KINGDOM OF SAUDI ARABIA
Northern Border University (NBU)
Journal of the North for Basic & Applied Sciences(JNBAS)
p - ISSN : 1658 -7022 / **e-ISSN:** 1658 - 7014
www.nbu.edu.sa
s.journal@nbu.edu.sa

SARS-CoV-2 Variants in the Eastern Region of Saudi Arabia: How Far from the Origin?

A. Khuzaim Alzahrani
Medical Laboratories Technology, Faculty of Applied Medical Sciences, Northern Border University, P.O. Box 1321, Arar 91431, Saudi Arabia.

(**Received:** : 10th February 2026; **Accepted:** 7th September 2026)

Abstract
Infectious viruses, including SARS-CoV-2, use different strategies, such as mutations, to alter their phenotype in ways that confer a fitness advantage. These mutations play a critical role in facilitating the virus's successful entry into host cells. In this study, we used 18 whole-genome sequences of SARS-CoV-2 variants circulating in the Eastern Region of Saudi Arabia and compared them to the Wuhan strain hCoV-19/Wuhan/WIV04/2019. The current study analyzed population expansion of SARS-CoV-2 transmission (Tajima's D = -1.061; P < 0.28 for the normal distribution, P < 0.311 for the beta distribution). In total, 18 haplotypes were detected, among which 17 were singletons. Changes in evolutionary relationships among the 18 Saudi SARS-CoV-2 variants indicate that the analyzed genotypes clustered into distinct groups relative to the hCoV-19/Wuhan/WIV04/2019 strain. Genomic variation may reflect the high rate of amino acid (aa) mutations, especially in the spike protein. Over 428 mutations were mainly localized in the spike, and 2 sequences (OV079467.1 and OV090001.1) showed the highest number of mutations (68 and 96, respectively). A total of 192 recorded mutations were unique to the studied genomes according to GISAID. The spike protein regulates SARS-CoV-2's ability to bind the human ACE2 receptor or to participate in other steps in the invasion of host cells. Thus, understanding the main drivers of the emergence, spread, and dominance of pathogens over human defenses is necessary, as these mutations may undermine all drug repurposing efforts.

Keywords: SARS-CoV-2, Eastern Region of Saudi Arabia, haplotype, amino acid (aa), mutations, spike protein

1658-7022© JNBAS. (1448 H/2026). Published by Northern Border University (NBU). All Rights Reserved.

DOI: 10.12816/0062683

(*) **Corresponding Author**:
A. Khuzaim Alzahrani
Medical Laboratories Technology, Faculty of Applied Medical Sciences, Northern Border University, P.O. Box 1321, Arar 91431, Saudi Arabia.

E-mail: akaalz@nbu.edu.sa

1. Introduction

In March 2020, the World Health Organization (WHO) declared the COVID-19 pandemic. The outbreak that began in Wuhan, Hubei Province, China, was caused by the emergence of a beta-coronavirus, SARS-CoV-2 (Huang et al., 2020; Wang et al., 2020; Zhu et al., 2020). By February 12, 2023, the viral infection had spread worldwide for over 2 years, resulting in over 677 million confirmed cases and 6.78 million deaths (https://www.who.int/). The Saudi Arabian government reported the first case of positive SARS-CoV-2 on 2 March 2020, and the first death was officially registered on 24 March 2020. Three years later, approximately 828 thousand confirmed cases and 9575 deaths were reported, placing Saudi Arabia as one of the most severely affected countries in the Middle East (https://covid19.moh.gov.sa/; https://covid19.who.int/data). During this period (March 2020 to the end of 2022), the official cumulative incidence rate was 1646 cases/100,000 inhabitants, while the mortality rate was 146 deaths/100,000 inhabitants, with a lethality of 8.9% (https://covid19.moh.gov.sa).

The contagiosity and pathogenic features, as well as the evolutionary trajectory of SARS-CoV-2, have been extensively studied, resulting in over 14 million sequences worldwide (https://gisaid.org/). Since the publication of the first viral genome reference (NCBI accession number NC-045512.2), the number of genome sequence publications has steadily increased. More than 2,000 variants/lineages are currently available in the GISAID database (http://www.gisaid.org), including recent reports of more contagious variants from the UK, South Africa, and Brazil.

Indeed, Saudi Arabia had already experienced a similar severe beta-coronavirus outbreak in 2012, when a pair of Saudi nationals were diagnosed with the Middle East respiratory syndrome coronavirus (MERS-CoV). The outbreak was closely monitored and dealt with in a professional way, which reduced mortality to 838 deaths among the reported cases (Rahman & Sarkar, 2019). Similar to SARS, MERS-CoV infection begins with minor upper respiratory symptoms and progresses to serious respiratory illness (Memish et al., 2013). The Eastern Region of Saudi Arabia was reported to be the first region hit by the pandemic (moh.gov.sa). The geographical distribution of cumulative COVID-19 cases in the country showed that the Eastern Region is among the most affected regions (Figure 1). For approximately 3 years, the region registered approximately 161744 confirmed cases and 1470 deaths (https://covid19.moh.gov.sa/).

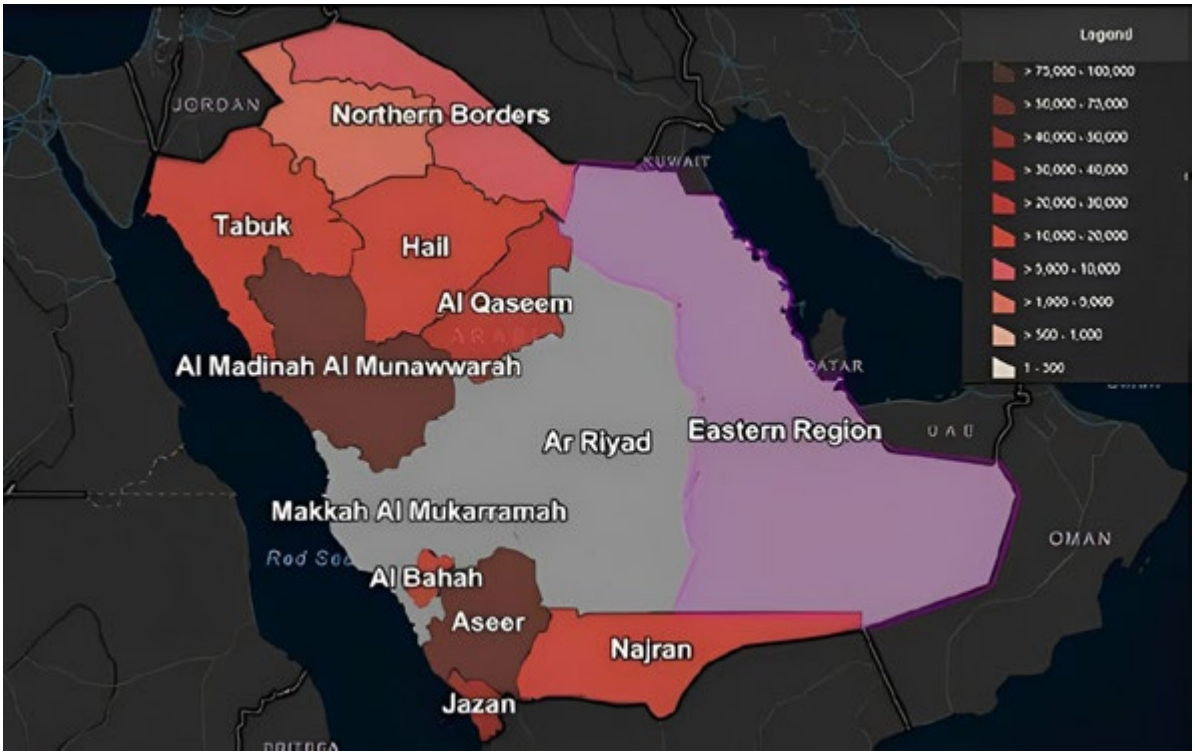


Figure 1. The thirteen regions of Saudi Arabia show the geographical distribution of cumulative COVID-19 cases. Riyadh region and Makkah Al Mukarramah region were the 2 regions with the highest cumulative confirmed cases. The eastern region, in pink, is the third region in the country with the highest cumulative number of affected cases in Saudi Arabia.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), an RNA coronavirus, is responsible for the COVID-19 outbreak (a, b). The virus has been identified as the cause of the disease and belongs to the beta-CoV genus (8,c). SARS-CoV-2 has a linear, single-stranded RNA genome (~30 kilobases, kb) encoding approximately 29 proteins (1,3,5,6,7,8). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) undergoes mutations as it replicates. These mutations have accumulated over time since the virus emerged, and new variants continue to appear as new mutations emerge.

In late 2020, after circulating for almost a year in the human population, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) exhibited a major step change in its adaptation to humans (9). The pathogenic potential of SARS-CoV-2 can be influenced by adaptive mutations in its genome. Mutations that can lead to amino acid modifications could affect the virus's ability to evade the immune system and complicate the recently developed vaccines (12). Like other RNA viruses, SARS-CoV-2 adapts to its new human hosts and evolves over time, developing mutations (13).

SARS-CoV-2 consists of four main structural proteins: (i) nucleocapsid protein (N), (ii) spike protein (S), (iii) membrane protein (M), and (iv) envelope protein (E). The E protein is the smallest, ranging from 8 to 12 kDa, and is involved in a wide functional repertoire (Schoeman & Fielding, 2019). The E protein includes three domains: (i) a short hydrophilic N-terminus domain, consisting of 7-12 amino acids; (ii) a hydrophobic transmembrane domain, which is approximately 25 amino acids long; and (iii) a long hydrophilic C-terminal region (Corse & Machamer, 2000; Surya et al., 2018). The S protein has two major domains: S1, which facilitates attachment and binding to the host cell receptor, and S2, which mediates fusion to the host cell membrane (Chu et al., 2020). Both domains play a critical role in establishing the successful entry of the virion into its host cell. SARS-CoV-2 is known to mutate rapidly, especially the spike protein, to escape the host's immune system and adapt to its environment (Saha et al., 2020). Unlike the S protein, the N gene, with 90% amino acid homology and fewer mutations over time, is more conserved and stable (Grifoni et al., 2020; Holmes & Enjuanes, 2003; Marra et al., 2003; Pyrc et al., 2007; Zhu et al., 2005). In many SARS-CoV-2 cases, the M and N proteins were reported to be the most abundant viral proteins expressed during infection and key proteins in the assembly of both naked and enveloped virus particles (Kuo & Masters, 2002). The M protein plays a critical role in intracellular virus particle assembly, for which S is not required (He et al., 2004).

Given that adaptive mutations could be naturally selected in broader populations, studying SARS-CoV-2 genomic variants and tracking them over time might help us understand viral evolution, behavior, and development. The present study aims to (1) characterize the genomic variations of SARS-CoV-2, (2) infer the evolutionary relationships of the Saudi strains, and (3) deduce the transmission history of Saudi SARS-CoV-2 from the Wuhan strain using genomic data of SARS-CoV-2 from the eastern region as "a starting point of infection in the country".

2. Methods

2.1 Data collection

The genomic sequences of SARS-CoV-2 patients were retrieved from the NCBI database as of January 2023. Eighteen complete whole-genome sequences were reported from 18 patients in the eastern region of Saudi Arabia, covering the period from the onset of the pandemic in 2020 through December 2022. Sequence selection criteria were as follows: (i) sequences must originate from human hosts; (ii) sequences must be complete whole-genome sequences with coverage above 99% (≥29,903 bp); (iii) sequences must have been collected from patients in the Eastern Region of Saudi Arabia; and (iv) sequences must have been deposited in the NCBI SARS-CoV-2 data hub by January 2023. Sequences were excluded if they were incomplete, had low coverage (<99%), or lacked metadata confirming their collection from the Eastern Region. As this study is entirely based on publicly available sequences, individual patient clinical or demographic metadata were not available; this is acknowledged as a limitation (see Limitations subsection). The sequence of Wuhan strain hCoV-19/Wuhan/WIV04/2019 was included as the reference sequence.

2.2 Bioinformatics

All 19 sequences were aligned against the reference Wuhan strain hCoV19/Wuhan/WIV04/2019 using the MUSCLE algorithm (Edgar, 2004) as implemented in MEGA X (Kumar et al., 2018). Ambiguous positions in each sequence pair were excluded using the pairwise deletion option. The most appropriate nucleotide substitution model was determined based on the lowest Bayesian Information Criterion (BIC) score, and the Tamura-Nei (TN93) model was selected accordingly.

The evolutionary history was inferred using the Maximum Likelihood method in MEGA X (Kumar et al., 2018). The reliability of the resulting phylogenetic tree was assessed using a bootstrap test with 1,000 replicates (Felsenstein, 1985). The final alignment comprised a total of 29,903 nucleotide positions (~29.9 kb), consistent with the full-length SARS-CoV-2 genome.

2.3 Extraction and visualization of haplotypes

The haplotypes and sequence variations were extracted from the aligned sequences as a matrix using the R base package (28) to compare sequences and to

determine haplotype numbers and frequencies. Unique haplotype sequences were calculated by ignoring common nucleotides between haplotypes. The Hamming distance method is a calculation of the pairwise distance matrix for the corresponding symbols in two equal-length strings (Pinheiro et al., 2005). Our dataset consisted of haplotype sequences, 18,368 base pairs long, used to assess phylogenetic relationships using the analysis of phylogenetics and evolution (APE) package in R (Paradiset al., 2004). The estimated differences between the two sequences are classified as strong for 85% and above, moderate for 70-85%, weak for 50-70%, and poor for 50% and below. Indeed, the evolutionary analyses were conducted in MEGA X (Kumar et al., 2018) using the same aligned sequences used to estimate transition/ transversion rate ratios in the current population.

2.4 Statistical Analyses

Tajima's D (25), a neutral mutation hypothesis test, was calculated as described in the population and evolutionary genetics analysis system "Pegas" (Paradis, 2010). Both nucleotide substitution and maximum-likelihood estimation of the gamma parameter for site rates were performed. The transition/transversion ratio (R) represents the ratio of the number of transitions to the number of transversions for the included sequences. The R value indicates the bias toward either transitional or transversional substitution.

Amino acid (aa) mutation analysis was conducted using GISAID CoVsurver (http://gisaid.org/database-features/covsurver-mutations-app/), a tool developed to aid the research community in the identification and interpretation of aa changes in coronavirus genomes. For the 16 SARS-CoV-2 sequences from Saudi Arabia that were successfully processed by GISAID CoVsurver (two sequences, OV090001.1 and OV079467.1, were excluded due to truncation of the spike protein exceeding the threshold accepted by the tool), mutations and their positions on the viral genome were presented along with their frequency change(s) in each structural model, compared with the reference genome of the Wuhan strain hCoV-19/Wuhan/WIV04/2019 (30).

3. Results

3.1 Phylogenetic analyses

In total, 18 sequences with a coverage higher than 99% (29.903 Kb) of the SARS-CoV-2 genome were obtained in this study. This work is an in silico study aimed at evaluating genetic variations in SARS-CoV-2 in the eastern region of Saudi Arabia collected between 2020 and December 2022. SARS-CoV-2 sequences were obtained from the NCBI SARS-CoV-2 data hub (Jan/2023). The sequence of Wuhan strain hCoV-19/Wuhan/WIV04/2019 was included as outgroup. Phylogenetic trees were constructed based on the estimation of the neighbor-joining method supported by the APE package (Paradis et al., 2004). Phylogenetic analysis of the 18 Saudi SARS-CoV-2 sequences, with the Wuhan reference sequence (hCoV-19/Wuhan/WIV04/2019) included as an outgroup, inferred using the Maximum Likelihood method with 1,000 bootstrap replicates, revealed two major clusters within the ingroup, supported by strong bootstrap values, along with several divergent sequences that did not group within either cluster (Figure 2). The 15 sequences forming the main ingroup clade were separated from the more basal isolates (OV090001.1, OV090750.1, and OV079467.1) with maximal bootstrap support (100%), and this clade subsequently diverged into two well-supported clusters (100% bootstrap support at the node separating them). The first cluster comprised seven sequences (OV090846.1, OV090864.1, OU998690.1, OV005890.1, OU993040.1, OV091501.1, and OU993652.1). Within this cluster, OU998690.1 and OV005890.1 formed a closely related pair, supported by a moderate bootstrap value of 72%. The second cluster comprised eight sequences (OV091476.1, OV090004.1, OV091486.1, OV091533.1, OV091491.1, OV091561.1, OV009572.1, and OU992750.1). Within this cluster, OV091561.1 and OV009572.1 formed a distinct pair, although this grouping received only weak bootstrap support (58%), suggesting limited confidence in their close evolutionary relationship relative to the rest of the cluster.

Three sequences OV090001.1, OV090750.1, and OV079467.1 occupied more basal positions in the tree and did not cluster with either of the two main groups, indicating greater genetic divergence from the remaining Saudi isolates. OV079467.1 branched closest to the root among all ingroup sequences. The 18 Saudi sequences were phylogenetically compared with the Wuhan reference strain hCoV-19/Wuhan/WIV04/2019 (NC-045512.2), which formed a clearly distinct outgroup, confirming the expected evolutionary divergence of the analyzed Saudi isolates from the ancestral Wuhan lineage.



Figure 2. Plotting the multiple sequence alignments with a phylogenetic tree. The phylogenetic tree of the 18 SARS-CoV-2 genomes and nucleotide differences has been reconstructed relative to Wuhan (NC-045512.2). The evolutionary history was inferred using the Maximum Likelihood method with the Tamura-Nei (TN93) substitution model as implemented in MEGA X. Bootstrap consensus tree inferred from 1,000 replicates is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1,000 replicates) are shown next to the branches. This analysis involved 18 Saudi SARS-CoV-2 sequences and one Wuhan reference sequence (NC-045512.2) as outgroup (19 nucleotide sequences in total). Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated (complete deletion option). There was a total of 29,661 positions in the final dataset. Evolutionary analyses were conducted in MEGA X.

3.2 Haplotype analysis

Using the R base package and the 18 aligned sequences of SARS-CoV-2 in the Eastern Region of Saudi Arabia as a matrix, we revealed 18 different haplotypes. The total length of the sequence of the generated haplotypes was 18,368 bp (61.43% of the aligned sequence and 61.23% of the SARS-CoV-2 genome). Seventeen of eighteen identified haplotypes were singletons (5.26%). Additionally, H11 (10.53%) was found in two sequences (OU998690.1 and OV091533.1). The distance matrix allowed analysis of differences in the number of nucleotides between haplotypes, showing the smallest difference of 1 nucleotide between 4 combinations of described haplotypes: H9-H11, H10-H11, H12-H15, and H15-H16. However, the largest difference was 35 nucleotides between haplotypes H5 and H14.

To understand the dynamics of these mutations, we constructed a frequency heatmap of the number of nucleotide differences between the 18 haplotypes. Each branch of the phylogenetic tree corresponds to the haplotype in the matrix; thus, it is fully compatible with the haplotype distance matrix. The results indicate a high rate of nucleotide differences between haplotypes.

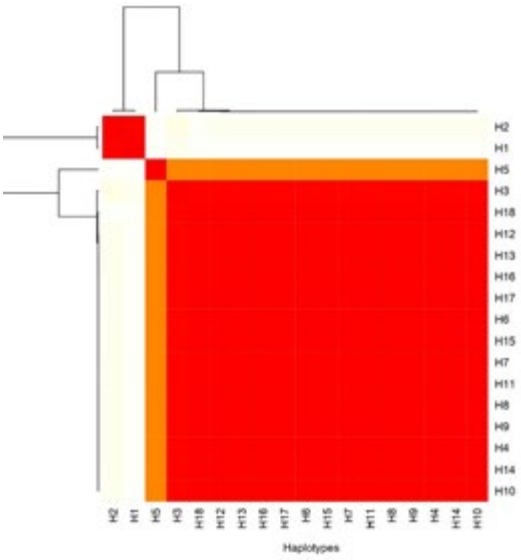


Figure 3. Heatmap of the number of nucleotide differences between haplotypes. Each branch of the phylogenetic tree corresponds to the haplotype in the matrix. Close relationships are defined by the "dark red" color, and distant relationships by the "light" color.

3.3 Maximum Likelihood Estimate of Transition/ Transversion Bias

The transition/transversion (Ti/Tv) ratio is a measure of the nucleotide mutation process and is a well-established aspect of substitution models (Kimura, 1980). The Ti/Tv ratio is calculated as the ratio of the transition of a single nucleotide polymorphism (changes from

one purine to another purine or from one pyrimidine to another pyrimidine) to the number of transversion SNPs (changes from a purine to a pyrimidine or vice versa).

The estimated transition/transversion bias parameter (R) in the Kimura 2-parameter model is 0.50, reflecting the ratio used in the substitution model parameterization. Note that the pairwise Ti/Tv values reported in Table 1 (ranging from 8.3285 to 8.3430) represent a separate maximum-likelihood estimation of nucleotide substitution rates between specific base pairs, and these two values are not directly comparable. Substitution patterns and rates were estimated under the Kimura (1980) 2-parameter model (+G). A discrete gamma distribution of SARS-CoV-2 genomes was used to model evolutionary rate differences among sites (5 categories, [+G], parameter = 5.2554). To estimate ML values, tree topology was automatically computed (Table 1). The maximum log-likelihood for this computation was -104971.438. This analysis involved 19 nucleotide sequences. The codon positions included were 1st+2nd+3rd+Noncoding. All positions with less than 95% site coverage were eliminated, i.e., fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position (partial deletion option). The current analysis included a total of 29677 positions in the final dataset. The estimated (Ti/Tv) bias analysis showed, relative to the reference genome, 51% as transition mutations and 49% as transversion mutations. As shown in Table 2, after examining all the variations, we noticed that the Ti/Tv ratio was high (the ratio varied between 8.3285 and 8.3430).

Table.1: The estimated Transition/Transversion bias.

From\To	A	T	C	G
A	-			
T	8.3285	-		
C	8.3285	8.3430	-	
G	8.3430	8.3285	8.3285	-

3.4 Genomic variation and population expansion

Phylogenetic analysis, haplotype networking, and genetic diversity analysis were performed on 18 genomic sequences from 18 patients confirmed to be SARS-CoV-2-positive in the eastern region of Saudi Arabia. These analyses revealed more than 428 mutations, mainly localized in the spike protein. Two sequences (OV079467.1 and OV090001.1) showed the highest number of mutations, with 68 and 96 modifications in the amino acid sequences, respectively. A total of 192 of the recorded 428 mutations were reported for the first time and have never been reported before. The phylogenetic analysis showed that 2 clusters can be recognized in Eastern Region Saudi Arabian SARS-CoV-2 genome sequences, with the first cluster (Cluster A) comprising the majority (84.21%) of the genotypes and represented by 16 of the 18 Saudi sequences included in the present study.

The second cluster included only 3 genotypes and was closely related to the Wuhan sequence (NC-045512.2). The 2 sequences (OV079467.1 and OV090001.1) appear closely related to the Wuhan sequence, NC-045512.2, with a high degree of genetic similarity. The study of these sequences can provide valuable insights into the virus's spread and evolution in the eastern region of Saudi Arabia and its potential impact on public health. Understanding the genetic differences between these sequences and the Wuhan sequence can also help inform the development of effective treatments and vaccines. Continuous surveillance and monitoring of these and other related SARS-CoV-2 sequences are crucial to our efforts to control the virus's spread and mitigate its impact on global health.

Similarly, the haplotype network analysis showed the same two clusters, comprising a total of 18 haplotypes. The mismatch distribution analysis of 18 SARS-CoV-2 genomes exhibited nonsignificant population expansion of SARS-CoV-2 transmission (Tajima's D= -1.061; P<0.28 for the normal distribution, P<0.311 beta distribution). Although no significant changes were observed, the nucleotide transition/transversion ratio was mostly observed in the viral spike protein, which accounts for less than 10% of the virus's total genome.

Indeed, amino acid (aa) mutations confirmed a significant effect on the viral spike protein. Both OV079467.1, with 68 mutations, of which 34% and 33 mutations are unique, and OV090001.1, with 96mutations, of which 62% and 33 mutations are unique, had their spike protein truncated by 64.5% of its original length (>10%). Mutations in the isolates ranged from 15 to 19, with >0.1% unique mutations; however, all were recorded at the spike protein except OV091533.1. All other included genomes had a Spike-D614G mutation recorded during this analysis, as well as Spike-L5F in OV090004.1. OV090004.1 had several other unique mutations that are mostly located on the spike protein. The aa mutation of OV091533.1 was recorded for the first time in Victoria, Australia, in March 2020, and in Trentino-Alto Adige, Italy, in March 2020 for mutations in OV090004.1. All other mutations in the remaining current genomes were recorded for the first time at Port of Spain, Trinidad and Tobago, in January 2020, GISAID.

4. Discussion

Mutations in RNA viruses, including SARS-CoV-2, are common, and they can alter several properties, such as the ability of the virus to cause disease, how easily it can infect its host, how strongly it binds to the host's receptors, and how it spreads (Harvey et al., 2021; J. Ou et al., 2021). Understanding these mutations in the

SARS-CoV-2 genome is important for understanding the nature of the infection and for documenting the virus's evolution and spread (Kim et al., 2020; Shen et al., 2020).

In the present study, SARS-CoV-2 strains in the eastern region of Saudi Arabia showed 100% bootstrap values separating the basal sequences from the two main clusters. Both clusters are close to Wuhan and showed nonsignificant population expansion or distribution, as indicated by Tajima's test. The phylogenetic topology obtained in this study clearly distinguished at the backbone level, with the separation of the three basal isolates (OV090001.1, OV090750.1, and OV079467.1) both main clusters, as well as the divergence between the two clusters themselves, both receiving maximal bootstrap support (100%). This indicates strong statistical confidence in the overall branching pattern and the major evolutionary relationships among the sampled Saudi isolates.

In contrast, two internal groupings within the clusters including OU998690.1/OV005890.1 (72%) and OV091561.1/OV009572.1 (58%) received comparatively lower bootstrap support. While a value of 72% falls within an acceptable range and lends moderate confidence to the close relationship between OU998690.1 and OV005890.1, the 58% support for the OV091561.1/OV009572.1 pairing falls below the conventional threshold of ≥70% typically used to designate a well-supported node (Hillis & Bull, 1993). This weaker support may reflect limited sequence divergence among isolates collected within a short sampling period, insufficient phylogenetic signal at the informative sites distinguishing these two sequences, or a degree of topological uncertainty inherent to closely related, recently diverged viral genomes. Consequently, while these two sequences cluster together in the inferred tree, their sister-taxon relationship should be interpreted with caution and would benefit from confirmation using additional data, such as whole-genome SNP analysis or a larger, more geographically or temporally diverse dataset.

Overall, the strong backbone support underscores the reliability of the major clustering pattern identified in this study, while the weaker internal node values highlight the limitations of resolving fine-scale relationships among highly similar SARS-CoV-2 isolates using a single alignment-based Maximum Likelihood approach. The differences in the recorded frequencies of certain haplotypes in our case may change rapidly due to founder effects caused by local outbreaks. Previous studies have shown that these changes rarely confer a selective advantage on SARS-CoV-2 strains (Hryhorowicz et al., 2021). Indeed, rapid changes in the frequency of certain haplotypes across populations may result from founder effects caused by local outbreaks, and this does not necessarily imply a selective advantage for these strains. Therefore, an unguided assumption about the prevalence of any given SARS-CoV-2 strain indicates that its advanced virulence should be avoided, and some coronavirus mutations require further investigation due to their global spread (Hryhorowicz et al., 2021). The SARS-CoV-2 genome evolves relatively slowly, thanks to the "proofreading" activity of the Nsp12 exonuclease; the SARS-CoV-2 mutation rate is slower than that of other RNA viruses (van Dorp et al., 2020). Previous results showed that the nucleotide mutation rate of the whole genome was 6.677×10^{-4} substitutions per site per year, and the nucleotide mutation rate of the S gene was 8.066×10^{-4} substitutions per site per year, which was at a medium level compared with other RNA viruses (Wang et al., 2022).

At 30 kb, coronaviruses possess the largest genome, which includes structural and accessory genes, ample replicases, and other nonstructural proteins (Nsps) (E et al., 2020). The majority of the SARS-CoV-2 genome contains the ORF1a/b region that is translated into 2 polyproteins, pp1a (Nsp1-Nsp11) and pp1ab (Nsp1-Nsp16) (Cui, Li, and Shi, 2019). The other 4 structural proteins, envelope (E), matrix (M), nucleocapsid (N) phosphoprotein, and spike (S), function together with viral RNA and Nsp1-16 to assist the replication of the virus within the host cell (Hakmaoui et al., 2021). Therefore, mutations in small proteins such as the spike may be unnoticed when analyzing the whole genome. Although the spike protein accounts for less than 10% of the SARS-CoV-2 genome, it plays a critical role in facilitating viral entry into the human body. Our results showed that the vast majority of mutations occurred in the gene encoding the spike protein, and 426 of 428 mutations (99.53%) were detected in this gene. This corroborates previous findings that the spike protein is one of the mutational hot spots (Kubik et al., 2021; Lee et al., 2021; Thakur et al., 2022).

A study of infectivity and reactivity to a panel of neutralizing antibodies and sera from convalescent patients confirmed several mutations and glycosylation site modifications in human SARS-CoV-2 spike proteins (Korber et al., 2020). The SARS-CoV-2 spike protein has an S1 domain at the N-terminus comprising a receptor-binding domain (RBD) and an S2 domain at the C-terminus containing fusion peptides. The spike RBD binds to the host cell's ACE2 receptor, triggering proteolytic cleavage of the S1 and S2 domains, exposing fusion peptides that then insert into the host plasma membrane, facilitating the virus's entry into the host cell (Huang et al., 2020). The D614G mutation is reported to be relatively common and known to increase infection rates (Singh et al., 2021). The p. D614G substitution in the spike protein-A23403G (Brufsky, 2020; Korber et al., 2020), is recorded in both genomes OV079467.1 with 68 mutations, of which 34% are unique, and OV090001.1 with 96 mutations, of which

62% are unique and have their spike protein truncated by 64.5% of its original length. This mutation is located between the receptor binding domain (amino acids 451-509) and the polybasic cleavage site (amino acids 682-685) (Boddy et al., 2007), which regulates the capability of SARS-CoV-2 to bind to the human host ACE2 receptor or is involved in other steps related to the invasion of host cells (Andersen et al., 2020; X. Ou et al., 2020). This might explain the rapid spread of infections in the eastern region at the beginning of the pandemic. Except for OV091533.1 and OV079467.1, the D614G mutation in the spike protein was observed in all genomes, among the other 428 aa mutations, with almost 50% of them unique based on GISAID.

A rare mutation, Spike-I472V, was also recorded in both genomes OV079467.1 and OV090001.1, which is known to work in favor of the virus by increasing both infectivity and resistance to neutralizing antibodies (Li et al., 2020). The current study also includes some deletion mutations at glycosylation sites in the spike proteins. These types of spike mutations are known to reduce infection, revealing the importance of glycosylation for viral infectivity (Guruprasad, 2021). A total of 22 deletion mutations were recorded in both genomes OV079467.1 and OV090001.1, and all were located on the spike protein. In addition, only one mutation (N-A313V) was recorded in the nucleocapsid (N) phosphoprotein in this study, which likely involves cooperative interactions between the dimerization domain and other regions of the protein, as well as the bound RNA (Surjit et al., 2004). However, this has yet to be confirmed, as patient antibodies targeting this domain may not be reliably detected with tests using the reference N protein (Ye et al., 2020).

The current study may enhance our understanding of COVID-19, especially during the ongoing pandemic, when the scientific community is learning about this new viral pathogen. This can be achieved through a better understanding of the main drivers of pathogen emergence, spread, and dominance over human defenses, as these mutations could adversely affect all efforts at drug repurposing. The SARS-CoV-2 S protein is an important component that determines virus virulence, tissue tropism, and host range, as well as a major target for neutralizing antibodies and vaccine development. Mutations at the interface between the spike protein and the ACE-2 receptor could affect vaccine performance and the efficacy of drugs targeting protein-protein interactions. Therefore, the mutations identified in this study should be considered in antibody, vaccine, and drug development.

5. Limitations

This study should be interpreted within the context of its scope and design. As a descriptive genomic surveillance analysis based on 18 publicly available sequences, the findings offer preliminary insights into the evolutionary relationships among circulating SARS-CoV-2 isolates in the Eastern Region of Saudi Arabia, rather than population-level epidemiological conclusions. Additionally, as the analysis relied on sequences deposited in NCBI, individual patient clinical and demographic metadata (such as age, sex, disease severity, and vaccination status) were not accessible, which is a common constraint of public sequence repository-based studies. Future work incorporating larger, longitudinal datasets with linked clinical annotation would help extend these findings and further clarify potential associations between genomic variation and clinical outcomes in this part of the world.

1. Consent for publication

- Not applicable.

2. Competing interests

- The authors declare that they have no competing interests.

3. Funding

- This work was not funded.

4. Authors' contributions

- The author prepared, analyzed, read, and approved the final manuscript.

5. Acknowledgements

- The author extends his appreciation to the University of Northern Border for its support during this work. The appreciation is also linked to Dr. Arbi Guetat for his assistance during the preparation and writing of this work.

6. References

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. In Nature Medicine. https://doi.org/10.1038/s41591-020-0820-9

Boddy, L., Frankland, J., & West, P. Van. (2007). Ecology of saprotrophic basidiomycetes.

Brufsky, A. (2020). Distinct viral clades of SARS-CoV-2: Implications for modeling of viral spread. In Journal of Medical Virology. https://doi.org/10.1002/jmv.25902

Chu, D. K. W., Pan, Y., Cheng, S. M. S., Hui, K. P. Y., Krishnan, P., Liu, Y., Ng, D. Y. M., Wan, C. K. C., Yang, P., Wang, Q., Peiris, M., & Poon, L. L. M. (2020). Molecular Diagnosis of a Novel Coronavirus (2019-nCoV) Causing an Outbreak of Pneumonia. Clinical Chemistry. https://doi.org/10.1093/clinchem/hvaa029

Corse, E., & Machamer, C. E. (2000). Infectious Bronchitis Virus E Protein Is Targeted to the Golgi Complex and Directs Release of Virus-Like Particles. Journal of Virology. https://doi.org/10.1128/jvi.74.9.4319-4326.2000

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792-1797. https://doi.org/10.1093/nar/gkh340

Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R. H., Peters, B., & Sette, A. (2020). A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host and Microbe. https://doi.org/10.1016/j.chom.2020.03.002

Guruprasad, L. (2021). Human SARS CoV-2 spike protein mutations. Proteins: Structure, Function and Bioinformatics. https://doi.org/10.1002/prot.26042

Hakmaoui, A., Khan, F., Liacini, A., Kaur, A., Berka, Y., Machraoui, S., Soualhine, H., Berka, N., Rais, H., & Admou, B. (2021). Relevant SARS-CoV-2 Genome Variation through Six Months of Worldwide Monitoring. BioMed Research International. https://doi.org/10.1155/2021/5553173

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. In Nature Reviews Microbiology. https://doi.org/10.1038/s41579-021-00573-0

He, R., Leeson, A., Ballantine, M., Andonov, A., Baker, L., Dobie, F., Li, Y., Bastien, N., Feldmann, H., Strocher, U., Theriault, S., Cutts, T., Cao, J., Booth, T. F., Plummer, F. A., Tyler, S., & Li, X. (2004). Characterization of protein-protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. Virus Research. https://doi.org/10.1016/j.virusres.2004.05.002

Hillis, D. M., & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology, 42(2), 182-192. https://doi.org/10.1093/sysbio/42.2.182

Holmes, K. V., & Enjuanes, L. (2003). The SARS coronavirus: A postgenomic era. In Science. https://doi.org/10.1126/science.1086418

Hryhorowicz, S., Ustaszewski, A., Kaczmarek-Ryś, M., Lis, E., Witt, M., Pławski, A., & Ziętkiewicz, E. (2021). European context of the diversity and phylogenetic position of SARS-CoV-2 sequences from Polish COVID-19 patients. Journal of Applied Genetics. https://doi.org/10.1007/s13353-020-00603-2

Huang, Y., Yang, C., Xu, X. feng, Xu, W., & Liu, S. wen. (2020). Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. In Acta Pharmacologica Sinica. https://doi.org/10.1038/s41401-020-0485-4

Kim, J. S., Jang, J. H., Kim, J. M., Chung, Y. S., Yoo, C. K., & Han, M. G. (2020). Genome-Wide Identification and Characterization of Point Mutations in the SARS-CoV-2 Genome. Osong Public Health and Research Perspectives. https://doi.org/10.24171/j.phrp.2020.11.3.05

Kimura, M. (1980). Journal of Molecular Evolution A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. In J. Mol. Evol (Vol. 16).

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., Angyal, A., Brown, R. L., Carrilero, L., … Montefiori, D. C. (2020). Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. https://doi.org/10.1016/j.cell.2020.06.043

Kubik, S., Arrigo, N., Bonet, J., & Xu, Z. (2021). Mutational hotspot in the sars-cov-2 spike protein n-terminal domain conferring immune escape potential. Viruses. https://doi.org/10.3390/v13112114

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msy096

Kuo, L., & Masters, P. S. (2002). Genetic Evidence for a Structural Interaction between the Carboxy Termini of the Membrane and Nucleocapsid Proteins of Mouse Hepatitis Virus. Journal of Virology. https://doi.org/10.1128/jvi.76.10.4987-4999.2002

Lee, S., Lee, M. K., Na, H., Ahn, J., Hong, G., Lee, Y., Park, J., Kim, Y., Kim, Y. T., Kim, C. K., Lim, H. S., & Lee, K. R. (2021). Comparative analysis of mutational hotspots in the spike protein of SARS-CoV-2 isolates from different geographic origins. Gene Reports. https://doi.org/10.1016/j.genrep.2021.101100

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., Qin, H., Wang, M., Lu, Q., Li, X., Sun, Q., Liu, J., Zhang, L., Li, X., Huang, W., & Wang, Y. (2020). The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. Cell. https://doi.org/10.1016/j.cell.2020.07.012

Marra, M. A., Jones, S. J. M., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S. N., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y., Cloutier, A., Coughlin, S. M., Freeman, D., Girn, N., Griffith, O. L., Leach, S. R., Mayo, M., McDonald, H., Montgomery, S. B., … Roper, R. L. (2003). The genome sequence of the SARS-associated coronavirus. Science. https://doi.org/10.1126/science.1085953

Memish, Z. A., Zumla, A. I., Al-Hakeem, R. F., Al-Rabeeah, A. A., & Stephens, G. M. (2013). Family Cluster of Middle East Respiratory Syndrome Coronavirus Infections. New England Journal of Medicine. https://doi.org/10.1056/nejmoa1303729

Ou, J., Zhou, Z., Dai, R., Zhang, J., Zhao, S., Wu, X., Lan, W., Ren, Y., Cui, L., Lan, Q., Lu, L., Seto, D., Chodosh, J., Wu, J., Zhang, G., & Zhang, Q. (2021). V367F Mutation in SARS-CoV-2 Spike RBD Emerging during the Early Transmission Phase Enhances Viral Infectivity through Increased Human ACE2 Receptor Binding Affinity. Journal of Virology. https://doi.org/10.1128/jvi.00617-21

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., Guo, L., Guo, R., Chen, T., Hu, J., Xiang, Z., Mu, Z., Chen, X., Chen, J., Hu, K., Jin, Q., Wang, J., & Qian, Z. (2020). Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nature Communications. https://doi.org/10.1038/s41467-020-15562-9

Paradis, E. (2010). Pegas: An R package for population genetics with an integrated-modular approach. In Bioinformatics. https://doi.org/10.1093/bioinformatics/btp696

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. https://doi.org/10.1093/bioinformatics/btg412

Pinheiro, H. P., de Souza Pinheiro, A., & Sen, P. K. (2005). Comparison of genomic sequences using the Hamming distance. Journal of Statistical Planning and Inference. https://doi.org/10.1016/j.jspi.2003.03.002

Pyrc, K., Berkhout, B., & van der Hoek, L. (2007). Identification of new human coronaviruses. In Expert Review of Anti-Infective Therapy. https://doi.org/10.1586/14787210.5.2.245

Rahman, A., & Sarkar, A. (2019). Risk factors for fatal Middle East respiratory syndrome coronavirus infections in Saudi Arabia: Analysis of the WHO Line list, 2013-2018. American Journal of Public Health. https://doi.org/10.2105/AJPH.2019.305186

Saha, P., Banerjee, A. K., Tripathi, P. P., Srivastava, A. K., & Ray, U. (2020). A virus that has gone viral: Amino acid mutation in S protein of Indian isolate of Coronavirus COVID-19 might impact receptor binding, and thus, infectivity. Bioscience Reports. https://doi.org/10.1042/BSR20201312

Schoeman, D., & Fielding, B. C. (2019). Coronavirus envelope protein: Current knowledge. In Virology Journal. https://doi.org/10.1186/s12985-019-1182-0

Shen, L., Dien Bard, J., Biegel, J. A., Judkins, A. R., & Gai, X. (2020). Comprehensive Genome Analysis of 6,000 USA SARS-CoV-2 Isolates Reveals Haplotype Signatures and Localized Transmission Patterns by State and by Country. Frontiers in Microbiology. https://doi.org/10.3389/fmicb.2020.573430

Singh, J., Ehtesham, N. Z., Rahman, S. A., & Hasnain, S. E. (2021). Structure-function investigation of a new VUI-202012/01 SARS-CoV-2 variant. BioRxiv.

Surjit, M., Liu, B., Kumar, P., Chow, V. T. K., & Lal, S. K. (2004). The nucleocapsid protein of the SARS coronavirus is capable of self-association through a C-terminal 209 amino acid interaction domain. Biochemical and Biophysical Research Communications. https://doi.org/10.1016/j.bbrc.2004.03.154

Surya, W., Li, Y., & Torres, J. (2018). Structural model of the SARS coronavirus E channel in LMPG micelles. Biochimica et Biophysica Acta - Biomembranes. https://doi.org/10.1016/j.bbamem.2018.02.017

Thakur, S., Sasi, S., Pillai, S. G., Nag, A., Shukla, D., Singhal, R., Phalke, S., & Velu, G. S. K. (2022). SARS-CoV-2 Mutations and Their Impact on Diagnostics, Therapeutics and Vaccines. In Frontiers in Medicine. https://doi.org/10.3389/fmed.2022.815389

van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., Owen, C. J., Pang, J., Tan, C. C. S., Boshier, F. A. T., Ortiz, A. T., & Balloux, F. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution. https://doi.org/10.1016/j.meegid.2020.104351

Wang, S., Xu, X., Wei, C., Li, S., Zhao, J., Zheng, Y., Liu, X., Zeng, X., Yuan, W., & Peng, S. (2022). Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. Journal of Medical Virology. https://doi.org/10.1002/jmv.27331

Ye, Q., West, A. M. V., Silletti, S., & Corbett, K. D. (2020). Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. Protein Science. https://doi.org/10.1002/pro.3909

Zhu, Y., Liu, M., Zhao, W., Zhang, J., Zhang, X., Wang, K., Gu, C., Wu, K., Li, Y., Zheng, C., Xiao, G., Yan, H., Zhang, J., Guo, D., Tien, P., & Wu, J. (2005). Isolation of virus from a SARS patient and genome-wide analysis of genetic mutations related to pathogenesis and epidemiology from 47 SARS-CoV isolates. Virus Genes. https://doi.org/10.1007/s11262-004-4586-9