyses indicating a reduction in molecular flexibility that may result in better stability of the targeted protein (Table 3, Fig. 6e). Atomic fluctuation depicts the amplitude of the absolute atomic motion and deformation energy provides a measure for the amount of local flexibility in a protein. In wild type and mutant NSP2, atomic fluctuation values were found to be 0.108 and 0.109, respectively, whereas deformation energy were found to be 3.001 and 2.982, respectively (Table 3). The lower resulting deformation energy of

Table 2
Statistical records of the infection and mortality rate associated with the most common SNPs found in highly infected countries.

				Italy (%)	Brazil (%)	USA (%)	Germany (%)	India (%)	Bangladesh (%)	Pakistan (%)	Sri Lanka (%)
Total case/1M population				5067	22039	21860	3384	4272	2168	1242	131
Infection rate (%)				0.5	2.20	2.18	0.33	0.42	0.21	0.12	0.013
Death/1M population				592	661	629	114	68	31	27	11
Mortality rate (%)				11.68	2.99	2.87	3.36	1.59	1.42	2.17	8.39
Genomic change	**Gene**	**Amino acid change**	**Impact**	**Italy (%)**	**Brazil (%)**	**USA (%)**	**Germany (%)**	**India (%)**	**Bangladesh (%)**	**Pakistan (%)**	**Sri Lanka (%)**
241C > *T*	*5'-UTR*	Non Coding	Modifier	96.12	100	99.03	87.38	72.0	96.74	50	50
3037C > *T*	*ORF1ab*	F924F/F106F	Low	96.12	100	99.03	89.32	72.0	96.74	33.33	50
14408C > *T*	*ORF1ab*	P4715L/P323L	Moderate	96.12	100	99.03	77.67	72.5	96.20	33.33	50
23403A > *G*	*S*	D614G	Moderate	96.12	100	99.03	88.35	71.6	96.20	33.33	50
1163A > *T*	*ORF1ab*	I300F/I120F	Moderate	0.00	0.00	0.00	0.00	0.00	78.80	0.00	0.00

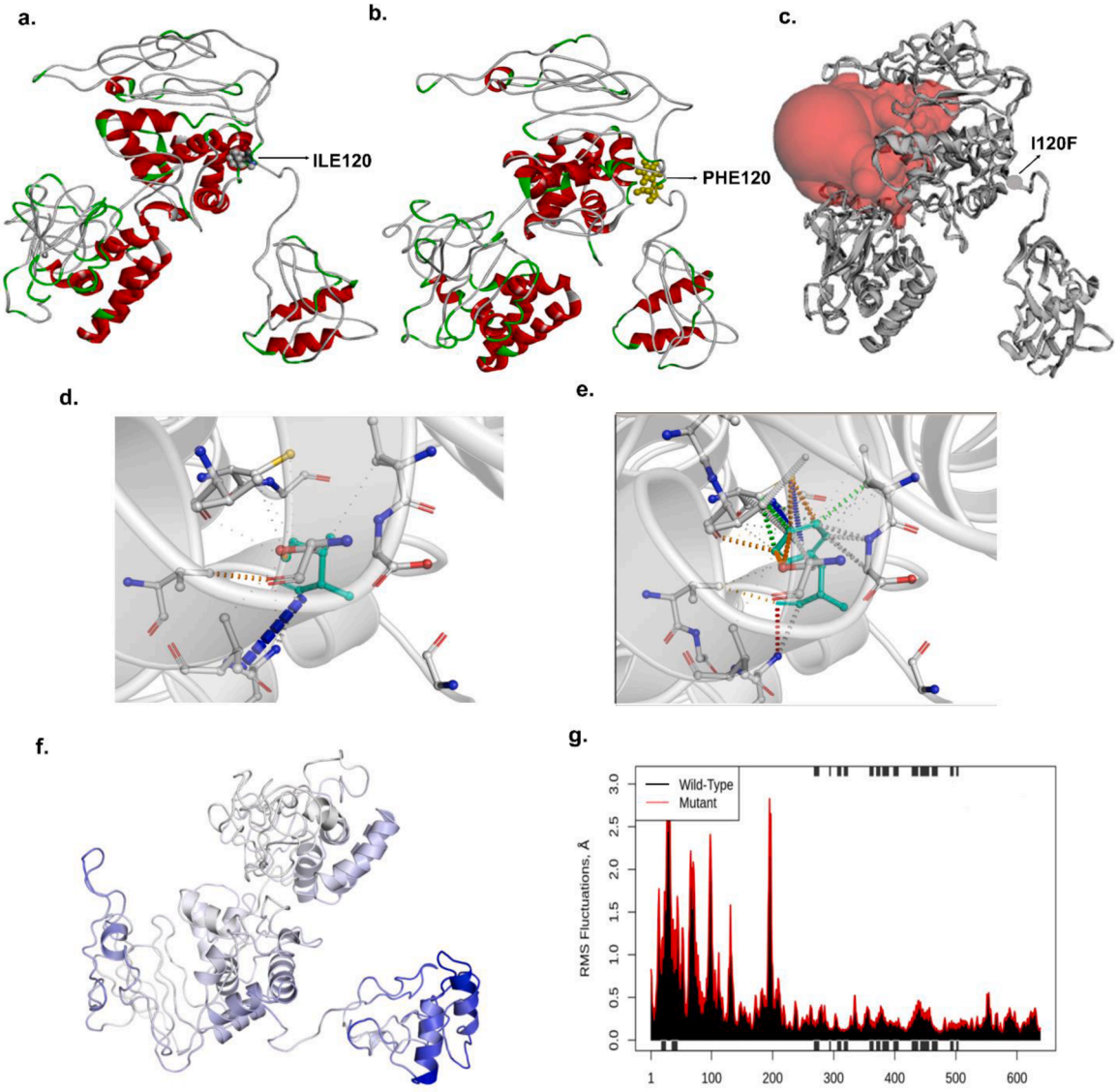

Fig. 6. Visual representation of the impact of 1163A > *T* mutation on the NSP2 protein structure. **(a,b)** Tertiary structure of the wild type and mutant NSP2, respectively **(c)** Structure of the NSP2 domain including its active site. The dome shaped active site of the protein marked in light red color. **(d,e)** Interatomic interaction of wild type and mutant NSP2, respectively. Wild type and mutant residues are represented by stick shape light green color alongside surrounding residues which are involved in any kind of interactions. **(f)** Vibrational entropy energy between the wild type and mutant NSP2 ($\Delta\Delta S_{Vib}$ = -0.145 kcalmol^{-1}k^{-1}). Amino acids are colored according to the changes upon mutations where blue and red color represents rigidity and flexibility of structure. **(g)** Root mean square fluctuation (RMSF) of the wild type and mutant NSP2 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 3
Comparative structural analysis of the wild type and mutant type (1163 A>T; I120F) NSP2 proteins of SARS CoV-2.

Protein Type	Amino Acid	Atomic fluctuation	Deformation energy	ΔΔG between wild and mutant type	$\Delta\Delta S_{Vib}$ between wild and mutant type
Wild type spike	D614	0.482	3.143		
Mutant type spike	G614	0.481	3.145	0.38	-0.02
Wild type nsp2	I120	0.108	3.001		
Mutant type nsp2	F120	0.109	2.982	1.602	-1.230

ΔΔG – Difference between the Gibbs Free Energy ($kcalmol^{-1}$) of Wild type & Mutant type NSP2.
$\Delta\Delta S_{Vib}$ – Difference between the Vibrational Entropy Energy ($kcalmol^{-1}k^{-1}$) of Wild type & Mutant type NSP2.

mutant NSP2 emphasizes that it is indeed more stable than the wild type. Root mean square fluctuation (RMSF) is the measure of structural variability and from the RMSF plot, it can be seen that mutant NSP2 fluctuated slightly more than the wild type NSP2 (Fig. 6f). Overall, the results predicted that the isoleucine to phenylalanine change in NSP2 resulted in the reduction of its molecular flexibility, which in turn formed as a more stable mutant as compared to the wild type.

4. Discussion

As previously published evidence has suggested, concurrent mutations in the SARS-CoV-2 genome can be harmful for the human host and can vary from population to population. Hence, identifying to what extent these variant strains exist in specific geographical regions can help to assess and undertake multiple therapeutic approaches both in general as well as region-wise. In this study the prevalent and important co-mutations (241C > *T*, 3037C > *T*, 14408C > *T*, 23403A > *G*) which were identified mostly prevailed in the South Asian SARS-CoV-2 isolates and aligned with those identified in the European countries (Yin, 2020). As the Indian sample pool was diversified in terms of geo-location, it contributed to the maximum of the identified SNPs (36/48) in the South Asian countries which included several missense mutations in a few SARS-CoV-2 genes (Hassan et al., 2020). Among the 11 genes, 85% of the identified SNPs were frequently found in ORF1ab, spike protein, and nucleocapsid genes, while membrane, envelope, ORF6, ORF7a and ORF7b were more conserved in the South Asian countries, which is consistent with the findings from other countries (Laha et al., 2020). A previous study conducted on isolates from 50 different countries found a significant correlation between the 614G variant of spike protein and ORF1ab 4715L missense mutation with high fatality rates in 28 countries and 17 states in the USA (Toyoshima et al., 2020). Another study conducted on 664 SARS-CoV-2 whole genome sequences also observed a higher frequency of non-synonymous mutations in both the spike protein and ORF1ab (Laha et al., 2020). ORF1ab has been found to have a contributing role in the early evolutionary phases of SARS-CoV-2 (Velazquez-Salinas et al., 2020), while its 3 non-structural proteins, NSP1, NSP3, NSP16 have been found to play a key role in suppressing the host immune response and promoting viral pathogenesis (Tang et al., 2020; Emam et al., 2021). Frequent non-synonymous mutations in regions that act as antigenic determinants of a virus also renders antibody recognition inefficacious (Gershoni et al., 2007; Irving et al., 2001; Gupta et al., 2020). Further analysis on ORF1ab region is required to assess its clinical significance.

Judging from the perspective of age, the highest numbers of mutations from India and Bangladesh were observed in the samples obtained from the 01 to 20 years and 41 to 60 years age cohort, respectively. However, since this concentration of mutations has not been identified elsewhere, the fact that we only had data for a small sample size that could be held accountable. Therefore, this area is a grey zone and calls for more analysis. When it comes to putting a pin on gender specific mortality and infection rates, according to a study on the Italian population, average COVID-19 fatality rate increase with age ranged from 0.27% to 34.68% in males and 0.16% to 20.88% in females which complied with the risk ratio as well; which increased up to 1.75 (Gadi et al., 2020) for males when compared with females. Akin to most of the countries, infection rate and clinical outcomes in the male population was depleting as compared to the female population in both India and Bangladesh (Saha et al., 2020; Shukla et al., 2020). In our study, we found that gender had no influence in regards to the prevalence of specific mutations in the SARS-CoV-2 isolates in populations of the subcontinental regions. This finding contradicts with the results of a study conducted in the United States of America, that found the 27964C > *T*-(S24L) mutation to be more prevalent in females than in males (Wang et al., 2021). As a bias has been observed in the infection and mortality rates between males and females from the very onset of community transmissions of COVID-19, further analysis on its gender specific mutation patterns is required.

Considering the clades-based analysis, we found that a large amount of samples (81.7%; 335/410 samples) were isolated in G-clade. Though the G-clade (D614G) mutations originated from China on 24 January 2020, in the next 30 days from the identification of this mutation the rapid transmission of this clade had occurred in the European countries (Eaaswarkhanth et al., 2020). Therefore, the frequent findings of this clade of mutation in the South Asian countries indicated a strong transmission linkage of this geographical region with the European countries (Eaaswarkhanth et al., 2020). One investigation has demonstrated that the D614G mutation in the Spike protein is associated with high fatality rates (Becerra-Flores and Cardozo, 2020). However, unlike many European countries and India, the mortality rate of Bangladesh was comparatively low, even though G clade mutation was prevalent as compared to other clades. Presence of D614G mutation in the first sequenced sample from Bangladesh indicates the presence of this mutation since the inception of COVID-19 infection and transmission in Bangladesh and explains the reason for the dominance of G-clade mutation in Bangladesh. We also found that isolates from Bangladesh had a high frequency (88.59%; 163/184 samples) of the KR-clade (RG203KR) which implies the possibility of transmissions in Bangladesh from European origins because of the shared homology (Mercatelli and Giorgi, 2020). The combination of Spike D614G and Nucleocapsid RG203KR mutations (Mercatelli and Giorgi, 2020) is currently the most eminent in SARS-CoV-2 positive population. Dissimilar to the readings in Bangladesh, frequency of KR-clade (RG203KR) was found to be relatively low in India (34.88%; 75/215). The variation created by the presence of different viral clades in Bangladesh and India needs to be considered with great attention for effectively scrutinizing the manifestations of vaccines when administered in these two nations. Lineage analysis revealed that during May – June sequences circulating in South Asian region represents the major lineages detected around the world. B.1.1.8, B.1.1.25, B.1.1.70, B.1.1.80, B.1.1.306 and B.1.1.316 lineages were found predominant in our analysis, all carrying D614G, RG203KR, Q675H mutations. These mutations linked with increased transmissibility and higher infection rate contributing to the severity of the disease (Parvin et al., 2021).

A missense mutation at 1163 A > *T* (I120F) causing an isoleucine to phenylalanine shift in the NSP2 was found to be unique in Bangladeshi isolates and persisted at an elevated level (~78.80%). A specific reason to be held accountable behind the unexpectedly high frequency of 1163 A > *T* mutation of NSP2 is yet to be found. One aspect that could relate to this high prevalence is the presence of this mutation type in the first

sequenced isolate from Bangladesh during the onset of COVID-19 transmission in the country (Saha et al., 2020; Rueca et al., 2021). Unfortunately, information on NSP2 protein of SARS-CoV-2, especially the impact of 1163 A > *T* (I120F) mutation is very limited. Therefore, a link between high frequency of 1163 A > *T* (I120F) mutation in NSP2 protein and low mortality rate against the infection rate in Bangladesh can be hypothesized. In the hindsight, transition of an aliphatic hydrophobic side chain amino acid (isoleucine) to an aromatic hydrophobic side chain amino acid (phenylalanine) may induce structural stability in that domain, which was observed by simulating the structure of the NSP2 protein harboring mutated allele through homology modeling. In this study, we found that the 1163 A > *T* (I120F) mutation was increasing the NSP2 protein stability, confirmed by analyzing deformation energy, atomic fluctuation, difference of Gibbs free energy and vibrational entropy. Recent scientific reports claimed that NSP2 is a vital protein that plays crucial role in disrupting host cell environment. It is found to interact with two host proteins prohibitin 1 (PH1) and prohibitin 2 (PH2) which contribute in cell differentiation, cell cycle progression, cellular apoptosis and mitochondrial biogenesis (Yoshimoto, 2020). Considering the relation of NSP2 function against low fatality rate in Bangladesh, we predicted that a frequently occurring mutation (1163A > *T*) in NSP2 protein can be a strong evidence to figure out the reason of low mortality rate in this country. However, we strongly recommend to extend in-depth research following this phenomenon for the proper validation of our findings.

To conclude, this study points to a relation between diverse mutations, including co-evolving ones, in specific SARS-CoV-2 proteins and specific South Asian countries. Further exhaustive studies, including genomic diversities with patient's epidemiological and clinical information, should be combined to identify strategies and therapies that can be useful to reduce the burden of COVID-19 in this region.

5. Conclusion

To date, eliminating and preventing COVID-19 through therapeutics is still a challenge, merely because of SARS-CoV-2's dynamic mutant nature and variating clinical manifestations in different geolocations as well as populations. In this study, through analyzing the patterns and localizations of both genomic and proteomic mutations in SARS-CoV-2 isolates from the South Asian regions, we found that the D614G (G-clade) mutation was the most prominent mutation (81.7%; 335/410 samples) in the South Asian isolates. Upon we were able to identify mutation types specific to countries in South Asia and also establish a phylogenetic and evolutionary analysis we found that the SARS-CoV-2 strains obtained from South Asian isolates are related to those found in South American and European isolates. relationship between those mutation types and the mutations identified in Europe and the Americas to trace the source of infection in terms of location. We also identified a missense mutation (1163A > *T*) which was unique to only Bangladeshi isolates. The unique mutations that were identified in our study specific to the South Asian region and populations can provide insights on the kinds of mutation that amplify and reduce the pathogenic potency of SARS-CoV-2.

Declaration of Competing Interest

The authors declare no competing interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.crmicr.2021.100065.

References

A. Gorbalenya, S. Baker, R. Baric, R. de Groot, C. Drosten, A. Gulyaeva, B. Haagmans, C. Lauber, A. Leontovich, B. Neuman, D. Penzar, S. Perlman, L.L.M. Poon, D. Samborskiy, I. Sidorov, I. Sola, J. Ziebuhr, Severe acute respiratory syndrome-related coronavirus: the species and its viruses – a statement of the coronavirus study group, in, bioRxiv, 2020.

Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., Liu, L., Shan, H., Lei, C.L., Hui, D.S.C., Du, B., Li, L.J., Zeng, G., Yuen, K.Y., Chen, R.C., Tang, C.L., Wang, T., Chen, P.Y., Xiang, J., Li, S.Y., Wang, J.L., Liang, Z.J., Peng, Y.X., Wei, L., Liu, Y., Hu, Y.H., Peng, P., Wang, J.M., Liu, J.Y., Chen, Z., Li, G., Zheng, Z.J., Qiu, S.Q., Luo, J., Ye, C.J., Zhu, S.Y., Zhong, N.S., 2020. Clinical characteristics of coronavirus disease 2019 in China. N. Engl. J. Med. 382, 1708–1720.

Mahmood, T.B., Chowdhury, A.S., Hossain, M.U., Hasan, M., Mizan, S., Aakil, M.M.U.I., Hossan, M.I., 2021. Evaluation of the susceptibility and fatality of lung cancer patients towards the COVID-19 infection: a systemic approach through analyzing the ACE2, CXCL10 and their co-expressed genes. Curr. Res. Microb. Sci. 2, 100022.

Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., Tan, W., 2020. Detection of SARS-CoV-2 in different types of clinical specimens. JAMA 323, 1843–1844.

Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K., Holmes, E.C., 2008. The genomic and epidemiological dynamics of human influenza A virus. Nature 453, 615–619.

Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L., Daly, J.M., Mumford, J.A., Holmes, E. C., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Sci. 303, 327–332 (New York, N.Y.).

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R.C., Zella, D., Ippodrino, R., 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J. Transl. Med. 18, 179-179.

Dediego, M.L., Pewe, L., Alvarez, E., Rejas, M.T., Perlman, S., Enjuanes, L., 2008. Pathogenicity of severe acute respiratory coronavirus deletion mutants in hACE-2 transgenic mice. Virology 376, 379–389.

Khalid, Z., Sezerman, O.U., 2020. A comprehensive study on identifying the structural and functional SNPs of human neuronal membrane glycoprotein M6A (GPM6A). J. Biomol. Struct. Dyn. 1–9.

Angeletti, S., Benvenuto, D., Bianchi, M., Giovanetti, M., Pascarella, S., Ciccozzi, M., 2020. COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis. J. Med. Virol. 92, 584–588.

André, N.M., Cossic, B., Davies, E., Miller, A.D., Whittaker, G.R., 2019. Distinct mutation in the feline coronavirus spike protein cleavage activation site in a cat with feline infectious peritonitis-associated meningoencephalomyelitis. J. Feline Med. Surg. Open Rep. 5, 2055116919856103.

Huang, X., Pearce, R., Zhang, Y., 2020. De novo design of protein peptides to block association of the SARS-CoV-2 spike protein with human ACE2. Aging 12, 11263–11276 (Albany NY).

McClymont, H., Hu, W., 2021. Weather variability and COVID-19 transmission: a review of recent research. Int. J. Environ. Res. Public Health 18.

Gupta, A., Madhavan, M.V., Sehgal, K., Nair, N., Mahajan, S., Sehrawat, T.S., Bikdeli, B., Ahluwalia, N., Ausiello, J.C., Wan, E.Y., Freedberg, D.E., Kirtane, A.J., Parikh, S.A., Maurer, M.S., Nordvig, A.S., Accili, D., Bathon, J.M., Mohan, S., Bauer, K.A., Leon, M.B., Krumholz, H.M., Uriel, N., Mehra, M.R., Elkind, M.S.V., Stone, G.W., Schwartz, A., Ho, D.D., Bilezikian, J.P., Landry, D.W., 2020. Extrapulmonary manifestations of COVID-19. Nat. Med. 26, 1017–1032.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34, 4121–4123.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A., 2010. Manipulation of FASTQ data with galaxy. Bioinformatics 26, 1783–1785.

H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, in, 2013, pp. arXiv:1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, G.P.D.P., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Li, H., 2011. Improving SNP discovery by base alignment quality. Bioinformatics 27, 1157–1158.

E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing, in, 2012, pp. arXiv:1207.3907.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012a. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6, 80–92.

Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M., Lu, X., 2012b. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front. Genet. 3, 35.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5, 1403–1407.

Zhang, C., Zheng, W., Huang, X., Bell, E.W., Zhou, X., Zhang, Y., 2020. Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. J. Proteome Res. 19, 1351–1360.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.

Letunic, I., Bork, P., 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47, W256–W259.
Roy, A., Kucukural, A., Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5, 725–738.
Rodrigues, C.H., Pires, D.E., Ascher, D.B., 2018. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. 46, W350–w355.
Kaya, E., Agca, M., Adiguzel, F., Cetin, M., 2019. Spatial data analysis with R programming for environment. Hum. Ecol. Risk Assess. Int. J. 25, 1521–1530.
Rahman, F., Mahmood, T.B., Amin, A., Alam, R., Jharna, J.F., Samad, A., Ahammad, F., 2020. A multi-omics approach to reveal the key evidence of GDF10 as a novel therapeutic biomarker for breast cancer. Inf. Med. Unlocked 21, 100463.
Teng, S., Sobitan, A., Rhoades, R., Liu, D., Tang, Q., 2021. Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. Brief. Bioinform. 22, 1239–1253.
Karamitros, T., Papadopoulou, G., Bousali, M., Mexias, A., Tsiodras, S., Mentis, A., 2020. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. J. Clin. Virol. 131, 104585.
Sanyaolu, A., Okorie, C., Marinkovic, A., Patidar, R., Younis, K., Desai, P., Hosein, Z., Padda, I., Mangat, J., Altaf, M., 2020. Comorbidity and its impact on patients with COVID-19. SN Compr. Clin. Med. 2, 1069–1076.
Takahiko, K., Daniel, P., Laxmi, P., 2020. Variant analysis of SARS-CoV-2 genomes. Bull. World Health Organ. 98, 495–504.
Wan, Y., Shang, J., Graham, R., Baric, R.S., Li, F., 2020. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. 94.
Eriksson, A., Baase, W., Zhang, X., Heinz, D., Blaber, M., Baldwin, E., Matthews, B., 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. Science 255, 178–183 (New York, N.Y.).
Chand, G.B., Banerjee, A., Azad, G.K., 2020. Identification of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure. PeerJ 8, e9492.
Chaudhuri, D., Majumder, S., Datta, J., Giri, K., 2021. In silico study of mutational stability of SARS-CoV-2 proteins. Protein J. 40, 328–340.
Goethe, M., Fita, I., Rubi, J.M., 2015. Vibrational entropy of a protein: large differences between distinct conformations. J. Chem. Theory Comput. 11, 351–359.
Teruel, N., Mailhot, O., Najmanovich, R.J., 2021. Modelling conformational state dynamics and its role on infection for SARS-CoV-2 Spike protein variants. PLoS Comput. Biol. 17, e1009286.
Yin, C., 2020. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics 112, 3588–3596.
Hassan, S.S., Choudhury, P.P., Roy, B., Jana, S.S., 2020. Missense mutations in SARS-CoV2 genomes from Indian patients. Genomics 112, 4622–4627.
Laha, S., Chakraborty, J., Das, S., Manna, S.K., Biswas, S., Chatterjee, R., 2020. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. Infect. Genet. Evol. 85, 104445.
Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., Kiyotani, K., 2020. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J. Hum. Genet. 65, 1075–1082.
Velazquez-Salinas, L., Zarate, S., Eberl, S., Gladue, D.P., Novella, I., Borca, M.V., 2020. Positive selection of ORF1ab, ORF3a, and ORF8 genes drives the early evolutionary trends of SARS-CoV-2 during the 2020 COVID-19 pandemic. Front. Microbiol. 11.
Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., Cui, J., Lu, J., 2020. On the origin and continuing evolution of SARS-CoV-2. Natl. Sci. Rev. 7, 1012–1023.
Emam, M., Oweda, M., Antunes, A., El-Hadidi, M., 2021. Positive selection as a key player for SARS-CoV-2 pathogenicity: insights into ORF1ab, S and E genes. Virus Res. 302, 198472.
Gershoni, J.M., Roitburd-Berman, A., Siman-Tov, D.D., Freund, N.T., Weiss, Y., 2007. Epitope mapping. BioDrugs 21, 145–156.
Irving, M.B., Pan, O., Scott, J.K., 2001. Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics. Curr. Opin. Chem. Biol. 5, 314–324.
Gupta, A.M., Chakrabarti, J., Mandal, S., 2020. Non-synonymous mutations of SARS-CoV-2 leads epitope loss and segregates its variants. Microbes Infect. 22, 598–607.
Gadi, N., Wu, S.C., Spihlman, A.P., Moulton, V.R., 2020. What's sex got to do with COVID-19? Gender-based differences in the host immune response to coronaviruses. Front. Immunol. 11.
Saha, A., Ahsan, M.M., Quader, T.U., Shohan, M.U.S., Naher, S., Dutta, P., Akash, A.S., Mehedi, H.M.H., Chowdhury, A.S.M.A.U., Karim, H., Rahman, T., Parvin, A., 2020. Characteristics, management and outcomes of critically ill covid-19 patients admitted to icu in hospitals in Bangladesh: a retrospective study. J. Prev. Med. Hyg. 62, E35–E44.
Shukla, U., Chavali, S., Mukta, P., Mapari, A., Vyas, A., 2020. Initial experience of critically ill patients with COVID-19 in western India: a case series. Indian J. Crit. Care Med. 24, 509–513.
Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., Wei, G.W., 2021. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Commun. Biol. 4, 228.
Eaaswarkhanth, M., Al Madhoun, A., Al-Mulla, F., 2020. Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? Int. J. Infect. Dis. 96, 459–460.
Becerra-Flores, M., Cardozo, T., 2020. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int. J. Clin. Pract. 74, e13525.
Mercatelli, D., Giorgi, F.M., 2020. Geographic and genomic distribution of SARS-CoV-2 mutations. Front. Microbiol. 11, 1800.
Parvin, R., Afrin, S.Z., Begum, J.A., Ahmed, S., Nooruzzaman, M., Chowdhury, E.H., Pohlmann, A., Paul, S.K., 2021. Molecular analysis of SARS-CoV-2 circulating in Bangladesh during 2020 revealed lineage diversity and potential mutations. Microorganisms 9, 1035.
Saha, S., Malaker, R., Sajib, M.S.I., Hasanuzzaman, M., Rahman, H., Ahmed, Z.B., Islam, M.S., Islam, M., Hooda, Y., Ahyong, V., Vanaerschot, M., Batson, J., Hao, S., Kamm, J., Kistler, A., Tato, C.M., DeRisi, J.L., Saha, S.K., 2020. Complete genome sequence of a novel coronavirus (SARS-CoV-2) isolate from Bangladesh. Microbiol. Resour. Announc. 9 e00568-00520.
Rueca, M., Di Caro, A., Gruber, C.E.M., Messina, F., Giombini, E., Valli, M.B., Lalle, E., Lanini, S., Vairo, F., Capobianchi, M.R., Bartolini, B., 2021. SARS-CoV-2 early screening at the point of entry: travelers from Bangladesh to Italy–July 2020. Front. Genet. 12.
Yoshimoto, F.K., 2020. The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. Protein J. 39, 198–216.