


ORIGINAL ARTICLE

Genome sequencing and analysis of genomic diversity in the locally transmitted SARS-CoV-2 in Pakistan

Muhammad Shakeel¹ | Muhammad Irfan¹ | Zaib un Nisa¹ | Saba Farooq² |
 Noor ul Ain³ | Waseem Iqbal⁴ | Niamatullah Kakar⁵ | Shah Jahan⁶ |
 Mohsin Shahzad⁷ | Saima Siddiqi³ | Ishtiaq Ahmad Khan¹ 

¹Jamil-ur-Rahman Center for Genome Research, Dr. Panjwani Center for Molecular Medicine and Drug Research, ICCBS, University of Karachi, Karachi, Pakistan

²National Institute of Virology, Dr. Panjwani Center for Molecular Medicine and Drug Research, ICCBS, University of Karachi, Karachi, Pakistan

³Institute of Biomedical and Genetic Engineering (IBGE), Islamabad, Pakistan

⁴Pathology Unit, Mardan Medical Complex, Mardan, Pakistan

⁵Center for Advanced Studies in Vaccinology and Biotechnology (CASVAB), University of Balochistan, Quetta, Pakistan

⁶Department of Immunology, University of Health Sciences Lahore, Lahore, Pakistan

⁷Department of Molecular Biology, Shaheed Zulfiqar Ali Bhutto Medical University, Islamabad, Pakistan

Correspondence

Ishtiaq Ahmad Khan, Jamil-ur-Rahman Center for Genome Research, Dr. Panjwani Center for Molecular Medicine and Drug Research, ICCBS, University of Karachi, Karachi 75270, Pakistan.

Email: ishtiaqchemist@gmail.com

Muhammad Shakeel, Muhammad Irfan, and Zaib un Nisa contributed equally to this study.

Funding information

The Searle Company Limited (TSCL), Pakistan; Government of Sindh, Pakistan

Abstract

Surveillance of genetic diversity of the SARS-CoV-2 is extremely important to detect the emergence of more infectious and deadly strains of the virus. In this study, we evaluated mutational events in the SARS-CoV-2 genomes through whole genome sequencing. The samples were collected from COVID-19 patients in different major cities of Pakistan during the four waves of the pandemic (May 2020 to July 2021) and subjected to whole genome sequencing. Using *in silico* and machine learning tools, the viral mutational events were analyzed, and variants of concern and of interest were identified during each of the four waves. The overall mutation frequency (mutations per genome) increased during the course of the pandemic from 12.19 to 23.63, 31.03, and 41.22 in the first, second, third, and fourth waves, respectively. We determined that the viral strains rose to higher frequencies in local transmission. The first wave had three most common strains B.1.36, B.1.160, and B.1.255, the second wave comprised B.1.36 and B.1.247 strains, the third wave had B.1.1.7 (Alpha variant) and B.1.36 strains, and the fourth waves comprised B.1.617.2 (Delta). Intriguingly, the B.1.36 variants were found in all the waves of the infection indicating their survival fitness. Through phylogenetic analysis, the probable routes of transmission of various strains in the country were determined. Collectively, our study provided an insight into the evolution of SARS-CoV-2 lineages in the spatiotemporal local transmission during different waves of the pandemic, which aided the state institutions in implementing adequate preventive measures.

KEYWORDS

COVID-19 pandemic, genetic evolution, SARS-CoV-2 lineages, spatiotemporal surveillance, viral variants

1 | INTRODUCTION

Since its emergence in Wuhan, China in December 2019, the SARS-CoV-2 has mutated continuously, resulting in the emergence of several more contagious variants of the virus, which have modulated the dynamics of the COVID-19 pandemic. The rapid evolution of the

virus during the pandemic, due to accumulation of mutations, has contributed to viral adaptation, drug resistance, and higher transmissibility of more virulent strains (Pachetti et al., 2020). However, very few mutations have been functionally characterized (Zhou & Wang, 2021). A massive genome sequencing drive is under way globally to document the genetic diversity of SARS-CoV-2. Several studies have

reported large number of mutations in various genes, including S, M, E, N, ORF1ab, ORF3a, ORF6, ORF7, ORF8, and ORF10 (Rahimi et al., 2020).

Genomic surveillance of a virus after it enters a naive population is crucial for designing effective strategies for disease control and prevention (Ladner et al., 2019). Continuous SARS-CoV-2 genomic characterization has aided the global health policy makers to determine and declare new mutations being variants of concern (VOC) and variants of interest (VOI) for better management and curtailing the transmissibility of the infection. The findings of such studies have supported contact tracing, social distancing, and travel restrictions to limit the spread of SARS-CoV-2. In a study conducted in Northern California from late January to mid-March 2020, using samples from 36 patients spanning nine counties and the Grand Princess cruise ship, phylogenetic analyses revealed the cryptic introduction of at least seven different SARS-CoV-2 lineages into California (Deng et al., 2020). Genetic surveillance of COVID-19 in Malaysia and other Asian countries has highlighted the presence of B.6 lineage in the Asia Pacific region (representing 95% of the world cases of B.6 strains) (Chong et al., 2020). Genomic-based surveillance of COVID-19 cases in Beijing, China till May 2020 has revealed transmission of SARS-CoV-2 in the city via three routes including Wuhan exposure group, foreigner imported cases, and locally transmitted cases (Du et al., 2020). In addition, genomic analysis of SARS-CoV-2 has indicated two separate introductions of the virus into the Republic of Congo during the first wave (April to July 2020) (Ntoumi et al., 2021). Phylogenetic investigation of 11,422 SARS-CoV-2 genomes (between January and September 2020) revealed 287 introductions into Washington, including 61% of the introduction from a source other than the United States (Tordoff et al., 2021).

Pakistan shares borders with world's most densely populated nations with strong movement of people from and to the hotspots of COVID-19. The country confirmed its first COVID-19 patient on February 26, 2020 in the southern city of Karachi. By 29th October 2021, there were 1,271,687 confirmed cases, 28,431 deaths, while 18.15% of the population was fully vaccinated in Pakistan (John Hopkins University). The Pakistani government adopted progressive disease prevention initiatives to restrict social contacts, reduce the spread of viruses, and avoid community-based transmissions. Pakistan witnessed a peak in June 2020 and cases fell from thousands to a few hundred per day soon in September. This was mostly due to the excellent containment strategies adopted by the Government of Pakistan to successfully contain the spread of the virus, which was due appreciated by the World Health Organization (WHO, 2020).

Given the specific demography and environmental conditions, the present study attempted to gain an insight into the mutational spectrum of SARS-CoV-2 genome in Pakistan. We performed 97 SARS-CoV-2 genome sequencings of virus isolated from patients from different major cities of the country during all the four waves of the pandemic and, by incorporating additional publicly available SARS-CoV-2 data from Pakistan, comprehensively analyzed the viral evolution. Spatiotemporal approach was adopted and samples were collected during May to August 2020, November to December 2020, March to April 2021, and July 2021. The focus of the epidemiological analysis

was to identify specific patterns of SARS-CoV-2 transmission through genomic analysis within local population before and during the containment stage of COVID-19 pandemic. Findings of this study helped to devise strategies for the surveillance of potential transmission routes to contain future outbreaks.

2 | MATERIALS AND METHODS

2.1 | Ethical consideration, recruitment of patients, and samples collection

The study was approved by the Research Ethics Committee of the International Center for Chemical and Biological Sciences, University of Karachi, and the study design adhered to the ethical considerations of the Declaration of Helsinki (Helsinki, 2013). For the current study, more than 1000 patients were recruited from different hotspot areas of the country during all the four waves of the infection. During the first wave, the samples were collected in Karachi from May to August 2020. During the second wave, the samples were also collected in the capital of the country, Islamabad, in addition to Karachi, from November to December 2020. During the third wave, further cities of the country were added, and the samples were collected from Karachi, Islamabad, Lahore, Mardan, and Quetta. The fourth wave samples were collected in Karachi only. For COVID-19 testing, nasopharyngeal swabs were collected in viral transport medium (VTM) according to the guidelines of the Center for Disease Control and Prevention (CDC, 2020). Only the patients who tested positive for COVID-19 by real-time PCR using SARS-CoV-2 specific primers and probes were proceeded for downstream investigation. Total RNA was isolated from the VTM in a biosafety level 3 (BSL-3) laboratory using the QIAamp viral RNA mini kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. Concentration of the total RNA was determined with a Qubit fluorometer using the Qubit RNA HS assay kit (Thermo Fisher Scientific, MA, USA).

2.2 | Complementary DNA synthesis

A total of 245 RNA samples were randomly selected for synthesizing double-stranded complementary DNA (cDNA) from the purified RNA. The samples with low threshold cycle (Ct) values in the real-time PCR assay ($Ct \geq 25$) were selected for library preparation. The double-stranded cDNA was synthesized from the total RNA by using Maxima H Minus Double-Stranded cDNA Synthesis Kit (Thermo Fisher Scientific) according to the manufacturer's protocol. For the first strand cDNA synthesis, the isolated total RNA and random hexamer primers were incubated at 65°C for 5 min followed by the addition of the 4X First Strand Reaction Mix along with the First Strand Enzyme Mix. The reaction mixture was incubated at 25°C for 10 min, 50°C for 30 min, followed by termination of the reaction at 85°C for 5 min. The second strand cDNA synthesis was performed by adding the 5X Second Strand Reaction Mix and the Second Strand Enzyme Mix to the first

strand cDNA synthesis reaction mixture. Final volume was adjusted with nuclease free water, and the reaction mixture was incubated at 16°C for 60 min. The reaction was stopped by adding 0.5 M EDTA. The double-stranded cDNA was purified by using the Agencourt AMPure XP beads (Beckman Coulter, CA, USA), and the concentration of the double-stranded cDNA was evaluated using the Qubit DNA HS assay kit (Thermo Fisher Scientific).

2.3 | DNA library preparation and whole genome sequencing

A total of 134 cDNA samples were selected, which passed the quality assessment criteria for performing the SARS-CoV-2 whole genome sequencing. Paired-end libraries were constructed from the double-stranded cDNA by using Illumina DNA Prep with Enrichment kit (Illumina Inc., San Diego, CA, USA) following the manufacturer's protocol. Briefly, the tagging of purified double-stranded cDNA (50 ng) was performed by using bead linked transposomes followed by adapters ligation to the tagged DNA. Unique indexes (IDT for Illumina DNA/RNA UD Indexes) were added to each tagged DNA in limited cycles of PCR. The amplified tagged DNA was purified using sample purification beads. Prior to the enrichment of SARS-CoV-2 genome, the libraries were pooled in 12 plex reactions (multiplexing of 12 samples). DNA fragments of the SARS CoV-2 genome were hybridized with biotinylated respiratory virus oligos (Illumina Inc.). The DNA fragments hybridized with the custom oligos were captured using streptavidin magnetic beads. The enriched library was amplified followed by purification with Agencourt AMPure XP beads (Beckman Coulter). The concentration of the enriched libraries was determined using Qubit DNA HS assay kit (Thermo Fisher Scientific). The libraries were denatured with 0.2 N NaOH followed by dilution to 12 pmole using the hybridization buffer (HT1). Paired-end sequencing (2 × 75 bases) using MiSeq reagent v3 kit was carried out on Illumina MiSeq (Illumina Inc.).

2.4 | Analysis of the sequencing data

The raw data in the binary base call format (.bcl) were converted into fastq format on the MiSeq instrument. The quality of the DNA short reads was assessed using FastQC tool (Andrews, 2010). The short reads were aligned with the reference SARS-CoV-2 genome of the isolate from Wuhan, China (Wuhan-Hu-1 genome, GenBank accession NC_045512) using the BWA-MEM algorithm (Li, 2013). For post-alignment processing, the sequence alignment map (SAM) was converted into binary format (BAM) using the Samtools package (Dhanapany et al., 2009). Among the paired-end reads mapping to same coordinates on the reference genome, only one paired-end read was retained at a given position. The base quality score recalibration (BQSR) and realignment of insertion/deletion (indels) were carried out employing the best practices of the Genome Analysis Toolkit (GATK) (DePristo et al., 2011). The samples with coverage of ≥85% were processed for downstream analysis. The variants with low quality score

(QUAL < 30) were filtered out. To determine functional consequences of the variants, a standalone perl-based algorithm ANNOVAR was employed using the Wuhan-Hu-1 reference genome-based annotation dataset (Yang & Wang, 2015).

2.5 | Assessment of the mutational frequencies

The mutation frequency (Mf) in the sequenced SARS-CoV-2 genomes was determined by calculating the average number of mutations in each genome, as $Mf = \sum M / \sum G$, where M is the number of total mutations, and G is the number of genomes sequenced. As, for a single base substitution mutation, a nucleotide base at a given site in the genome can be replaced, with respect to the reference genome, with either of the three other bases, so we calculated the possible 12 distinct mutational types, that is, A>C, A>G, A>U, C>A, C>G, C>U, G>A, G>C, G>U, U>A, U>C, and U>G in the sequenced genomes. Further, these mutational types were also determined for different functional categories of the mutations, that is, non-coding, silent, missense, and nonsense. In addition, in order to categorize the mutations on the basis of their number of descendants (passed into samples of later dates), we also determined the type of mutations with zero descendants, with few descendants (2–3), and with many descendants (>3). The rationale for categorizing the identified mutations on the basis of their number of descendants was to determine the mutations of different functional consequences transmitted in the local population given the specific demographic and environmental factors of the region (De Maio et al., 2021). The transitions (Ti) to transversions (Tv) ratio was calculated as $\sum Ti / \sum Tv$, as stated previously (Roy et al., 2020).

2.6 | Phylogenetic analysis

For building the whole genome, the consensus sequences were generated from the binary alignment map (bam) files, as described previously (Sah et al., 2020). For comparison and validation, DNA short reads were assembled through de novo assembly with Velvet 1.0.0 (Zerbino, 2010) and SPAdes (Prjibelski et al., 2020) tools using the default parameters. For inferring phylogenetic relationship, the assembled genomes were aligned with the Wuhan-Hu-1 genome using the MUSCLE multiple sequence alignment tool (Edgar, 2004). The phylogeny was constructed using the RAxML 8.2.12 tool (Stamatakis, 2014) by employing the maximum likelihood (ML) algorithm and 100 bootstrap replicates. The phylogeny tree was further manipulated with MEGA X (Kumar et al., 2018) and FigTree v.1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>) tools.

For determining epidemiologically important SARS-CoV-2 variants, a previously proposed dynamic nomenclature system for assigning lineages of the SARS-CoV-2 by Rambaut et al. (2020) was employed using the Phylogenetic Assignment of named Global Outbreak Lineages (PANGOLIN 2.0) software suite. According to this nomenclature system, the genomes having two characteristic mutations (8782C>T, and 28144T>C) with respect to Wuhan-Hu-1 genome were designated as Lineage A, and the genomes lacking these two mutations were

designated as Lineage B. The sequences of Lineage A have been shown to be more closely related with the bat corona virus. The Wuhan-Hu-1 genome relates to the Lineage B, as proposed by Rambaut et al. (2020), and the Lineage B having D614G mutation in the spike protein gene is categorized into sub-lineage B.1. The emergence of further recurrent mutations in B.1 lineages divides the genomes into further sub-lineages such as B.1.1, B.1.36, B.1.36.6, and so on.

3 | RESULTS

3.1 | Description of the cohort

We selected more than 1000 COVID-19 patients from the public sector hospitals during the peak time of four waves of the pandemic (from May to August 2020 [first wave], November to December 2020 [second wave], February to April 2021 [third wave], and July 2021 [fourth wave]). The samples of the first wave were collected from the COVID-19 patients in the metropolitan city of Karachi (Sindh Province). During the second wave, in addition to Karachi, the capital city Islamabad, which is ~1140 km away from Karachi, was also included for samples collection. For the samples of third wave, in addition to Karachi, and Islamabad, other major cities of Pakistan, that is, Lahore (Punjab province), Mardan (Khyber Pakhtunkhwa province), and Quetta (Baluchistan province), were also included. The fourth wave samples were collected in Karachi. The details of samples collected in different cities of Pakistan have been presented in Table S1. The samples were included in the study after confirmed diagnosis with a real-time PCR test from the nasopharyngeal swab specimens. The disease symptoms varied in the patients including mild symptoms of low-grade fever and flu to moderate symptoms of fever, flu, and difficulty in breathing.

3.2 | Genetic characteristics of the SARS-CoV-2

To determine the genomic characteristics and diversities in the virus responsible for the COVID-19 transmission in Pakistan, deep SARS-CoV-2 genome sequencing was employed using a custom oligos panel designed for the enrichment of respiratory virus sequences. Through massively parallel sequencing, we obtained, after removing the samples in which viral genome coverage was below 85%, 97 complete or near to complete viral genomes with average depth of coverage of >1000X. After alignment of the short reads with the reference SARS-CoV-2 genome of Wuhan isolate (Wuhan-Hu-1, NC_045512.2), the genetic variations were obtained employing the on-instrument default pipeline of BWA, Samtools, and GATK tools. After filtering out the low-quality mutations (mutations with QUAL < 30), the first wave samples contained cumulative 451 mutations at 122 genomic sites (including 120 single nucleotide variation [SNV] sites, and two deletion sites) (Table S2), with a mutation frequency of 12.19 mutations per sample. The samples from the second wave contained 520 mutations at 214 genomic sites (including 210 SNV sites, and two insertion sites), with a mutation frequency of 23.63 mutations per genome (Table S3). In the

third wave samples, there were 900 mutations at 247 genomic sites (including 237 SNV, one insertion, and nine deletion sites), with a mutation frequency of 31.03 mutations per genome (Table S4). In the fourth wave samples, there were 376 mutations at 84 sites, with a mutation frequency of 41.22 (Table S5). Analysis of the site frequency mutational spectrum showed that the increased mutational load (mutations/genome) was due to emergence of large number of singletons (Figure 1). The proportion of singletons to recurrent mutations sites was 3.08 in the first wave samples, and 9.44 and 9.42 in the second and third wave samples, respectively.

Great diversity in mutational sites among the genome assemblies was observed. There were, on average and median mutational events per genome, 12.19 (SD \pm 3.57), and 12 (IQR 11–14) in the first wave samples, 23.17 (SD \pm 5.5) and 24 (IQR 20–27) in the second wave samples, 31.03 (SD \pm 8.13) and 35 (IQR 26–37) in the third wave samples, and 41.22 (SD \pm 2.22) and 41 (IQR 39–43) in the fourth wave samples, respectively. The genome-wide non-synonymous/synonymous ratio observed was 1.54 in the first wave samples, this ratio decreased to 1.52 and 1.35 in the second wave and third wave samples, respectively, but increased to 2.67 in the fourth wave genomes. Each of these ratios is lower than the precedented non-syn/syn ratio (1.88) in previous global scale studies (van Dorp et al., 2020). Further bifurcation indicated that the non-syn/syn ratio at singleton and polymorphic sites was 1.48 and 1.52 in the first wave, 1.56 and 1.05 in the second wave, 1.12 and 2.33 in the third wave, and 1.86 and 3.8 in the fourth wave samples. These analyses indicated that the SARS-CoV-2 continued to acquire new non-synonymous mutational sites and raised their frequencies within the studied cohort.

3.3 | Mutational descendants

Meta-analysis of large number of SARS-CoV-2 genomes across multiple countries has demonstrated that one SARS-CoV-2 genome differs from the Wuhan-Hu-1 reference strain (NC_045512.2) at maximum of 32 sites (van Dorp et al., 2020). We constructed mutational landscape to decipher the types of mutations, and genes which are more recurrently mutated than the others. This analysis highlighted that the C>U mutations were predominantly higher in the genomes of all the four waves (Wilcoxon rank-sum p < .05), followed by G > U, and U>C mutations (Figure S1). The deamination of nucleotide bases, primarily the cytosine, and conversion into uracil by the host RNA editing system such as apolipoprotein B mRNA editing-enzyme, catalytic polypeptides (APOBECs), is a typical hall mark of SARS-CoV-2 genome (Kosuge et al., 2020). Notably, the higher proportion (18%) of U>C in the third wave genomes is the first to be observed here. We further classified these mutation types on the basis of number of descendants. There was almost similar pattern of mutations with no descendants, few descendants, and many descendants among the genomes of the three waves (Figure S2). The pattern of C>U transitions was found predominantly higher in the mutations with no descendants, few descendants, and many descendants, indicating that this higher proportion was not due to the sequencing artefacts. Variations in mutational dynamics were

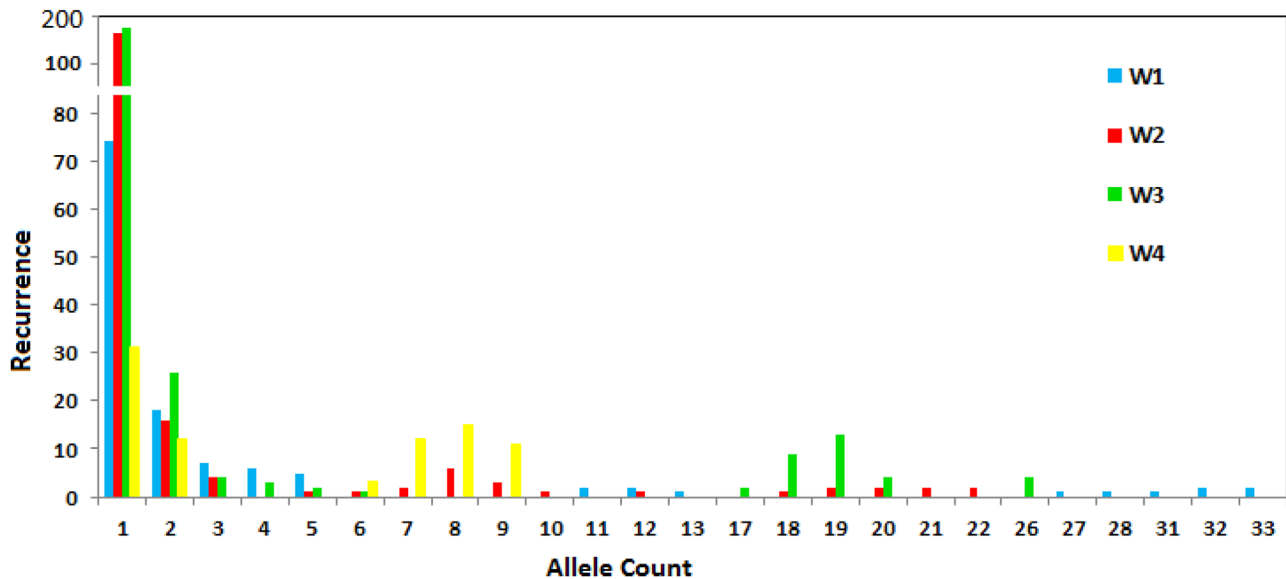


FIGURE 1 Allelic frequency spectrum of the mutations identified in the SARS-CoV-2 genomes of this study. The larger proportion of mutations comprised singletons. Notably, more numbers of highly recurrent mutations were observed in the third and fourth waves genomes

also observed among the genomes of the three waves. For example, the proportion of G>U alterations was higher in the mutations with many descendants in the first wave genomes, whereas it decreased in the second and third wave samples. Contrarily, the A>U transversions were observed in the mutations without descendants only in the first wave genomes, whereas their proportion increased in mutations with many descendants in the second and third wave samples. The C>A transversion was not observed in the mutations with many descendants in the samples of the first and second waves, whereas considerable proportion of C>A was observed in mutations with many descendants in the third wave samples.

3.4 | Mutational landscape

In addition to the mutational pattern, to determine specific mutations, which were found recurrently in certain genes, mutational landscape plots were constructed. In the first wave genomes, two genes, that is, Spike protein and non-structural protein 3 (nsp3) contained the highest number of recurrent mutations, that is, 35 mutational incidences each (Figure 2a). Notably, the spike protein contained highest number of homoplasic sites (six mutations) including two missense SNVs, that is, the D614G mutation observed in 32 samples (86.5% of the cohort), and Q677H mutation observed in four samples (10.8% of the cohort). The other highly recurrent variations included an upstream 5'-UTR mutation 241:C>T (observed in 33 samples), a silent mutation F924F in nsp3 (observed in 31 samples), a missense mutation in nsp12 P4715L (observed in 31 samples), a silent mutation L227L in nsp13 (observed in 12 samples), a silent mutation L280L in nsp14 (observed in 28 samples), silent mutations D294D and G880G in spike protein (observed in 12 and 13 samples, respectively), a missense mutation Q57H in orf3a (observed in 31 samples), silent mutation Y71Y and missense muta-

tion H125Y in membrane protein (observed in 27 and four samples, respectively), a stop-gain mutation E39X in orf7b (observed in four samples), and missense mutations S194L and R209I in nucleocapsid protein (observed in 11 samples each).

In the second wave samples, nsp3, N, and M genes contained the highest number of mutational events (23 mutations each), followed by S, nsp14, and nsp12 genes (22 mutations) (Figure 2b). Here, the number of homoplasies was the highest in N gene (six mutations) including the S194L mutation in 19 genomes, followed by the S gene (five mutations). In the spike protein, two missense homoplasies, that is, R102S, and A222V, in addition to the D614G, were observed in eight genomes each. In N gene, S194L was observed in 19 genomes, and in M gene, silent mutation Y71Y was found in 20 genomes.

In the third wave samples, the highest number of mutations were observed in the S, and nsp3 genes (39 and 35 mutational sites respectively), followed by nsp12 (28 mutations), nsp2, and N genes (27 mutations each) (Figure 2c). Many mutations being reported having considerable significance in viral transmission were also observed, for example, in the spike protein, p.68_70del, and N501Y were observed in 19 genomes, and p.143_144del and A570D were observed in 18 genomes, in addition to the D614G being observed in all the genomes. The p.68_70del and N501Y mutations are the characteristic mutations of the UK variant (B.1.1.7). The South African strain (B.1.351) specific mutation E484K was observed in two genomes. The fourth wave comprised Delta variant of the virus, and all the samples contained characteristic mutations of the Delta variant, including T19R, G142D, 156_158del, L452R, T478K, D614G, P681R, and D950N mutations of the spike protein.

An amalgamated view of the mutations identified in all the four waves genomes revealed that the genomes of third and fourth waves harboured higher proportion of high-frequency (AF > 0.75) mutations (Figure 3). The high frequency mutations were assessed for either

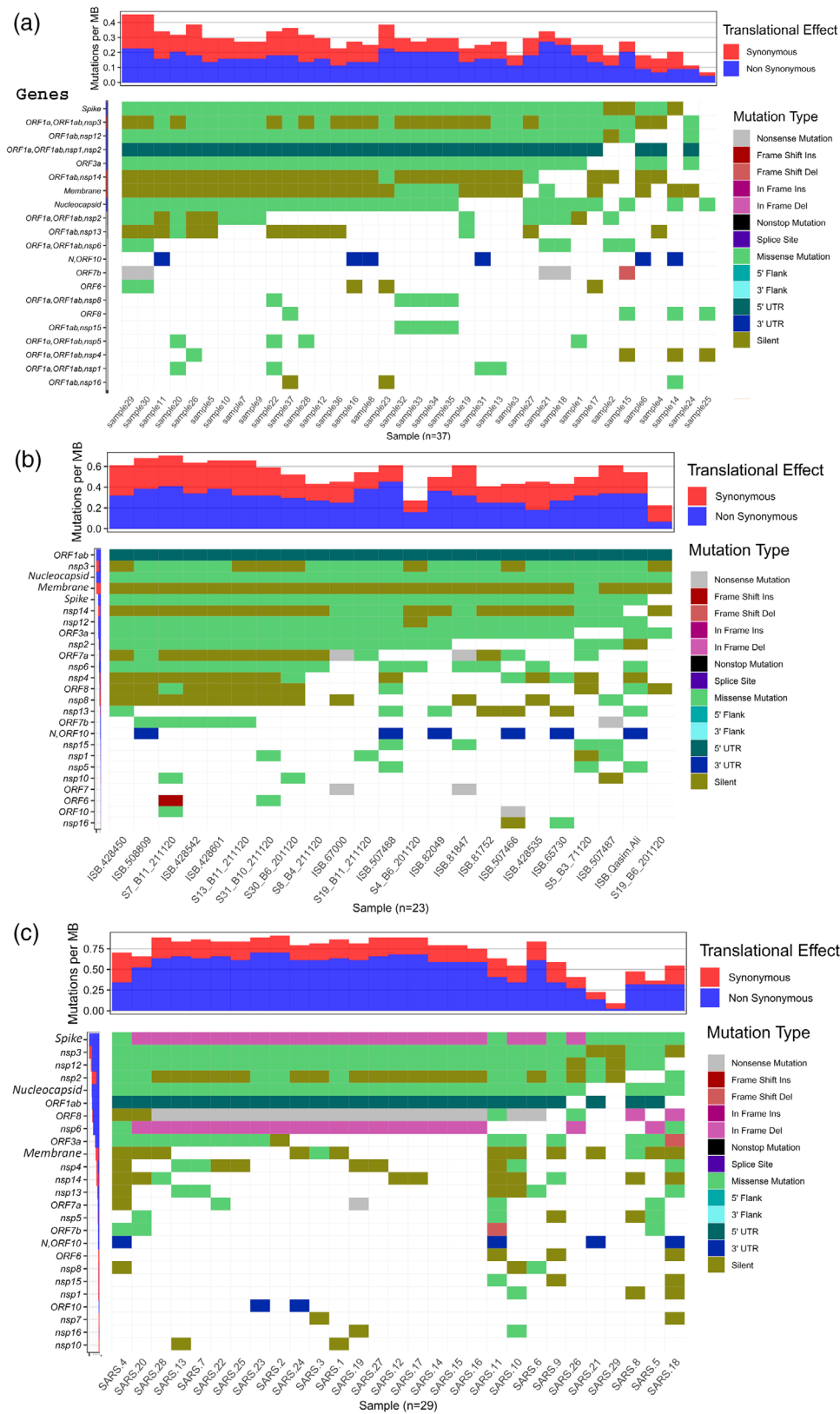


FIGURE 2 Mutational landscape, recurrence of mutations in the SARS-CoV-2 genomes of different waves

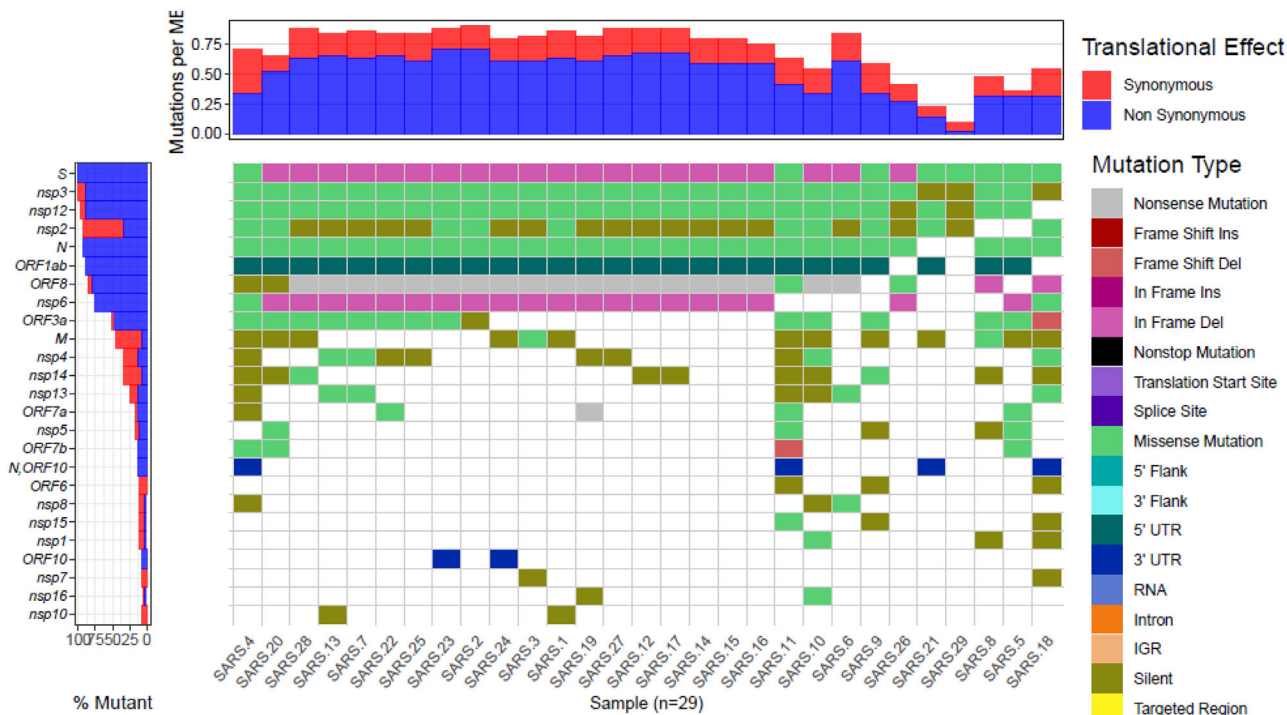


FIGURE 2 Continued

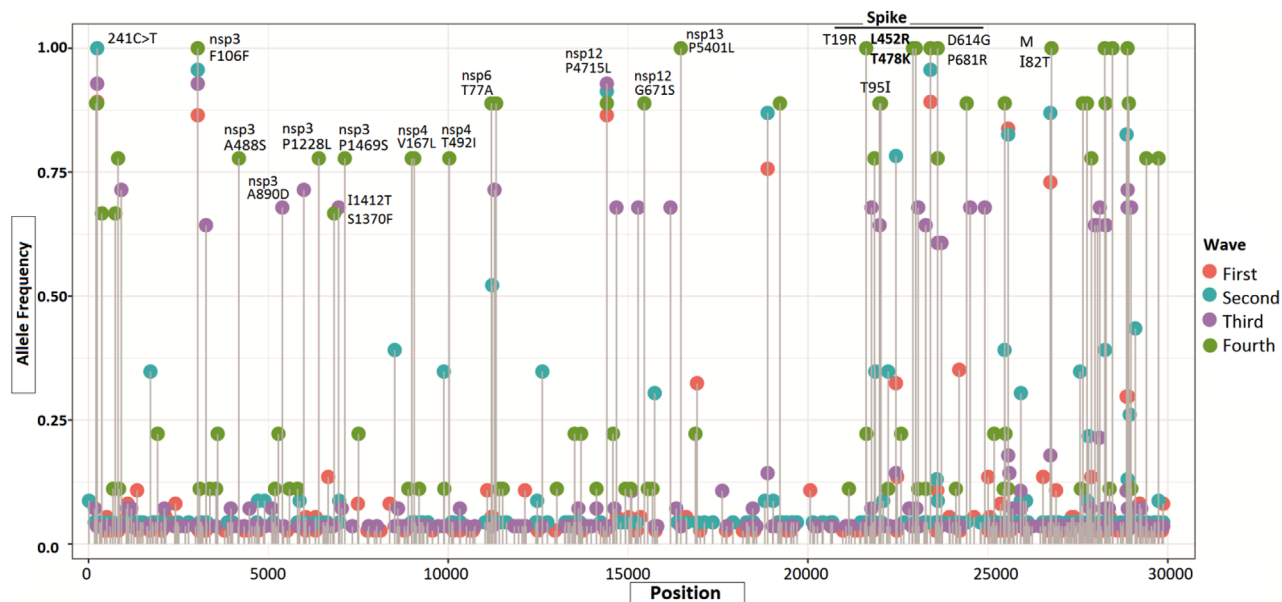


FIGURE 3 Lollipop plot indicating the high-frequency variants in all the four waves. Notably, the fraction of high-frequency mutations (allele frequency > 0.75) was greater in the genomes of the third and fourth waves, largely due to the local transmission of the Alpha (UK) and Delta (Indian) variants, respectively

VOC, VOI, variants under monitoring (VUM), or de-escalated variants according to the criteria set by the European Centre for Disease Prevention and Control (ECDC) (European Centre for Disease Prevention and Control, 2021). This manual curation revealed that the first wave genomes contained a VOC D614G (recurrence 85.6%) and a

VUM Q677H (10.8%). In the second wave genomes, the VOC D614G (recurrence 95.6%), a VOI P681H (4.3%), and the VUM Q677H (4.3%) were observed. In the third wave genomes, six VOCs were detected including the D614G (recurrence 100%), K417N (7.1%), L452R (3.6%), E484K (7.1%), N501Y (67.8%), and A701V (3.6%). Among the VOI, two

mutations including P681H (60.7%) and S477N (3.6%) were observed. Among these mutations, the D614G, P681H, K417N, L452R, E484K, and N501Y have also been declared as variants under monitoring status by the ECDC. The fourth wave genomes carried all the VOC of the Delta variants including the L452R, T478K, D614G, and P681R mutations in all the samples, whereas the VOI P681H was observed in 37.8% samples.

3.5 | Phylogenetic relationship and lineages in SARS-CoV-2

Given that the samples were collected in different cities of the country, we conducted a phylogenetic analysis to delineate clustering among the SARS-CoV-2 genomes and finding the likely route of transmission into the country. The phylogenetic analysis revealed that the majority of the samples of all the four waves belonged to various strains/descendants of the B lineages according to the taxonomic nomenclature proposed by Rambaut et al. 2020 (Figure S3).

Among the genomes of the first wave, one genome was of the parental B lineage and had highest similarity with the Wuhan-Hu-1 genome, two genomes were of B.6 (Singapore lineage), whereas remaining 33 were either of B.1 lineage (3 genomes) or further descendants of B.1 lineage. The descendants of B.1 comprised three high prevalent lineages, that is, B.1.36 (10 genomes), B.1.160 (11 genomes), and B.1.255 (five genomes). On average, the B.1.36 and B.1.160 lineages contained higher number of mutational sites per genome compared with the parental B.1 and B.1.255 lineages (Figure S4). We retrieved >400 SARS-CoV-2 whole genome sequences of similar time stretch of other countries including China, Iran, Afghanistan, India, the United Arab Emirates, Saudi Arabia, Qatar, Israel, Italy, France, the United Kingdom, and the United States from the Global initiative on sharing all influenza data (GISAID) repository (Shu & McCauley, 2017), and constructed phylogenetic tree to infer the routes of transmission of virus in the country (Figure S5). This analysis revealed two monophyletic clusters comprising of 14 and 6 genomes, respectively. The first monophyletic group comprised predominantly of B.1.160 genomes (12) and two B.1.255 genomes. The second monophyletic group contained B.1.36.6 genomes. The B.1.160 genomes appeared between the monophyletic groups comprising genomes of Saudi Arabia and France. The divergence of B.1.160 from the most recent common ancestor (MRCA) appeared a bit earlier than the divergence of SARS-CoV-2 genomes from France and Saudi Arabia (Figure 4). The B.1.36.6 genomes were found in close proximity of India and UK sub-clades. Four B.1.36 genomes appeared at a single branching point near Saudi Arabia. The B.1 genomes seemed to have two different introductions, where two genomes appeared in close proximity of Saudi Arabia and European countries (Italy, France, and the United Kingdom). The B.1.1 strain was observed in close proximity of genomes from the United Kingdom, Italy, and Saudi Arabia. The SARS-CoV-2 genome of lineage A appeared in the sub-clade evolved from Chinese SARS-CoV-2 genomes. The B.6 genome was observed in a distinct sub-clade comprising genomes from India.

The SARS-CoV-2 genomes of the second wave comprised B lineages only, with predominant proportion of B.1.36 variants (14 genomes), followed by B.1.1 (two genomes), B.1.247 (two genomes), and one each of B, B.1.1.1, and B.1.160 lineages. The SARS-CoV-2 whole genome sequences (>400) of the countries were the same as in first wave analysis and retrieved from GISAID repository for phylogenetic analysis (Figure S6). Through this analysis, at least seven distinct introductions of SARS-CoV-2 into the two cities of Pakistan were observed. Notably, most of the genomes of B.1.36 and B.1.36.6 lineages (13 genomes) appeared in one monophyletic cluster in close proximity of the United Arab Emirates and Indian sub-clades (Figure 5). The B.1.247 strain of Islamabad was observed near to a sub-clade of Iran, the B.1.160 strain of Islamabad appeared in a different sub-clade of India, and the B.1.36 strain in Karachi appeared in close proximity of sub-clades of Iran and India. One B.1.1 genome in Karachi was in close proximity of sub-clades of Iran, Israel, and the United Arab Emirates, and the other B.1.1 genome was found in close proximity of Indian and Israeli sub-clades.

The genomes of the third wave comprised primarily of the UK variant of the virus (B.1.1.7), accounting for 69% of the genomes. Other lineages included three genomes of B.1.36 strains, one genome of South African variant (B.1.351), one genome of B.1.468 lineage, one genome of B.1.1.413 lineage, and one genome of A.27 variant. It was important to note that genomes of one city, Quetta, did not contain the Alpha variant. This city, the capital of Baluchistan province, lies in the North-West of the country and has little to and fro movement with other provinces of the country. The South African variant was observed in Karachi, the metropolitan city of Pakistan, which caters to major shipping activities of the country through commercial ports. To find out potential route of transmission of these variants, we conducted ML phylogeny by taking into account SARS-CoV-2 genome sequences of other countries, as mentioned above (Figure S7). The three B.1.36 genomes, one each from Lahore, Quetta, and Mardan, clustered between the Indian nodes (Figure 6). The South African variant detected in Karachi was found clustered with SARS-CoV-2 genomes of France, England, and Italy. Among the B.1.1.7 variants in Lahore and Mardan, one genome in each of the cities showed close proximity of India and England. The other B.1.1.7 genomes in Karachi, Lahore, and Mardan were clustered near the SARS-CoV-2 nodes of England, Italy, and the United States. These findings suggested two major routes of transmission of the third wave virus of the pandemic into the country, that is, India and Europe.

We also determined the viral strains according to the Nextstrain clades (Aksamentov et al., 2021). The predominant clade in the first and second waves was 20A, whereas the third wave was the result of 20I (Alpha) variant, and for the fourth wave 21J (Delta) variant was responsible (Figures S8, and S9).

4 | DISCUSSION

Understanding the evolution of SARS-CoV-2 is a fundamental effort in coping with the COVID-19 pandemic. The reason for the rapid evolution of the SARS-CoV-2 is that it is infecting millions of people of

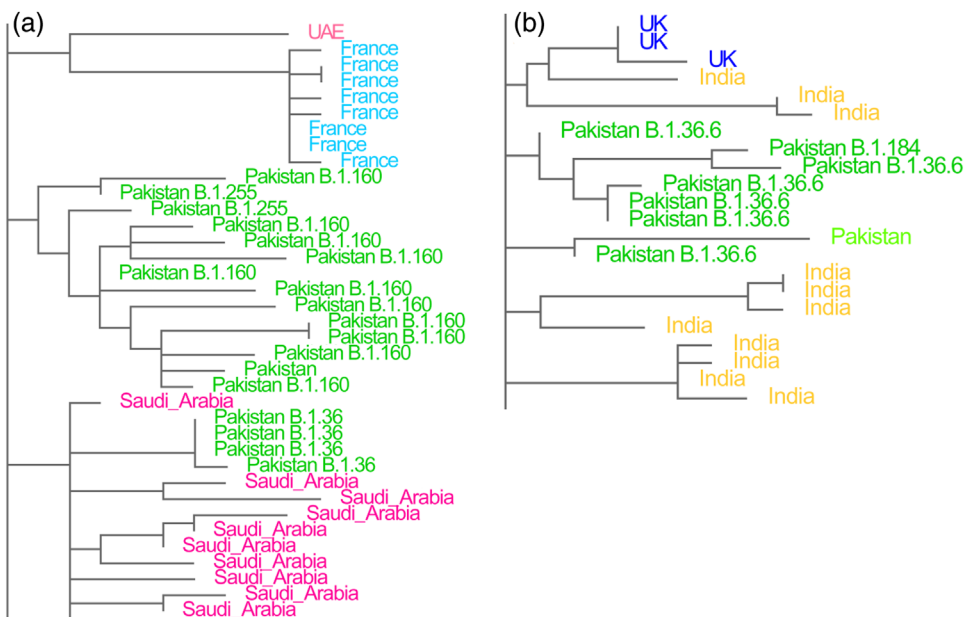


FIGURE 4 Phylogenetic relationship and likely routes of transmission of SARS-CoV-2 in Karachi during the first wave (May to August 2020) of the COVID-19 pandemic

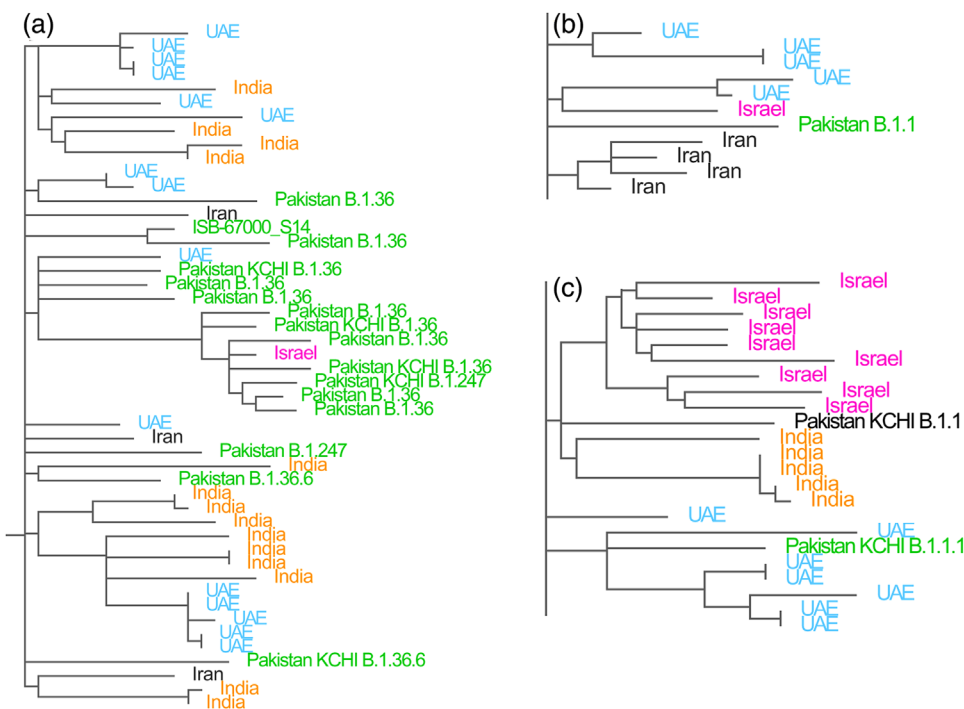


FIGURE 5 Phylogenetic relationship and likely routes of transmission of SARS-CoV-2 in Karachi and Islamabad during the second wave of COVID-19

diverse ethnicities of the world (Azgari et al., 2021). Several studies have demonstrated that new mutations affecting transmissibility and immune escape are the result of long-term infection in immune suppressed patients (Greaney et al., 2021). Mutant spectrum dynamics has a decisive impact on virus behaviour (Gregori et al., 2016).

We observed fluctuation in the mutation frequencies of viral lineages defining mutations (Figure 3; Figure S4). Notably, the frequency of synonymous mutation F106F (nsp3) raised from 85% in first wave to 100% in the fourth wave. The frequency of D614G (Spike) was 86% in the first wave, and became 100% in the fourth wave genomes. The D614G

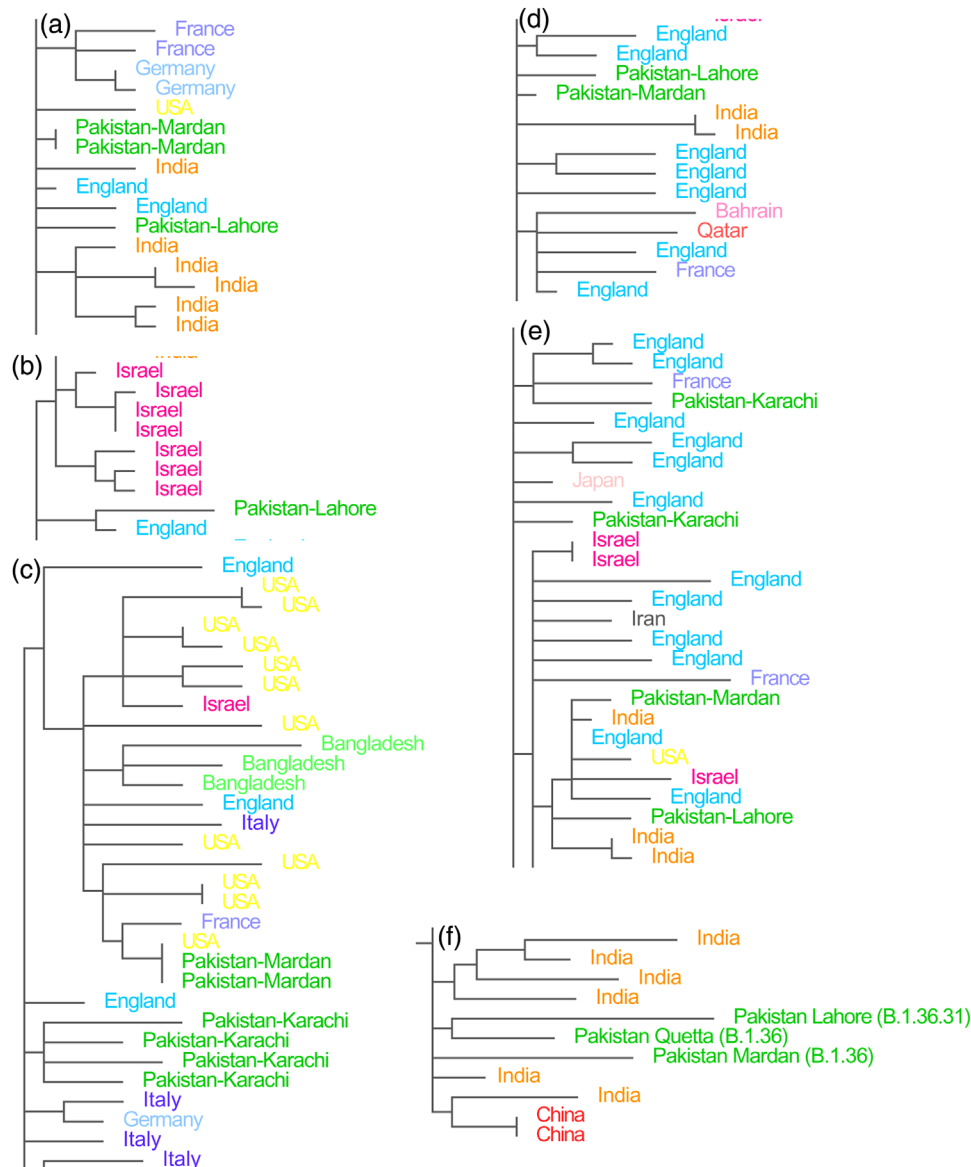


FIGURE 6 Phylogenetic relationship and likely routes of transmission of SARS-CoV-2 in Pakistan during the second wave of COVID-19

has been reported to alter SARS-CoV-2 fitness (Plante et al., 2020). The L452R was observed in the fourth wave samples only and was recently found to enhance viral fusogenicity and infectivity as well as host glycolysis (Zhang et al., 2022). Another mutation P4715L (nsp12) was observed with higher mutation frequency in the second and third waves; however, its frequency decreased in the fourth wave, indicating a purifying selection.

Pakistan, being in the subtropical belt, and sharing borders with the largest populations of the world, i.e., China, and India, has witnessed four waves of the pandemic so far. In this context, the present study is presenting the first comprehensive report on genetic diversity and evolution in the country by carrying out temporal whole genome sequencing of the SARS-CoV-2 isolated in different cities of the country during all the four waves of the pandemic. The increasing mutation frequency with the passage of time, presented here, correlates with

the previous observations (Martinez et al., 2020). It was noteworthy that a predominant lineage was observed in each pandemic wave, suggesting advantageous properties of the predominant lineage over other contemporary co-circulating variants (Andrés et al., 2022) (Figure S3). To the best of our knowledge, this is the first multi-wave comprehensive study deducing potential variants of interest and variants of concern, and viral transmission in the country.

During the first wave, the three types of B.1 sub-lineages seem to have spread in the metropolitan city of Karachi during independent transmission events. The B.1.36 is a global lineage with lots of representation of sequences from India, Saudi Arabia, and the United Kingdom (Ishtiaque et al., 2020; Joshi et al., 2020). The genomes of B.1.36 variant of the present study also appeared in the clade of Saudi Arabia and India (Figure 4). The B.1.160 lineage is the recent split from B.1.36 lineage and has mostly been reported from the European

countries with major representation in the United Kingdom followed by Denmark, France, and Switzerland (Rambaut et al., 2020). In the phylogeny, the B.1.160 variant of the present study clustered with France, Italy, and the United Kingdom. The third major lineage in our genomes B.1.255 has major global representation with significantly higher representation in the United States. Notably, one genome was observed from B.1.184 lineage, which has 100% representation in India (Andrew & Áine, 2020). These findings were validated in the hierarchical clustering where three clusters of genomes were observed (Figure S10). The founding lineages (B, B.1) appeared in the centre of the principal component analysis (PCA) plot, whereas the three sub-lineages clustered at the peripheries of the plot. These findings are suggestive of multiple routes of transmission of the virus during the first wave in the country including India, Europe, Saudi Arabia, and the United Arab Emirates.

From the three high-frequency SARS-CoV-2 variants observed in the first wave, only B.1.36 was found at a high frequency in the second wave, representing a selection advantage for B.1.36 in the region. The clustering of B.1.36 genomes of the second wave in close proximity of the United Arab Emirates and India (Figure S5) in the phylogeny may also indicate recurrent transmissions of the virus from these territories. Pakistan has large to and fro movement of the people from India as religious pilgrimages, and with the United Arab Emirates which caters large numbers of South Asian expatriates. Leveraging the publicly available genome sequences of SARS-CoV-2 reported from the same city (Karachi) and the second mostly populated city of Pakistan (Lahore, around 1030 km from Karachi), we determined concordance of SARS-CoV-2 transmission. The sequences from Karachi around June 2020 (GISAID accession numbers EPI_ISL_708839, EPI_ISL_708840, and EPI_ISL_709044) and October 2020 (EPI_ISL_709542) belonged to B.1.36 and B.1.160 lineages, which substantiate the findings of the present study. The sequences from Lahore around May 2020 (GISAID accession numbers EPI_ISL_548942, EPI_ISL_548943, EPI_ISL_548944, and EPI_ISL_548945) belonged to B.1.1.1 lineages. This analysis indicated that transmission of the virus within the two major cities was not identical and was through independent routes of transmission. Notably, the B.1.1.1 lineage is Europe specific with lot of representation (82.0%) from the United Kingdom (Rambaut et al., 2020).

Earlier this year (2021), the country witnessed the first UK variant (B.1.1.7) (Umair et al., 2021) in the capital Islamabad. Given the high transmissibility of the UK variant, this strain became the predominant virus in the country during the third wave, accounting for 69% of the genomes identified in the current study. However, the introduction of B.1.1.7 in the country was not primarily from the United Kingdom, many other transmission routes were also identified through the phylogeny. The B.1.1.7 strains of the present study were found in close proximity with India and European countries (France, Italy, and England) strains indicating the share of Indian pilgrimages in spreading the UK variant in the country. Notably, the B.1.36 strains managed their survival and contributed to infectivity in various cities. However, the mutations of B.1.36 strains have not been declared as variants of concern by the WHO.

Despite a marginal health care set-up in the country, Pakistan is a success story in controlling the pandemic. Having characterized the genetic diversity and lineages of the locally transmitted SARS-CoV-2 genomes, the data were shared with the National Command Operation Center (NCOC) of the country, which is the state-governed authority to manage and adopt appropriate measures to combat the pandemic. In addition to major lockdowns in the major cities during the peak of a COVID-19 wave, smart lockdowns were imposed in municipalities having COVID-19 cases of high transmissible strains. The employment of door-to-door community health workers enabled to closely monitor infected persons/families for restricting their stay inside their houses. These efforts lead to a decrease in local transmission of the COVID-19 viral strains in the country. The director general of the WHO in the opening remarks at a media briefing on COVID-19 (07 September 2020) applauded Pakistan's efforts in combating the pandemic within the country (World Health Organization, 2020).

5 | CONCLUSION

To the best of our knowledge, this study presents the first comprehensive report on the surveillance of genomic evolution in the SARS-CoV-2 transmitted locally in Pakistan during the different waves of the COVID-19 pandemic. The schematic analysis provided meaningful insight into the lineages being transmitted within the country. A continuous mutational tracking showed that none of the pandemic waves in the country was due to some locally evolved SARS-CoV-2 strain. Instead, the introduction of different variants of the virus through the travellers led to four episodes of infection in the country. The higher prevalence of specific types of B.1 sub-lineages highlighted that the viral strains of Europe, the United States, Middle East, and India had higher local transmission in the country. The findings of this study enabled the state authorities to monitor the spread of higher transmissible SARS-CoV-2 strains in the country and adopting timely and appropriate individual-level prophylactic measures. By now, about 10% of the population has been vaccinated. The identified lineages have not been reported to escape the vaccine protection. Further studies are underway to find more emerging strains within the country so that appropriate measures would be adopted by the respective state authorities.

ACKNOWLEDGEMENTS

This study was conducted through the institution strengthening support by The Searle Company Limited (TSCL), Pakistan, and the Government of Sindh, Pakistan to the Dr. Panjwani Center for Molecular Medicine and Drug Research (PCMD), ICCBS, University of Karachi, Pakistan. The funding agencies had no any role in the study design, data interpretation, or presentation of the results. We also acknowledge the National Institute of Virology, PCMD, ICCBS, University of Karachi for providing the biosafety laboratory grade 3 (BSL-3) facility.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

The study was approved by the Research Ethics Committee of the International Center for Chemical and Biological Sciences, University of Karachi, Karachi, Pakistan.

AUTHOR CONTRIBUTIONS

Experimental work, data analysis, data visualization, and writing of the manuscript: Muhammad Shakeel. *Experimental work, data analysis, data visualization, and writing of the manuscript:* Muhammad Irfan. *Experimental work, data analysis, and writing of the manuscript:* Zaib un Nisa. *Samples collection:* Saba Farooq. *Experimental work and data analysis:* Noor ul Ain. *Samples collection:* Waseem Iqbal. *Experimental work:* Niamatullah Kakar. *Samples collection and manuscript improvement:* Shah Jahan. *Experimental work, manuscript review and improvement:* Mohsin Shahzad. *Experimental work, manuscript review and improvement:* Saima Siddiqi. *PI of the study, conceptualization and overall supervision:* Ishtiaq Ahmad Khan.

DATA AVAILABILITY STATEMENT

The whole genome sequence data reported in this paper have been deposited and freely accessible in the Genome Warehouse in National Genomics Data Center (Genome Warehouse, 2021), Beijing Institute of Genomics (China National Center for Bioinformatics), Chinese Academy of Sciences, under BioProject accession number PRJCA004109 (<https://bigd.big.ac.cn/gwh>), and NCBI GenBank accession numbers MW447609-MW447645, MW960273-MW960294, MZ328025-MZ328045, and MZ676659-MZ976786.

ORCID

Ishtiaq Ahmad Khan  <https://orcid.org/0000-0003-2421-2625>

REFERENCES

- Aksamentov, I., Roemer, C., Hodcroft, E. B., & Neher, R. A. (2021). Nextclade: Clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, 6(67), 3773.
- Andrés, C., Piñana, M., Borrás-Bermejo, B., González-Sánchez, A., García-Cehic, D., Esperalba, J., Rando, A., Zules-Oña, R.-G., Campos, C., & Codina, M. G. (2022). A year living with SARS-CoV-2: An epidemiological overview of viral lineage circulation by whole-genome sequencing in Barcelona city (Catalonia, Spain). *Emerging Microbes & Infections*, 11(1), 172–181.
- Andrew, B., & Áine, O. T. (2020). SARS-CoV-2 lineages—Lineage B. https://cov-lineages.org/lineages/lineage_B.html
- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>
- Azgari, C., Kilinc, Z., Turhan, B., Circi, D., & Adebali, O. (2021). The mutation profile of SARS-CoV-2 is primarily shaped by the host antiviral defense. *Viruses*, 13(3), 394.
- CDC. (2020). Interim guidelines for collecting, handling, and testing clinical specimens for COVID-19. <https://www.cdc.gov/coronavirus/2019-ncov/lab/guidelines-clinical-specimens.html>
- Chong, Y. M., Sam, I.-C., Chong, J., Bador, M. K., Ponnampalavanar, S., Omar, S. F. S., Kamarulzaman, A., Munusamy, V., Wong, C. K., & Jamaluddin, F. H. (2020). SARS-CoV-2 lineage B. 6 was the major contributor to early pandemic transmission in Malaysia. *PLoS Neglected Tropical Diseases*, 14(11), e0008744.
- European Centre for Disease Prevention and Control. (2021). SARS-CoV-2 variants of concern as of 9 September 2021. <https://www.ecdc.europa.eu/en/covid-19/variants-concern>
- De Maio, N., Walker, C. R., Turakhia, Y., Lanfear, R., Corbett-Detig, R., & Goldman, N. (2021). Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biology and Evolution*, 13(5), evab087. <https://doi.org/10.1093/gbe/evab087>
- Deng, X., Gu, W., Federman, S., du Plessis, L., Pybus, O. G., Faria, N. R., Wang, C., Yu, G., Bushnell, B., Pan, C.-Y., Guevara, H., Sotomayor-Gonzalez, A., Zorn, K., Gopez, A., Servellita, V., Hsu, E., Miller, S., Bedford, T., Greninger, A. L., ... Chiu, C. Y. (2020). Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science*, 369(6503), 582–587. <https://doi.org/10.1126/science.abb9263>
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., & Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491.
- Dhandapan, P. S., Sadayappan, S., Xue, Y., Powell, G. T., Rani, D. S., Nallari, P., Rai, T. S., Khullar, M., Soares, P., & Bahl, A. (2009). A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. *Nature Genetics*, 41(2), 187–191.
- Du, P., Ding, N., Li, J., Zhang, F., Wang, Q., Chen, Z., Song, C., Han, K., Xie, W., Liu, J., Wang, L., Wei, L., Ma, S., Hua, M., Yu, F., Wang, L., Wang, W., An, K., Chen, J., ... Chen, C. (2020). Genomic surveillance of COVID-19 cases in Beijing. *Nature Communications*, 11(1), 5503. <https://doi.org/10.1038/s41467-020-19345-0>
- Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1), 113. <https://doi.org/10.1186/1471-2105-5-113>
- Genome Warehouse. (2021). Database resources of the national genomics data center, China National Center for Bioinformatics in 2021. *Nucleic Acids Research*, 49(D1), D18–D28.
- Greaney, A. J., Starr, T. N., Gilchuk, P., Zost, S. J., Binshtein, E., Loes, A. N., Hilton, S. K., Huddleston, J., Eguia, R., & Crawford, K. H. (2021). Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host & Microbe*, 29(1), 44–57. e49.
- Gregori, J., Perales, C., Rodríguez-Frias, F., Esteban, J. I., Quer, J., & Domingo, E. (2016). Viral quasispecies complexity measures. *Virology*, 493, 227–237.
- Helsinki, D. O. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310(20), 2191–2194. <https://doi.org/10.1001/jama.2013.281053>
- Ishtiaque, A., Mohammad Uzzal, H., Arittra, B., Zeshan Mahmud, C., Tabassum Hossain, E., Golam, M., Keshob Chandra, D., Chaman Ara, K., & Salimullah, M. (2020). Comparative genomic study for revealing the complete scenario of COVID-19 pandemic in Bangladesh. *medRxiv*, 16, e0258019. <https://doi.org/10.1101/2020.11.27.20240002>
- Joshi, M., Puvar, A., Kumar, D., Ansari, A., Pandya, M., Raval, J., Patel, Z., Trivdi, P., Gandhi, M., & Pandya, L. (2020). Genomic variations in SARS-CoV-2 genomes from Gujarat: Underlying role of variants in disease epidemiology. *bioRxiv*, .
- Kosuge, M., Furusawa-Nishii, E., Ito, K., Saito, Y., & Ogasawara, K. (2020). Point mutation bias in SARS-CoV-2 variants results in increased ability to stimulate inflammatory responses. *Scientific Reports*, 10(1), 17766. <https://doi.org/10.1038/s41598-020-74843-x>
- Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, 35(6), 1547.
- Ladner, J. T., Grubaugh, N. D., Pybus, O. G., & Andersen, K. G. (2019). Precision epidemiology for infectious disease control. *Nature Medicine*, 25(2), 206–211.
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. <https://arxiv.org/abs/1303.3997>

- Martinez, I., Llinás, D., Romero, M., & Salazar, L. (2020). High mutation rate in SARS-CoV-2: Will it hit us the same way forever. *Journal of Infectious Diseases and Epidemiology*, 6, 176.
- Ntoumi, F., Mfoutou Mapanguy, C. C., Tomazatos, A., Pallerla, S. R., Linh, L. T. K., Casadei, N., Angelov, A., Sonnabend, M., Peter, S., Kreamsner, P. G., & Velavan, T. P. (2021). Genomic surveillance of SARS-CoV-2 in the Republic of Congo. *International Journal of Infectious Diseases*, 105, 735–738. <https://doi.org/10.1016/j.ijid.2021.03.036>
- Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., & Gallo, R. C. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine*, 18(1), 1–9.
- Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., Zhang, X., Muruato, A. E., Zou, J., Fontes-Garfias, C. R., Mirchandani, D., Scharton, D., Bilello, J. P., Ku, Z., An, Z., Kalveram, B., Freiberg, A. N., Menachery, V. D., Xie, X., ... Shi, P. Y. (2020). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, 592, 116–121. <https://doi.org/10.1038/s41586-020-2895-3>
- Prijbelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes de novo assembler. *Current Protocols in Bioinformatics*, 70(1), e102. <https://doi.org/10.1002/cpbi.102>
- Rahimi, A., Mirzazadeh, A., & Tavakolpour, S. (2020). Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. *Genomics*, 113, 1221–1232.
- 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. *Nature Microbiology*, 5(11), 1403–1407. <https://doi.org/10.1038/s41564-020-0770-5>
- Roy, C., Mandal, S. M., Mondal, S. K., Mukherjee, S., Mapder, T., Ghosh, W., & Chakraborty, R. (2020). Trends of mutation accumulation across global SARS-CoV-2 genomes: Implications for the evolution of the novel coronavirus. *Genomics*, 112(6), 5331–5342. <https://doi.org/10.1016/j.ygeno.2020.11.003>
- Sah, R., Rodriguez-Morales, A. J., Jha, R., Chu, D. K. W., Gu, H., Peiris, M., Bastola, A., Lal, B. K., Ojha, H. C., Rabaan, A. A., Zambrano, L. I., Costello, A., Morita, K., Pandey, B. D., & Poon, L. L. M. (2020). Complete genome sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. *Microbiology Resource Announcements*, 9(11), e00169–00120. <https://doi.org/10.1128/mra.00169-20>
- Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data—From vision to reality. *Euro Surveillance*, 22(13), 30494. <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494>
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Tordoff, D. M., Greninger, A. L., Roychoudhury, P., Shrestha, L., Xie, H., Jerome, K. R., Breit, N., Huang, M.-L., Famulare, M., & Herbeck, J. T. (2021). Phylogenetic estimates of SARS-CoV-2 introductions into Washington State. *The Lancet Regional Health—Americas*, 1, 100018. <https://doi.org/10.1016/j.lana.2021.100018>
- Umair, M., Ikram, A., Salman, M., Alam, M. M., Badar, N., Rehman, Z., Tamim, S., Khurshid, A., Ahad, A., & Ahmad, H. (2021). Importation of SARS-CoV-2 variant B. 1.1. 7 in Pakistan. *Journal of Medical Virology*, 93, 2623–2625.
- van Dorp, L., Richard, D., Tan, C. C. S., Shaw, L. P., Acman, M., & Balloux, F. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nature Communications*, 2020(11), 5986. <https://doi.org/10.1038/s41467-020-19818-2>
- World Health Organization. (2020). WHO Director-General's opening remarks at the media briefing on COVID-19 - 7 September 2020. <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-7-september-2020>
- Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols*, 10(10), 1556–1566.
- Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics*, 31(1), 11–15. <https://doi.org/10.1002/0471250953.bi1105s31>
- Zhang, Y., Zhang, T., Fang, Y., Liu, J., Ye, Q., & Ding, L. (2022). SARS-CoV-2 spike L452R mutation increases Omicron variant fusogenicity and infectivity as well as host glycolysis. *Signal Transduction and Targeted Therapy*, 7(1), 1–3.
- Zhou, W., & Wang, W. (2021). Fast-spreading SARS-CoV-2 variants: Challenges to and new design strategies of COVID-19 vaccines. *Signal Transduction and Targeted Therapy*, 6(1), 1–6.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Shakeel, M., Irfan, M., Nisa, Z. u., Farooq, S., Ain, N. u., Iqbal, W., Kakar, N., Jahan, S., Shahzad, M., Siddiqi, S., & Khan, I. A. (2022). Genome sequencing and analysis of genomic diversity in the locally transmitted SARS-CoV-2 in Pakistan. *Transboundary and Emerging Diseases*, 1–13. <https://doi.org/10.1111/tbed.14586>