Medical Science

Phylogenetic analysis of SARS-CoV-2 strains from different countries: Insights into genetic variation

Muhammad Sajid Mustafa[1], Farrakh Mehmood Alvi[1], Nabeela Jabeen[2], Nasir Ali[3], Ansa Ijaz[4], Sadia Arshad[2], Ramsha Aamir[2], Muhammad Abid Mustafa[2], Namra Rasheed[2], Muhammad Zahid Iqbal[4*]

To Cite:
Mustafa MS, Alvi FM, Jabeen N, Ali N, Ijaz A, Arshad S, Aamir R, Mustafa MA, Rasheed N, Iqbal MZ. Phylogenetic analysis of SARS-CoV-2 strains from different countries: Insights into genetic variation. *Medical Science* 2024; 28: e117ms3427
doi: https://doi.org/10.54905/disssi.v28i150.e117ms3427

Authors' Affiliation:
[1]Department of Paramedical Sciences, Faculty of Allied Health Sciences, Institute of Public Health, Lahore, Pakistan
[2]Department of Pharmacognosy, Faculty of Pharmaceutical Sciences, Lahore University of Biological and Applied Sciences, Lahore, Pakistan
[3]Department of Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Lahore University of Biological and Applied Sciences, Lahore, Pakistan
[4]Department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Lahore University of Biological and Applied Sciences, Lahore, Pakistan

[*]Corresponding Author
Department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Lahore University of Biological and Applied Sciences, Lahore, Pakistan
Email: drmmziqbal@gmail.com

Peer-Review History
Received: 01 June 2024
Reviewed & Revised: 05/June/2024 to 19/August/2024
Accepted: 22 August 2024
Published: 27 August 2024

Peer-review Method
External peer-review was done through double-blind method.

Medical Science
pISSN 2321–7359; eISSN 2321–7367

© The Author(s) 2024. Open Access. This article is licensed under a Creative Commons Attribution License 4.0 (CC BY 4.0)., which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

DISCOVERY SCIENTIFIC SOCIETY

ABSTRACT

Introduction: Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was found in Wuhan, China, in November 2019. The World Health Organization proclaimed the illness a global pandemic. *Aim and Objectives:* This study determined mutations over a period of time in SARS-CoV-2 variants from different countries and the phylogenetic relationship between them. *Material and Methods:* Genomic sequences from Specified Countries Sequenced in 2020, 2021, and 2022 were downloaded from GenBank and NCBI and analyzed using specialized bioinformatics software such as BLAST. Variations within specific regions were examined with the aid of the reference genome (NC_045512.2). A phylogenetic tree was generated using the software MEGA 11. *Results:* A total of 659 variations were found, of which 582 (88.31%) were mismatches, 65 (9.86%) were gaps (deletions), and 12 (1.82%) were insertions. Most variations were found in ORF-1ab, which is 266 (45.70%), and the most common nucleotide variation found was C>T, commonly found in Australian isolates. Phylogeny showed that the reference genome and the South Africa 2020 isolate were homologous. *Conclusion:* It was concluded that SARS-CoV-2 mutates over time, changing its genomic sequence and leading to changes in its structure. This changes its affectivity, resulting in changes in its pathogenicity.

Keywords: BLAST, MEGA, Phylogenetic tree, SARS-CoV-2, Genetic Variations

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19) belongs to a group of epidemics with human-to-human transmission that are caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) coronavirus (Hao et al., 2022). June Almeida and Tyrrell conducted electron microscopy on B814-infected organ culture fluids, and they discovered particles that looked similar to the contagious chicken bronchitis virus. The particles had a membrane coating, were medium in size (80–150 nm), pleomorphic, and covered in widely spaced surface projections in the shape of clubs. Tyrrell oversaw a team of virologists who were working with human strains and various animal viruses in the late 1960s, all of which had been shown to be morphologically identical as observed through electron microscopy. As a result of the surface projections' "Crown-Like Appearance", this new group of viruses was given the name CORONAVIRUS and eventually recognized as a new genus of viruses (Kahn and McIntosh, 2005).

The positive-sense single-stranded RNA genome of enveloped coronaviruses measures 26–32 kb in length (Yang and Rao, 2021). The coronavirus virion particle usually has a diameter of 120–160 nm and a triple Spike (S) protein protrusion in the shape of a petal (King et al., 2011). The coronavirus genomes contain three structural proteins in addition to the distinctive S protein, including the Membrane (M) protein, the Envelope (E), and the Nucleocapsid (N) protein (Shi et al., 2020). Current research indicates that SARS-CoV-2 is spread from person to person when infectious particles are transmitted from an infected person's respiratory tract and enter the respiratory tract of a vulnerable person (Leung, 2021). The three main routes for SARS-CoV-2 transmission are (i) Airborne transmission, (ii) Direct contact, and (iii) indirect contact.

SARS-CoV-2 has a high rate of human-to-human transmission through intimate contact with infected individuals when the contagious virus is released during talking, breathing, coughing, or sneezing by an infected person (Da-Silva et al., 2022). The mean period from the commencement of the disease to the onset of symptoms for COVID-19 is around 5 days, and pneumonia often develops within a median time of 8 days (Li et al., 2020). Initial clinical manifestations of SARS-CoV-2 infection vary and resemble those of other respiratory viruses, such as influenza and parainfluenza viruses. Dry coughing, fever, and fatigue are the most typical signs of SARS-CoV-2 infection. Less frequent signs and symptoms include headache, sore throat, myalgia or arthralgia, shortness of breath, vomiting, dyspnea, chills, ageusia, dysgeusia, as well as changes in hyposmia, anosmia (Elmas et al., 2020).

As of 11 December 2021, five SARS-CoV-2 VOCs had been discovered since the start of the pandemic, according to the WHO's epidemiological update: Alpha (B.1.1.7): First variant of concern described in the United Kingdom (UK) in late December 2020, Beta (B.1.351): First reported in South Africa in December 2020, Gamma (P.1): First reported in Brazil in early January 2021, Delta (B.1.617.2): First reported in India in December 2020, Omicron (B.1.1.529): First reported in South Africa in November 2021 (Aleem et al., 2021). This study aims to find genetic mutations in different SARS-CoV-2 variants sequenced in specific regions of the world and to find a phylogenetic relationship between the SARS-CoV-2 strain genome and the reference genome. The objective is to find different COVID-19 strain genome sequences from selected countries, namely Pakistan, India, Iraq, South Africa, Australia, the United Kingdom, and Brazil. Then, genomic analysis will be done to determine SARS-CoV-2 genetic variation.

2. MATERIALS AND METHODOLOGY

The research was conducted over ten months, from 15 March 2023 to 16 December 2023. It received ethical approval from the Research Ethics Committee at the Institute of Public Health in Lahore, Pakistan, with reference number 79/ERC/IPH, confirming compliance with ethical standards and protocols. The study design was a Retrospective Observational Study, and this study includes Genomic sequences from 2020, 2021, and 2022 from specified regions (Pakistan, India, Iraq, Australia, South Africa, United Kingdom, and Brazil). Sequences with different nucleotide base pairs and the most minor percentage identity are included. Collected sequences with similar nucleotide base pairs and sequences with similarity (maximum percentage identity) are excluded. Sequences from years other than 2020, 2021, and 2022 and from other than specified regions are not collected.

Data Collection

Genomic Sequences Collection

Genomic sequences from Pakistan, India, Iraq, Australia, South Africa, the United Kingdom, and Brazil Sequenced in 2020, 2021, and 2022 were downloaded from GenBank, NCBI (ncbi.nlm.nih.gov) (Khailany et al., 2020).

Data Analysis

Genomic Analysis

Genomic Data collected from GenBank was analyzed by using specialized bioinformatics software such as BLAST (Almubaid and Al-Mubaid, 2021). Genes within the whole genome were considered to align the reference genome with the sequence of isolated strains. ORF-1ab, the S gene, the N gene, and the E gene were selected for genomic analysis. By using BLAST, the reference genome isolates sequences of genes in specific regions aligned to find genetic variations.

Genetic Variations

With the aid of the reference genome (NC_045512.2), variations were examined within specific regions using the genomic data that has been collected (Almubaid and Al-Mubaid, 2021). Genetic variation was examined in specific genes as the reference genome and Isolates were aligned as shown in Figure 1, and substitution, insertion and deletion mutations were extracted.

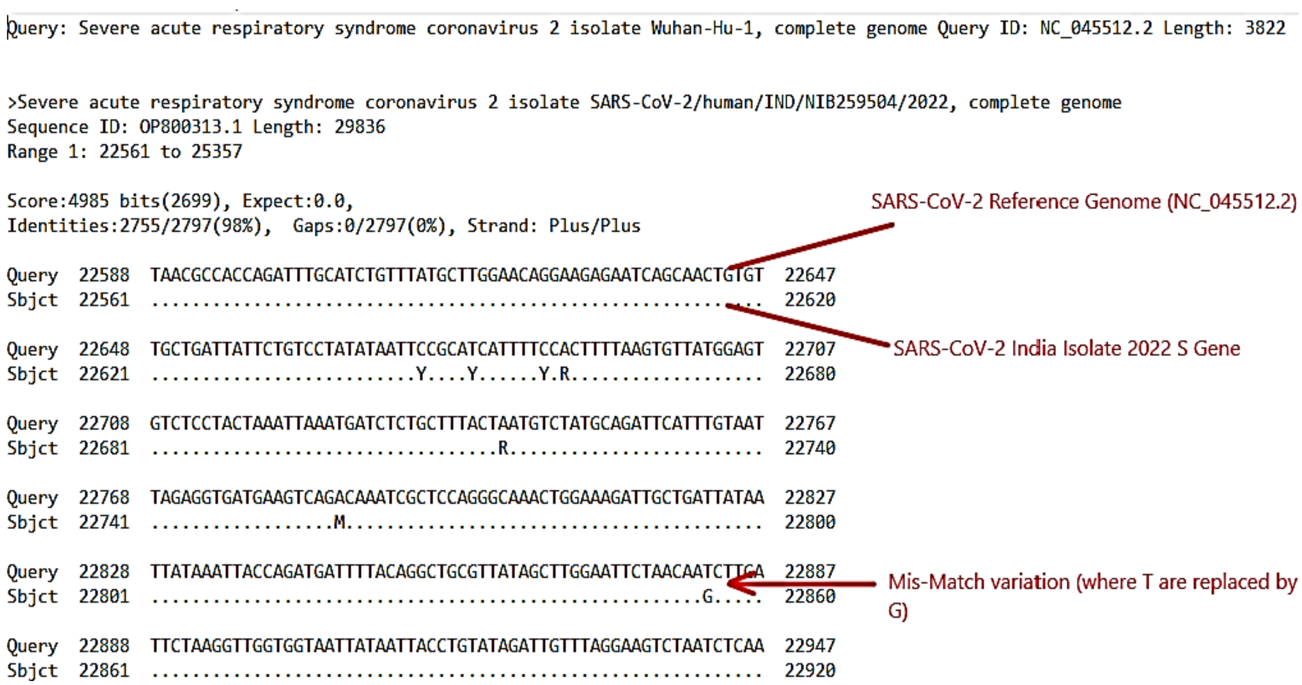

Figure 1 SARS-CoV-2 reference genome aligned with SARS-CoV-2 isolate from which different mutations are observed

Phylogenetic Analysis

A phylogenetic tree was generated by using the software MEGA 11 to analyze the relationship between isolates that have been collected (Stefanelli et al., 2020). A phylogenetic tree was generated of SARS-CoV-2 isolates, including outgroup Middle East Respiratory Syndrome (MERS), on the software MEGA 11. The maximum likelihood method is used with Bootstrap 1000 and the Kimura 2 parameter model to generate a phylogenetic tree, as shown in (Figure 1).

3. RESULTS

Sequences from specified countries are downloaded randomly from the NCBI (GenBank). 725 sequences were downloaded, of which 68 were from Pakistan, 122 from India, 133 from Iraq, 153 from Australia, and 249 from South Africa in the years 2020, 2021, and 2022. These sequences were run on Clustal W multiple sequence alignment at the European Bioinformatics Institute (https://www.ebi.ac.uk/). Sequences were aligned from each country of the year. A percentage identity matrix was generated as shown in (Table 1).

Table 1 Percentage identity matrix generated on Clustal W

				PAK 2022					
	1: OM327536.1	2: OM327535.1	3: OM327526.1	4: OM327532.1	5: OM327527.1	6: OM327533.1	7: OM327531.1	8: OM327534.1	9: OM327528.1
1: OM327536.1	100	99.97	99.99	99.99	99.99	99.98	99.98	99.99	99.99
2: OM327535.1	99.97	100	99.98	99.98	99.98	99.98	99.98	99.99	99.99
3: OM327526.1	99.99	99.98	100	99.99	99.99	99.99	99.99	99.99	99.99
4: OM327532.1	99.99	99.98	99.99	100	99.99	99.99	99.99	99.99	99.99
5: OM327527.1	99.99	99.98	99.99	99.99	100	99.99	99.99	99.99	99.99
6: OM327533.1	99.98	99.98	99.99	99.99	99.99	100	99.99	99.99	99.99
7: OM327531.1	99.98	99.98	99.99	99.99	99.99	99.99	100	100	100
8: OM327534.1	99.99	99.99	99.99	99.99	99.99	99.99	100	100	100
9: OM327528.1	99.99	99.99	99.99	99.99	99.99	99.99	100	100	100

Through a percentage identity matrix, sequences were analyzed to determine how much they were similar, and sequences with the most minor similarity were selected from each country in a year, as shown in (Table 2). 17 representative sequences were considered for further genomic variation analysis, of which 15 were from Pakistan, India, Iraq, South Africa and Australia (3 from each country each year), while 2 sequences were from the United Kingdom and Brazil in 2022. Table 3 represents a number of country-specific sequences together with the accession number and year.

Table 2 Sequences Selection

No. of Sequence Downloaded Randomly	**Included**	**Excluded**			**Reason of Exclusion/Inclusion**
			Similar	Dissimilar	
727	17	710	257	453	• Sequences having similarities between themselves. • The Most dissimilar sequences are selected for analysis.

Table 3 Number of sequences from countries with year and accession number

Sr. No.	**Country & Year**	**Accession No.**
1.	Pakistan 2020	MW433723.1
2.	Pakistan 2021	MZ328030.1
3.	Pakistan 2022	OM327535.1
4.	India 2020	MZ336023.1
5.	India 2021	OP599863.1
6.	India 2022	OP800313.1
7.	Iraq 2020	MT940495.1
8.	Iraq 2021	MZ366448.1
9.	Iraq 2022	OQ332738.1
10.	Australia 2020	MW321328.1
11.	Australia 2021	OL869974.1
12.	Australia 2022	OP604177.2
13.	South Africa 2020	OM725609.1
14.	South Africa 2021	OM721081.1

15.	South Africa 2022	OM857899.1
16.	Brazil 2022	ON972833.1
17.	United Kingdom 2022	OP680473.1

Genomic Variation in SARS-CoV-2 Isolates in different genes

A total of 659 variations were found during the analysis, of which 582 (88.31%) were mismatches (substitutions), 65 (9.86%) were gaps (deletions), and 12 (1.82%) were insertions. A total of 582 mismatches (substitutions) consist of 419 (71.99%) synonymous mutations and 163 (28%) non-synonymous (unique) mutations. These variations were found in specific genes (ORF-1ab, N gene, S gene, and E gene) as the SARS-CoV-2 reference genome (NC_045512.2) aligned with SARS-CoV-2 isolated from different countries by using BLAST. Most of the variations found were synonymous, but some of the variations were unique. This study shows that over time, a virus mutates, and changes in its genomic sequence and variations are produced, as shown in (Table 4). These variations lead to changes in its structure and its infectivity.

Table 4 SARS-CoV-2, over a period of time, mutates, and its sequence changes.

Country	Year	Genes	Nucleotide Variations
Iraq	2020	Orf1ab	10440 C>T, 10834 C>T, 14408 C>T, 1058 C>T, 1301 C>T, 3036 C>T
		N gene	28887 C>T
		E gene	No
		S gene	23127 C>T, 23403 A>G
Iraq	2021	Orf1ab	913 C>T, 3037 C>T, 3267 C>T, 5388 C>A, 5986 C>T, 6954 T>C, 14408 C>T, 14676 C>T, 15096 T>C, 15279 C>T, 15601 C>T, 16176 T>C,19079 A>G, 20465 A>G
		N gene	28881 G>A, 28882 G>A, 28883 G>C
		E gene	No
		S gene	21765—21770 (DEL),21991—21993 (DEL) 23271 C>A, 23403 A>G. 23604 C>A, 23709 C>T 24506 T>G, 24914 G>C
Iraq	2022	Orf1ab	583 C>T, 670 T>G, 2790 C>T, 3037 C>T, 3796 C>T, 3927 C>T, 4184 G>A, 4321 C>T, 4586 C>T, 5183 C>T, 7768 C>T, 9344 C>T, 9424 A>G, 9534 C>T, 9866 C>T, 10029 C>T, 10198 C>T, 10447 G>A, 10448 C>A, 10480 T>C, 12444 A>G, 12727 T>C, 12880 C>T, 13335 C>T, 14408 C>T, 15451 G>A, 15714 C>T, 17410 C>T, 18163 A>G, 19955 C>T, 20055 A>G, 20235 C>T
		N gene	28881 G>A ,82 G>A, 28883 G>C, 29510 A>C
		E gene	No
		S gene	21987 G>A, 22001 A>G, 22016 T>C, 22033 C>A, 22190 A>G, 22200 T>G, 22331 G>A, 22577 G>C, 22578 G>A, 22599 G>C, 22629 A>C, 22674 C>T, 22679 T>C, 22686 C>T, 22688 A>G, 22775 G>A, 22786 A>C, 22813 G>T, 22882 T>G, 22898 G>A, 22942 T>G, 22992 G>A, 22995 C>A, 23013 A>C, 23031 T>C, 23055 A>G, 23063 A>T, 23075 T>C, 23403 A>G, 23525 C>T 23599 T>G, 23604 C>A, 23854 C>A, 23948 G>T, 24424 A>T, 24469 T>A, 24764 G>A, 25000 C>T

According to Table 4, in the SARS-CoV-2 isolate from Iraq in 2020, we found only 9 variations; in 2021, only 32 variations; and in 2022, 74 variations. This shows that viruses mutate over time, which leads to changes in their structure and furthers their infectivity and affectivity.

Nucleotide Variations found in different countries & years in different genes

Aligning the SARS-CoV-2 reference genome and isolates from different countries and years revealed different nucleotide variations. Table 5 summarises the variations in nucleotides.

Adenine to Thymine Nucleotide Variations

A>T nucleotide variation in 7 different countries in specific genes at different locations was found 17 times as sequences aligned with the reference sequence. It shows 14 synonymous variations in the S gene at 23063 and 24424 and 3 unique variations in the orf-1ab and the S gene at 21405, 1163, and 22310.

Adenine to Guanine Nucleotide Variations

A>G nucleotide variation in 7 different countries in specific genes at different locations was found 73 times as sequences aligned with the reference sequence. It shows 59 synonymous variations at 2832, 9424, 11537, 18163, 20055, 28461, 23403, 23040, 23055, 22688 and 14 unique variations at 22190, 22001, 29301, 28370, 28881, in orf-1ab, S gene and N gene.

Adenine to Cytosine Nucleotide Variations

A>C nucleotide variation in 7 different countries in specific genes at different locations was found 20 times as sequences aligned with the reference sequence. It shows 18 synonymous variations at 29510, 22198, 23013, and 22786 and 2 unique variations at 23014 and 22629 in the N gene and S gene.

Cytosine to Adenine Nucleotide Variations

C>A nucleotide variation in 7 different countries in specific genes at different locations was found 41 times as sequences aligned with the reference sequence. It shows 37 synonymous variations at 10449, 24130, 23202, 23604, 22995, and 23854 and 4 unique variations at 5700, 5388, 23271, and 22033 in the orf-1ab, S gene.

Cytosine to Guanine Nucleotide Variations

C>G nucleotide variation in 7 different countries in specific genes at different locations was found 5 times as sequences aligned with the reference sequence. It is the least common nucleotide variation seen in our study. It shows 2 synonymous variations at 23604 and 3 unique variations at 9448, 21530, and 21618 in the orf-1ab, S gene.

Cytosine to Thymine Nucleotide Variations

C>T nucleotide variation in 7 different countries in specific genes at various locations was found 183 times as sequences aligned with the reference sequence. It is the most common nucleotide variation seen in our study. It shows 132 synonymous variations at 14408, 18877, 3037, 16466, 10198, 9866, 12880, 9344, 9534, 15714, 2790, 4321, 15240, 17410, 28311, 10029, 21846, 22686, 19955, 23525, 22674, 21618, 25000, 24503 and 51 unique variations at 12534, 13724, 5312, 16915, 1190, 15601, 13335, 20235, 2091, 6402, 13680, 4543, 1301, 5183, 313, 10834, 3267, 14230, 3927, 10440, 11514, 18555, 1058, 8991, 12025, 913, 13019, 5986, 18744, 583, 18742, , 3796, 7768, 20719, 9891, 3036, 15279, 6538, 12076, 28887, 14676, 4586, 28854, 16575, 2508, 22444, 23191, 23127, 23709, 22227, 26270 in orf-1ab, N gene, S gene and E gene.

Guanine to Adenine Nucleotide Variations

G>A nucleotide variation in 7 different countries in specific genes at various locations was found 74 times as sequences aligned with the reference sequence. It shows 70 synonymous variations at 15451, 11291, 10447, 8393, 4184, 23048, 22775, 28882, 28881, 22992, 22575, 24410, 22898, 22578, 21987, and 4 unique variations at 23401, 22331, 11873, 24764 in orf-1ab, N gene and S gene.

Guanine to Cytosine Nucleotide Variations

G>C nucleotide variation in 7 different countries in specific genes at various locations was found 17 times as sequences aligned with the reference sequence. It shows 11 synonymous variations at 28883 and 6 unique variations at 2164, 10523, 28423, 22599, 24914, and 22577 in the orf-1ab, N gene, and S gene.

Guanine to Thymine Nucleotide Variations

G>T nucleotide variation in 7 different countries in specific genes at various locations was found 30 times as sequences aligned with the reference sequence. It shows 15 synonymous variations at 28881, 22813, 23948 and 15 unique variations at 19677, 4907, 16647, 3692, 4300, 1145, 3114, 4180, 29468, 15760, 17562, 29212, 29402, 28899, 29314 in orf-1ab, N gene, S gene and E gene.

Thymine to Adenine Nucleotide Variations

T>A nucleotide variation in 7 different countries in specific genes at various locations was found 10 times as sequences aligned with the reference sequence. It shows 10 synonymous variations at 22204 and 24469 and no unique variations in the S gene.

Thymine to Guanine Nucleotide Variations

T>G nucleotide variation in 7 different countries in specific genes at various locations was found 34 times as sequences aligned with the reference sequence. It shows 31 synonymous variations at 5386, 670, 22197, 23599, 22882, 22917, 22195, and 22200 and 3 unique variations at 24506, 11296, and 22942 in orf-1ab and S gene.

Thymine to Cytosine Nucleotide Variations

T>C nucleotide variation in 7 different countries in specific genes at various locations was found 34 times as sequences aligned with the reference sequence. It shows 18 synonymous variations at 13195, 23075, 22679, 22673 and 16 unique variations at 3898, 15096, 16176, 8365, 12727, 23031, 2346, 11418, 6954, 10480, 7540, 9823, 10135, 24202, 22016, 21633, in orf-1ab and S gene.

Ambiguous Nucleotide Variations

In 7 different countries in specific genes to varying locations as sequences aligned with the reference sequence: A>R nucleotide variation was found 7 times at 11201, 11332, 20262, 15034, 18068, 22742, and 22788 in the orf-1ab and S gene. C>Y nucleotide variation was found 18 times; it shows 1 synonymous variation at 22786 and 17 unique variations 1191, 1267, 5188, 6539, 7124, 8986, 9891, 10029, 20320, 1473, 2790, 4321, 7083, 13487, 24745, 22674, 25000 in orf-1ab and S gene. T>Y nucleotide variation was found 7 times at 11418, 12946, 20497, 21270, 15030, 22679, 22416 in orf-1ab and S gene.

A>W nucleotide variation was found 2 times at 24424 and 28363 in the N gene and S gene. G>K nucleotide variation was found 5 times at 4182, 9053, 20274, 28916, and 24558 in the orf-1ab, N gene, and S gene. T>K nucleotide variation was found 1 time at 22882 in S gene. C>M nucleotide variation was found 2 times at 17430, 23854 in orf-1ab and S gene. A>M nucleotide variation was found 2 times; It shows 1 synonymous variation at 22786 and 1 unique variation at 5782 in the orf-1ab and S gene. These nucleotide variations were different in different genes at various locations.

Number of deletions and insertions found in Isolates

75 deletions and 12 insertions were found in the S gene and ORF-1ab, of which 30 synonymous deletions were found in the S gene and ORF-1ab, as shown in (Table 6). 12 insertions found in the S gene in Pakistan 2022 and Brazil 2022 were synonymous, as shown in (Table 6.) These insertions and deletions change the length of the SARS-CoV-2 genome, which affects its affectivity.

Table 5 The summary table shows nucleotide variation

		No. of times variation	Genes	No.	%	Synonymous & Unique Variation
Nucleotide Variations	A>T	17	ORF-1ab	2	12 %	14 synonymous & 3 Unique
			S Gene	15	88 %	
	A>G	73	ORF-1ab	32	44 %	59 synonymous & 14 Unique
			N Gene	6	8 %	
			S Gene	35	48 %	
	A>C	20	N Gene	5	25%	18 synonymous & 2 Unique
			S Gene	15	75%	
	C>A	41	ORF-1ab	10	24%	37 synonymous & 4 Unique
			S Gene	31	76%	
	C>G	5 (Most Least Variation)	ORF-1ab	2	40%	2 synonymous & 3 Unique
			S Gene	3	60%	
	C>T	183 (Most Common Variation)	ORF-1ab	135	74%	132 synonymous & 51 Unique
			N Gene	6	3%	
			S Gene	41	22%	
			E Gene	1	1%	
	G>A	74	ORF-1ab	20	27%	70 synonymous & 4 Unique
			N Gene	20	27%	
			S Gene	34	46%	
Nucleotide Variations	G>T	30	ORF-1ab	10	33%	15 synonymous & 15 Unique
			N Gene	6	20%	
			S Gene	13	43%	
			E Gene	1	3%	
	T>A	10	S Gene	10	100%	10 synonymous
	T>G	34	ORF-1ab	9	26%	31 synonymous & 3 Unique
			S Gene	25	74%	
	T>C	34	ORF-1ab	15	44%	18 synonymous & 16 Unique
			S Gene	19	56%	
Ambiguous Nucleotide Variations	A>R	7	ORF-1ab S-Gene	5 2	71% 29%	7 Unique
	C>Y	18	ORF-1ab S-Gene	14 4	78% 22%	1 synonymous & 17 Unique
	T>Y	7	ORF-1ab S-Gene	5 2	71% 29%	7 Unique
	A>W	2	N Gene S Gene	1 1	50% 50%	2 Unique
	G>K	5	S Gene	1	100%	2 Unique
	C>M	2	ORF-1ab S-Gene	1 1	50% 50%	2 Unique
	A>M	2	ORF-1ab S-Gene	1 1	50% 50%	2 Unique

	G>C	17	ORF-1ab	2	12%	11 synonymous & 6 Unique							
			N Gene	12	71%								
			S Gene	3	18%								

Table 6 Number of Deletions and Insertion found in isolates

Deletion (GAP)		
Country & Year	**Gene**	**Location**
PAK 2021	S Gene	21765—21770
PAK 2022	S Gene	21762, 21764, 21766, 21767, 21769, 21770, 21987—21995
India 2021	S Gene	22029—22034
Iraq 2021	S Gene	21765—21770 21991—21993
AUS 2021	S Gene	22194-22196
AUS 2022	Orf-1ab	11288—11296
SA 2022	Orf-1ab	11288—11296
Brazil 2022	Orf-1ab S Gene	11279—11287 21987—21995
Insertion		
Country & Year	Gene	Location
PAK 2022	S Gene	22194>T, 22202>G, 22203>A, 22205>C, 22207>A, 22209>A
Brazil 2022	S Gene	22194>T, 22202>G, 22203 >A, 22205>C, 22207>A, 22209>A

Number of total variations found in different Genes

Most variations were found in ORF-1ab, which is 266 (45.70%), then in the S gene, there were 257 (44.15%), and in the N gene, 57 (9.79%), while in the E gene, just 2 (0.34%) variations were found. Table 7 presents the findings.

Table 7 Number of totals, unique and synonymous nucleotide variation in different genes

	Total	**Non-Synonymous (Unique)**	**Synonymous (Duplicate)**
ORF-1ab	266	112	154
S Gene	257	36	221
N Gene	57	13	44
E Gene	2	2	0
Final Total	582	163	419

Above the following, most non-synonymous (unique) variations, 112, were found in ORF-1ab, while the least non-synonymous variations were found in the E gene, which is 2.

Frequency and Number of nucleotide variations found in different countries and year

This study shows different nucleotide variation frequencies in other countries over the years. The most common variation is C>T, which is 183, while the least common variation is C>G, which is 5, as shown in (Figure 2).

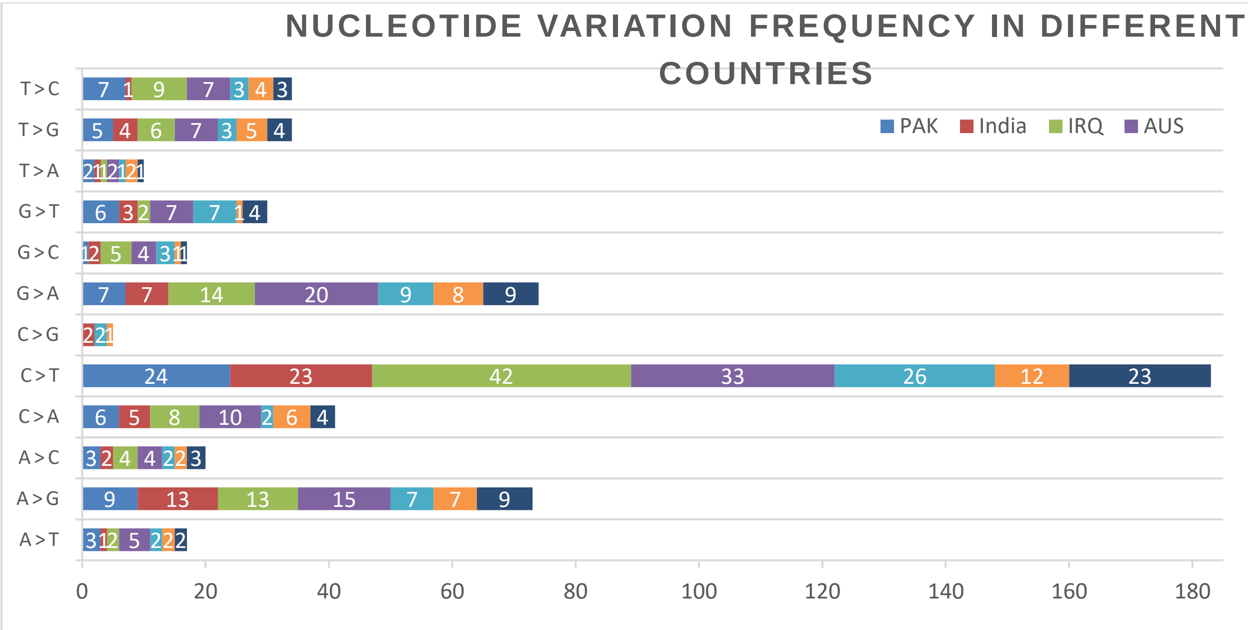

Figure 2 The frequency of different nucleotide variations found in different countries.

We found different numbers of nucleotide variations in different countries, which is demonstrated in the following (Figure 3). According to the above figure, most variation is found in Australia, and the least variation is found in Brazil, while in the remaining countries, variations are in between them.

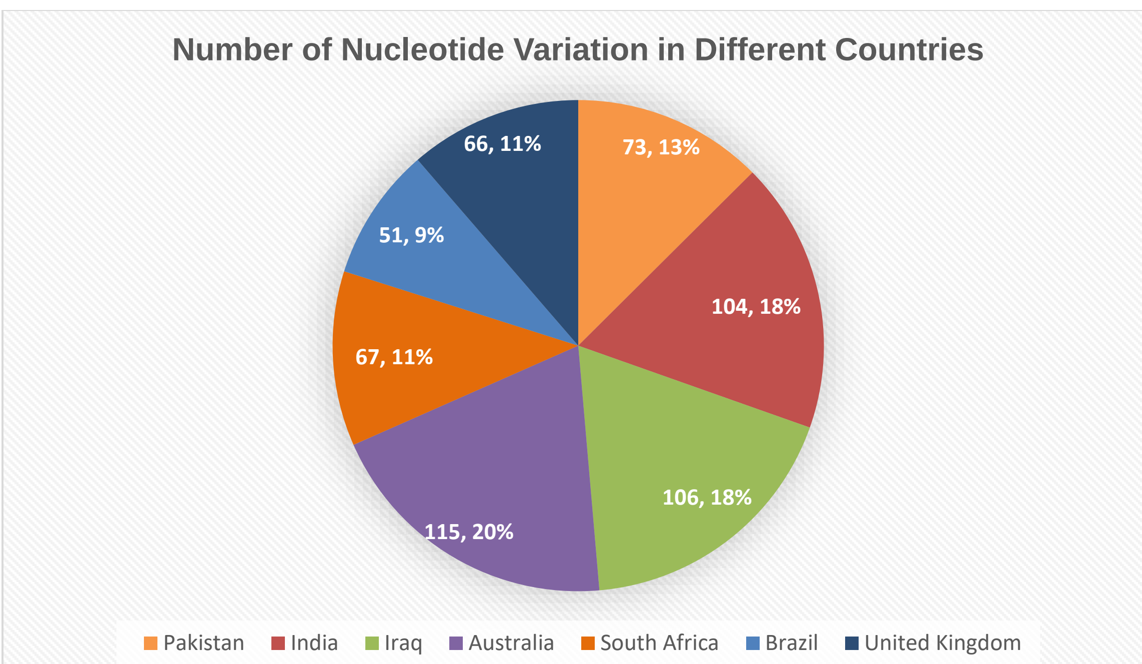

Figure 3 Number of different nucleotide variations in different countries

Phylogenetic Analysis

The phylogenetic tree from the sequence collected was generated using MEGA 11 with a maximum likelihood method having bootstrap 1000 and a Kimura 2 parameter model. In the Phylogenetic tree, one outgroup from Middle East Respiratory Syndrome from UAE (2016) was added, while the ingroup contained all the SARS-CoV-2 isolates sequence of interest that were collected. A phylogenetic tree, also known as a phylogeny, is a diagram that depicts the evolutionary descent of different species, organisms, or genes from a common ancestor as shown in (Figure 4).

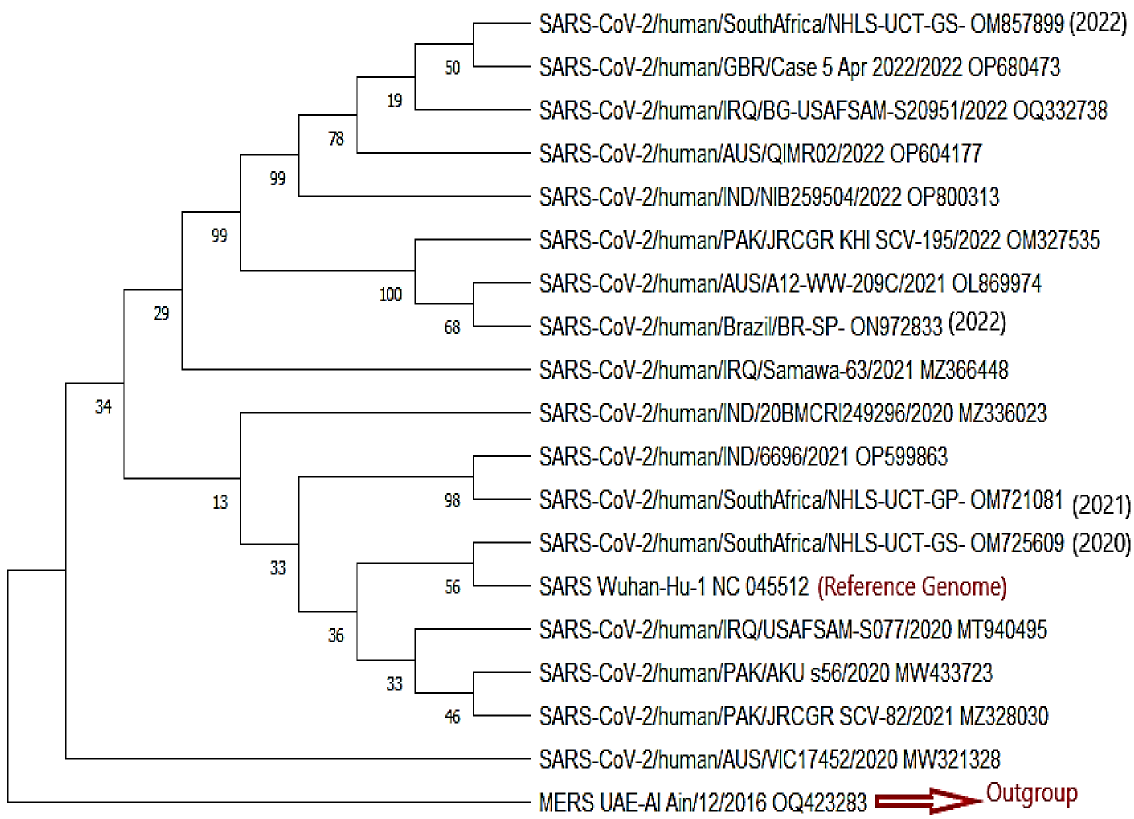

Figure 4 Phylogenetic tree of SARS-CoV-2 sequences collected from different countries

According to the phylogenetic tree, the sequences of SARS-CoV-2 collected from South Africa and the United Kingdom in 2022 were the most common recent ancestor and are most closely related to each other. The isolates from Iraq 2022, Australia 2022 and India 2022, as well as the South African 2022 and UK 2022, fall into the same clade, while the isolates from Pakistan 2022, Brazil 2022 and Australia 2021 fall into a different clade that is parallel to the above clade (sister clade), and both have evolved from a common ancestor. Isolate from Iraq 2021 has a common ancestor with the above-mentioned clades, and from this common ancestor, SARS-CoV-2 falls into two different clades, and from there, one further falls into two sister clades. Isolates from Pakistan in 2020 and 2021 are related to each other and fall in the same clade, and isolates from India and South Africa in 2021 are related and fall in the same clade.

Reference genome Wuhan-hu-1 (NC_045512.2) is most closely related to the South African isolate from 2020. These clades, along with Iraq in 2020, fall into one clade, and all of the isolates with India 2020 isolate and the remaining 2022 isolates evolved with Australian 2020 isolates from a common ancestor. So, from a common ancestor, 2020, 2021, and 2022 isolates evolved. This common ancestor of SARS-CoV-2 isolates evolved with MERS (outgroup) from a root ancestor. So, we found that the Australian 2020 isolate was the first to evolve from a common ancestor of SARS-CoV-2 isolates and is least relative to others, while the remaining isolates evolved over a period of time and fall into different clades and are related to each other.

4. DISCUSSION

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) was responsible for Coronavirus Disease 2019 (COVID-19) all over the world. Five strains of SARS-CoV-2 emerged from different regions of the world. The World Health Organization announced the COVID-19 pandemic in March 2019. After that, different SARS-CoV-2 strains were isolated from different areas of the world and sequenced and analyzed to create the structure of the virus. Different isolate sequences were submitted to Genbank NCBI and collected to find variations with the reference genome. In our study, we found different variations in which most of the variations were found in ORF-1ab, which is 266/582 (45.70%); similarly, most of the variations in ORF-1ab were reported by (Ahmed-Abakur and Alnour, 2020; Khailany et al., 2020). In genomic sequences, most of region 2/3 is covered by ORF-1ab (21290 nucleotides), and most mutations were reported in this region (Abduljalil and Abduljalil, 2020; Koyama et al., 2020).

Most nucleotide variations were found, which are C>T variations reported by (Khailany et al., 2020) which is also found in our study, which is 183/582 (31.44%). The ORF-1ab gene is involved in SARS-CoV-2 disease Khailany et al., (2020) is involved in virus replication, and is also involved in virus immune evasion (Abduljalil and Abduljalil, 2020; Ahmed-Abakur and Alnour, 2020). Variations found in the S gene are less than ORF-1ab, which is 257/582 (44.15%), as spike proteins play a key role in SARS-CoV-2 infection by virus invasion through cell membrane attachment Huang et al., (2020), and variations in the S gene can lead to changes in spike protein structure, which can affect virus pathogenicity Ahmed-Abakur and Alnour, (2020) and variation in spike protein is essential to understanding virus antigenicity and for vaccine development or therapeutic interventions (Alquraan et al., 2023; Lokman et al., 2020).

As reported, deletion in the N gene can likely impair the efficiency of primer annealing and can lead to impaired amplification, which results in the diagnostic escape of SARS-CoV-2, and deletion in the S gene (69–70) also affects the SARS-CoV-2 allplex PCR assay. The concurrence of these mutations can cause issues in the identification of a SARS-CoV-2 positive sample (Zannoli et al., 2022). In our study, there was no deletion in the N gene, while in the S gene, 48 deletions were found. Most of the deletions were found in the S gene. We found only two variations in the E gene in the UK 2022 and Brazil 2022 isolates. Gamma (P.1) first time reported in Brazil in 2021 and Alpha (B.1.1.7) first time reported in the United Kingdom in 2020. These isolates are of importance due to the variation in the E gene. As Special attention should be paid to the pathogenicity of the E gene variant in the future. The E gene is a small but important structural protein for coronaviruses.

It is involved in many viral-cycle processes. The E gene is conserved in coronaviruses (Sun et al., 2020). As reported by Nyagupe et al., (2023), the E gene of all nine lineages was the least mutated diagnostic gene, and in Belgium, a single nucleotide polymorphism (SNP) in the E gene caused a diagnostic dropout (Artesi et al., 2020). In phylogenetic analysis, we found homology between different isolates and reference genome, in which South African and UK 2022 are closely related to each other and South Africa 2020 and Wuhan Hu-1 Reference genome were more homologous. Australian 2020 isolate were different from the rest of the isolates and fell into different clades. As reported by (Stefanelli et al., 2020). The sequences of an Italian patient and a Chinese tourist were analyzed and compared to the sequence of the COVID-19 patient in Wuhan.

Phylogenetic analysis consistently placed the Italian patient's strain in a distinct cluster with other viral strains identified in Germany and Mexico, while the strain from the Chinese tourist, related to the Wuhan virus strain, clustered with different European strains and a strain from Australia. Some limitations in our study need to be addressed, such as the fact that it did not mention amino acid variations that lead to changes in specific protein structures and result in changes in viral structure and affectivity. Amino acid variation also specifies mutation types. The phylogenetic analysis did not mention the virus route in different countries and each isolate's evolutionary time from a common ancestor to the time of collection.

5. CONCLUSION

From our study, it is concluded that SARS-CoV-2 mutates over a period of time and changes its genomic sequence, leading to changes in its structure. Different nucleotide variations were found, most of which were C>T, and most variations were found in ORF-1ab. Variations found in Australian isolates are more significant than those in other isolates. The phylogeny shows homology between different isolates and their evolution from a common ancestor, from which the South African 2020 and reference genome fall into the same clade and are closely related, while the Australian 2020 isolates were much more different from the rest of the isolates.

As SARS-CoV-2 mutates and changes its structure, this results in changes in its effectivity, resulting in changes in its pathogenicity. It can be a source of difficult-to-treat disease, and in the future, it can be a deadly disease like COVID-19. So, studying variations and mutational analysis helps in the development of drugs and vaccines and prevents diseases like COVID-19 and others in the future.

Acknowledgement

We thank the participants who all contributed samples to the study. We want to extend our sincere gratitude to the Deanship of the Institute of Public Health Lahore for providing the necessary resources and support that made this research possible. Their commitment to fostering academic excellence has been invaluable.

Authors Contribution

Contribution	Muhammad Sajid Mustafa	Farrakh Mehmood Alvi	Nabeela Jabeen	Nasir Ali	Ansa Ijaz	Sadia Arshad	Ramsha Aamir	Muhammad Abid Mustafa	Namra Rasheed	Muhammad Zahid Iqbal
Conceived & designed	✓	✓	✓	✓	×	✓	✓	×	×	✓
Collected & analyzed data	✓	✓	✓	×	✓	✓	×	✓	✓	✓
Wrote manuscript	✓	✓	✓	✓	×	×	✓	✓	✓	✓
Read & approved manuscript	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Funding

This study has not received any external funding.

Conflict of interest

The authors declare that there is no conflict of interests.

Data and materials availability

All data sets collected during this study are available upon reasonable request from the corresponding author.

REFERENCES

1. Abduljalil JM, Abduljalil BM. Epidemiology, genome, and clinical features of the pandemic SARS-CoV-2: a recent view. New Microbes New Infect 2020; 35:100672. doi: 10.1016/j.nmni.2020.100672
2. Ahmed-Abakur EH, Alnour TMS. Genetic variations among SARS-CoV-2 strains isolated in China. Gene Rep 2020; 21:100925. doi: 10.1016/j.genrep.2020.100925
3. Aleem A, Akbar AB, Slenker AK. Emerging variants of SARS-CoV-2 and novel therapeutics against coronavirus (COVID-19). In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2021. doi: 10.1016/j.genrep.2020.100925
4. Almubaid Z, Al-Mubaid H. Analysis and comparison of genetic variants and mutations of the novel coronavirus SARS-CoV-2. Gene Rep 2021; 23:101064. doi: 10.1016/j.genrep.2021.101064
5. Alquraan L, Alzoubi KH, Rababa'h SY. Mutations of SARS-CoV-2 and their impact on disease diagnosis and severity. Inform Med Unlocked 2023; 39:101256. doi: 10.1016/j.imu.2023.101256
6. Artesi M, Bontems S, Göbbels P, Franckh M, Maes P, Boreux R, Meex C, Melin P, Hayette MP, Bours V, Durkin K. A Recurrent Mutation at Position 26340 of SARS-CoV-2 Is Associated with Failure of the E Gene Quantitative Reverse Transcription-PCR Utilized in a Commercial Dual-Target Diagnostic Assay. J Clin Microbiol 2020; 58(10):e01598-20. doi: 10.1128/JCM.01598-20
7. Da-Silva SJR, Do-Nascimento JCF, Germano-Mendes RP, Guarines KM, Targino-Alves-da-Silva C, Da-Silva PG, De-Magalhães JJF, Vigar JRJ, Silva-Júnior A, Kohl A, Pardee K, Pena L. Two Years into the COVID-19 Pandemic: Lessons

Learned. ACS Infect Dis 2022; 8(9):1758-1814. doi: 10.1021/acsinfecdis.2c00204
8. Elmas ÖF, Demirbaş A, Özyurt K, Atasoy M, Türsen Ü. Cutaneous manifestations of COVID-19: A review of the published literature. Dermatol Ther 2020; 33(4):e13696. doi: 10.1111/dth.13696
9. Hao YJ, Wang YL, Wang MY, Zhou L, Shi JY, Cao JM, Wang DP. The origins of COVID-19 pandemic: A brief overview. Transbound Emerg Dis 2022; 69(6):3181-3197. doi: 10.1111/tbed.14732
10. Huang Y, Yang C, Xu XF, Xu W, Liu SW. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin 2020; 41(9):1141-1149. doi: 10.1038/s41401-020-0485-4
11. Kahn JS, McIntosh K. History and recent advances in coronavirus discovery. Pediatr Infect Dis J 2005; 24(11 Suppl):S223-7, discussion S226. doi: 10.1097/01.inf.0000188166.17324.60
12. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene Rep 2020; 19:100682. doi: 10.1016/j.genrep.2020.100682
13. King AM, Lefkowitz E, Adams MJ, Carstens EB. Virus taxonomy: classification and nomenclature of viruses. Ninth report of the International Committee on Taxonomy of Viruses. Elsevier 2011.
14. Koyama T, Weeraratne D, Snowdon JL, Parida L. Emergence of Drift Variants That May Affect COVID-19 Vaccine Development and Antibody Treatment. Pathogens 2020; 9(5):324. doi: 10.3390/pathogens9050324
15. Leung NHL. Transmissibility and transmission of respiratory viruses. Nat Rev Microbiol 2021; 19(8):528-545. doi: 10.1038/s41579-021-00535-6
16. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 2020; 382(13):1199-1207. doi: 10.1056/NEJMoa2001316
17. Lokman SM, Rasheduzzaman M, Salauddin A, Barua R, Tanzina AY, Rumi MH, Hossain MI, Siddiki AMAMZ, Mannan A, Hasan MM. Exploring the genomic and proteomic variations of SARS-CoV-2 spike glycoprotein: A computational biology approach. Infect Genet Evol 2020; 84:104389. doi: 10.1016/j.meegid.2020.104389
18. Nyagupe C, De-Oliveira-Martins L, Gumbo H, Mashe T, Takawira T, Maeka KK, Juru A, Chikanda LK, Tauya AR, Page AJ, Kingsley RA, Simbi R, Chirenda J, Manasa J, Ruhanya V, Mavenyengwa RT. SARS-CoV-2 mutations on diagnostic gene targets in the second wave in Zimbabwe: A retrospective genomic analysis. S Afr Med J 2023; 113(3):141-147. doi: 10.7196/SAMJ.2023.v113i3.16762
19. Shi Y, Wang G, Cai XP, Deng JW, Zheng L, Zhu HH, Zheng M, Yang B, Chen Z. An overview of COVID-19. J Zhejiang Univ Sci B 2020; 21(5):343-360. doi: 10.1631/jzus.B2000083
20. Stefanelli P, Faggioni G, Lo-Presti A, Fiore S, Marchi A, Benedetti E, Fabiani C, Anselmo A, Ciammaruconi A, Fortunato A, De-Santis R, Fillo S, Capobianchi MR, Gismondo MR, Ciervo A, Rezza G, Castrucci MR, Lista F; ISS COVID-19 study group. Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. Euro Surveill 2020; 25(13):2000305. doi: 10.2807/1560-7917.ES.2020.25.13.2000305
21. Sun YS, Xu F, An Q, Chen C, Yang ZN, Lu HJ, Chen JC, Yao PP, Jiang JM, Zhu HP. A SARS-CoV-2 variant with the 12-bp deletion at E gene. Emerg Microbes Infect 2020; 9(1):2361-2367. doi: 10.1080/22221751.2020.1837017
22. Yang H, Rao Z. Structural biology of SARS-CoV-2 and implications for therapeutic development. Nat Rev Microbiol 2021; 19(11):685-700. doi: 10.1038/s41579-021-00630-8
23. Zannoli S, Dirani G, Taddei F, Gatti G, Poggianti I, Denicolò A, Arfilli V, Manera M, Mancini A, Battisti A, Sambri V. A deletion in the N gene may cause diagnostic escape in SARS-CoV-2 samples. Diagn Microbiol Infect Dis 2022; 102(1):115540. doi: 10.1016/j.diagmicrobio.2021.115540