JIDC | THE JOURNAL OF INFECTION IN DEVELOPING COUNTRIES

Coronavirus Pandemic

SARS-CoV-2 genetic diversity and variants of concern in Saudi Arabia

Dalia Abdullah Obeid[1,2], Madain Saleh Alsanea[1], Rawan Talal Alnemari[2], Ahmed Ali Al-Qahtani[1,3], Sahar Isa Althawadi[4], Maysoon Saleh Mutabagani[4], Reem Saad Almaghrabi[5], Faten Mohammed Alhadheq[1], Basma Mohammed Alahideb[1], Fatimah Saeed Alhamlan[1,3,4]

[1] Department of Infection and Immunity, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[2] Department of infectious Genome, Public Health Authority, Riyadh, Saudi Arabia
[3] College of Medicine, Alfaisal University, Riyadh, Saudi Arabia
[4] Department of Pathology and Laboratory Medicine, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[5] Department of Medicine, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

Abstract

Introduction: In December 2019, a new severe acute respiratory syndrome coronavirus, SARS-CoV-2, emerged in China, causing coronavirus disease 2019. The present study investigated genetic profiles and variations of SARS-CoV-2 distributed in different regions of Saudi Arabia to begin to understand the pathogenesis and transmission of SARS-CoV-2 in this country and analyzed associations of these variations with host factors.

Methodology: In total, 774 SARS-CoV-2 genomic sequences obtained and annotated by the Global Initiative on Sharing All Influenza Data (GISAID) were captured and analyzed.

Results: The most common SARS-CoV-2 clades in Saudi Arabia were GH followed by O, GR, G, and S. Statistically significant associations were detected between clades and patient outcome. Age, as a host factor, was significantly associated with many variables, including virus geographical location, clade, and patient outcome. The most common variants detected were the NSP12_P323L mutation 94.9%, followed by the D614G mutation (76%) and the NS3_Q57H mutation (71.4%). The concerned variants B.1.1.7, B.1.351, and P.1 were not detected in our population. D614G was associated with higher morbidities than the wild-type virus, including higher rates of death and hospitalization. The NS3_Q57H mutation was the only variant associated with better patient outcome than the wild type. Risk of death was highest with the NSP12_P323L mutation (OR = 1.84; 95% CI = 0.37-9.30) and lowest with the NS3_Q57H mutation (OR = 0.43; 95% CI = 0.25-0.727).

Conclusions: SARS-CoV-2 has evolved uniquely and independently in Saudi Arabia. Our findings provide evidence to begin linking the evolutionary implications to host factors and their effects on the virus severity and transmission.

Key words: SARS-CoV-2; COVID-19; genome variants; genome diversity.

J Infect Dev Ctries 2021; 15(12):1782-1791. doi:10.3855/jidc.15350

(Received 18 May 2021 – Accepted 21 September 2021)

Introduction

In late 2019, China reported the first case of coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus that rapidly became a global pandemic [1,2]. SARS-CoV-2 causes a respiratory syndrome with a variable degree of severity, ranging from a mild upper respiratory tract illness to severe acute respiratory syndrome [3,4]. As of May 15, 2021, more than 126 million cases, including 3.35 million deaths, were reported to the World Health Organization worldwide, while 431,000 cases, including 7,134 deaths, were reported in Saudi Arabia. The first gene sequence analysis for SARS-CoV-2 revealed that the virus has sequence similarity of over 80% with severe acute respiratory syndrome (SARS) in 2002 and 50% with MERS-CoV that caused the Middle East respiratory syndrome (MERS) in 2012 [4].

The severity and fatality of COVID-19 seemed to be determined by several factors some of which include but are not limited to old age (> 65 years) and pre-existing comorbidities such as cardiovascular diseases, immunological diseases, diabetes mellitus and respiratory diseases [2,5,6]. Genetic variations in the SARS-CoV-2 genome may occur due to random genetic drift or to natural selection [6,7]. At the beginning of 2020, the genetic sequence of SARS-CoV-2 was published [8]. The virus has a single positive-stranded RNA—which has a higher mutation rate than DNA viruses—consisting of approximately 30,000 nucleotides [2,3,9]. The genome contains 14 open reading frames (ORFs) encoding structural proteins and

non-structural proteins (NSPs) [2,4,10,11]. Structural proteins include spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins with eight accessory proteins, five of which are encoded by ORF3a, ORF6, ORF7a ORF8, and ORF10 genes. The ORF1 and ORF2a on the other hand encode 15 non-structural proteins important for virus replication and transcription [4,11].

A few months after the first reported case of COVID-19, 329 naturally occurring variants in the S protein were reported in the public domain [3]. Some implications of mutations in S protein might be increased or reduce the binding efficiency to host cell receptor which may enhance or suppress intracellular viral entry virus. In the latter case, reduced binding may aid immune evasion. The most significant variant in the SARS-CoV-2 S protein detected to date is a mutation of the aspartate (D) at position 614, which is found in nearly all Chinese variants, to a glycine (G), which is enriched in European variants [8]. The D614G variant may have caused fatal infections in European populations because although Germany and Kuwait have a substantial number of this mutation and high mortality rates, viruses with the wild-type 614D, which have lower mortality rates, are dominant in these countries. Novel mutations continue to emerge, potentially resulting in variants with greater virulence and higher mortality rates or strains resistant to treatment [2]. At the beginning of 2021, new variants of concerns emerged worldwide, these included B.1.1.7 in England, B.1.351 in South Africa, and P.1 in Japan. These variants of concern were linked to higher transmission rate [12–14].

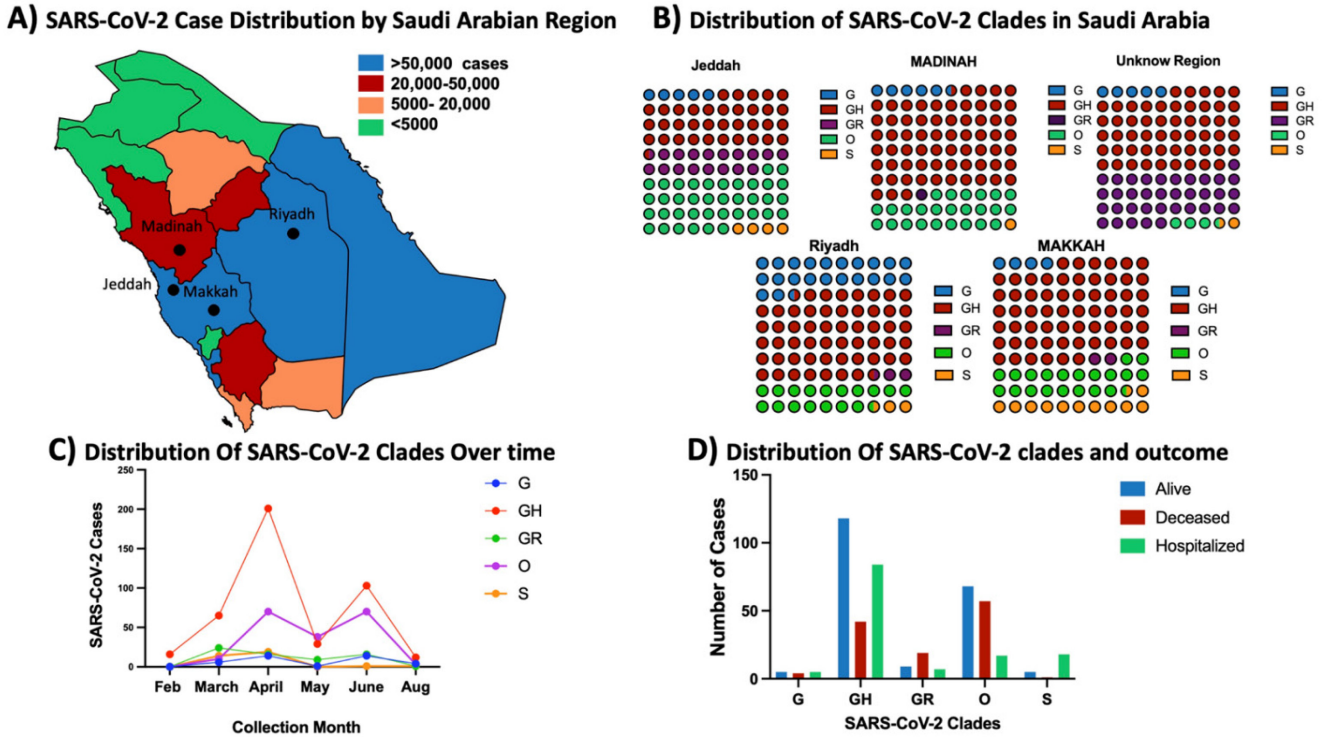
In Saudi Arabia, more than 431,000 cases of SARS-CoV-2 have been detected as of May 15th, 2021 [15]. However, the genetic variations within the Saudi population have not yet been evaluated. Therefore, the present study investigated the genetic variations of SARS-CoV-2 distributed in different areas of Saudi Arabia. Analyzing these genetic diversities will identify markers vital to better understanding the pathogenesis and transmission of this virus. By studying the genetic profile of SAR-CoV-2 and its common mutations in Saudi Arabia, the evolutionary profile of the virus will be known and may assist clinicians in COVID-19 treatment decisions and in predicting infected patients' outcomes.

Methodology

Data collection

In total, 774 SARS-CoV-2 genomic sequences were obtained from the Global Initiative On Sharing All

Figure 1. A) Map of Saudi Arabia showing the distribution of SARS-CoV-2 cases. B) Distribution of SARS-CoV-2 clades in Saudi Arabia, each circle represents 1% of the total number of cases represented in each region. C) Distribution of SAR-CoV-2 cases over time showing that most belonged to the GH clade; however, the O clade was higher in June. D) Distribution of SARS-CoV-2 clades in Saudi Arabia stratified by patient outcome.



Influenza Data (GISAID) website [16]. The submitters to this website are provided in the Annex section. Only viruses widely distributed throughout Saudi Arabia and with complete genomes (> 29,000 nucleotides) were included in the analysis. Variables, including sample collection site location, gene mutations, virus lineage, and clade as well as patient sex, age, and clinical outcome, were collected as identified on the GISAID website. Other classifications for the number of mutations were determined using the Nextclade tool on the Nextstrain website [17]. This study was approved by the ethics committee of King Faisal Specialist Hospital and Research Centre, which waived the need for obtaining informed consent because all data used in the study were de-identified and are publicly available.

Data and statistical analyses

Data were analyzed using the Statistical Analytical Software SAS, version 9.4 (SAS Institute Inc.; Cary, NC, USA) and GraphPad, version 8 (San Diego, CA, USA). Data were stratified by age for all the categorical variables, and a *t* test or an analysis of variance (ANOVA) was conducted as appropriate to assess significance by clinical group. Descriptive analysis was conducted for the distribution of cases by group with time using bar graphs and dot plots. Logistic model was used to estimate the risk of each mutation the outcome. Mutations as determined with the GISAID pipeline and Nextclade tool were tested for their significant association with patient outcomes using the chi-square test. Gene sequences were aligned with international sequences, and a phylogeny tree was constructed using the Nextstrain phylogeny tree tool. A 2-sided *p*-value of alpha < 0.05 was considered statistically significant.

Results

Patient demographic characteristics and SARS-CoV-2 case and clade distributions

The mean (standard deviation) age of infected patients was 47.9 (15) years. Most genome sources (379 [49%]) were from males. The primary collection sites were Jeddah (222 [28.7%]), Madinah (263 [33.9%]), Makkah (180 [23.3%]), Riyadh region (40 [5.1%]), and Qatif region (8 [1.1%]), while around (61 [7.9%]) were unknown.

The highest numbers of patients with SARS-CoV-2 infections (cases) in Saudi Arabia were located in Riyadh, Makkah, Jeddah, and Eastern Province [15]. In Jeddah, the most common clade was GISAID clade O, and in Makkah, Madinah, and the unknown region, the most common clade was GISAID clade GH. Figure 1 shows the distributions of both SARS-CoV-2 infections (cases) and clades by region. Overall, the most common clade in Saudi Arabia was GH, followed by O, GR, G, and S. Figure 1. D shows the distribution of SARS-CoV-2 clades by patient outcomes. Clade O was associated with the highest death rate, and clade GH was associated with the most hospitalized patients who and patients who survived. The association between clade and patient outcome was statistically significant ($\chi^2 = 116$, $p < 0.001$).

Association of age with other reported variables

A test for association between age and different patient parameters such as clade classification, location, numbers of mutations and patient outcome indicated a significant difference as shown in Supplementary Table 1. Age differed by geographical location, with patients in Jeddah significantly older than those in Madinah or Makkah (ANOVA test *p*-value = 0.0004), and by clade classification. Clades O and G were mostly detected in

Table 1. Distribution of mutations in the SARS-CoV-2 genome detected in Saudi Arabia stratified by patient outcome.

Mutation		Patient Outcome, No.					χ^2 (*p*-value)[A]
		Alive	Deceased	Hospitalized	Unknown	Total	
Spike D614G	Negative	64	50	28	42	184	10.6 (0.0049)*
	Positive	141	73	103	273	590	
NSP12_P323L	Negative	7	2	16	14	39	18.6 (0.0001)*
	Positive	198	121	115	301	735	
N_S194L	Negative	165	103	122	72	462	9.9 (0.007)*
	Positive	29	16	9	7	71	
NS8_L84S	Negative	200	121	113	303	737	27.2 (< 0.001)*
	Positive	5	2	18	12	37	
NS3_Q57H	Negative	35	58	41	87	221	34.8 (< 0.001)*
	Positive	170	65	90	228	553	
N_R203K	Negative	181	74	119	273	647	49.9 (< 0.001)*
	Positive	24	49	12	42	127	
N_G204R	Negative	184	70	117	274	645	60.9 (< 0.001)*
	Positive	21	53	14	41	129	

[A] *p*-value calculated without unknown patient data< * Statistically significant with alpha equal to 0.05.

older patients and clade S in the youngest (ANOVA test p -value = 0.0033). Patient outcome was significantly associated with age, with deceased patients associated with older age (ANOVA test p-value < 0.05). Age group of patients younger and older than 50 years old showed no significant difference with mutation load (above 8 or below 8 mutations per sample), overall younger patients reported 29.9% higher load of mutation compared to 9.3% reported in the older age group (χ^2=0.14, p-value = 0.71).

Synonymous mutations and their association with patient outcome

The mean number of gene mutations per sample was 8.5. Mutations detected by the GISAID pipeline were collected and re-analyzed by patient outcome. The most common gene mutations in our study population are summarized in Table 1. The first mutation was found in the region coding for the Spike protein, which is the D614G mutation. This variation is caused by the replacement of aspartic acid with glycine. The mutation was detected in 76.2% of our samples and was strongly associated with higher morbidities, including death and hospitalization (p-value = 0.0049). The second mutation NSP12_P323L was in a non-structural protein, NSP12, which encodes the viral RNA-dependent RNA polymerase. This mutation was also common in our population (94.9%) and was associated with severe morality (p-value < 0.0001). The third mutation was located in the N gene N_S194L. This mutation was less commonly detected in our population (19.5%) but was still associated with higher morbidities, including death and hospitalization (p-value = 0.007). The fourth mutation NS8_L84S is located in the non-structural NS8 gene, which is part of the ORF8 polyprotein. This mutation is rarer compared with 4.8% occurrence frequency and it is only associated with higher hospital admissions (p-value < 0.0001). The fifth mutation, NS3_Q57H, is located in the NS3 gene portion of the ORF3a polyprotein. It is a high frequency of occurrence at 71.5% within the Saudi population and it is associated with favorable patient outcomes (p-value < 0.0001). The sixth and seventh most common mutations are located in the N gene, N_R203K and N_G204R. These two mutations were each detected in approximately 16% of the population and were associated with severe morbidities. Other less common mutations in SARS-CoV-2 are reported by patient outcome in Supplementary Table 2. All of those mutations were found in the N and NSP genes, and many were associated with severe morbidities.

Geographical phylogenetic analysis of the SARS-CoV-2 variants

Overall, the summary of the phylogeny analysis is shown in Figure 2. A summary of the phylogeny analysis revealed that the first viruses detected in Saudi Arabia sequenced in Riyadh were located at the 19B node, with many of them characterized by an ORF8 gene mutation L84S. The 19B node viruses were mostly from Makkah, with the first virus isolated in Saudi Arabia in February 2020. The second major node was 20A and was characterized by an ORF1b mutation located at P314L. The divergence rate of the clade from

Table 2. Patient outcome predictions stratified by SARS-CoV-2 variant and mutation load (likelihood test with 95% Wald odds ratios).

Mutation	Outcome, Alive = 0	Odds Estimate (95% Confidence Limits)	Global χ^2 p-value
SPIKE D614G			
Negative (0)	Hospitalized	0.63(0.38-1.1)	0.0048*
	Deceased	1.56(0.98-2.48)	
Positive (1)	Hospitalized	1.59(0.95-2.65)	
	Deceased	0.64(0.40-1.03)	
NSP12_P323L	Outcome, Alive = 0		
Negative (0)	Hospitalized	4.69(1.85-13.26)	0.0002*
	Deceased	0.54(0.11-2.73)	
Positive (1)	Hospitalized	0.215(0.082-0.56)	
	Deceased	1.84(0.37-9.30)	
NS3_Q57H	Outcome, Alive = 0		
Negative (0)	Hospitalized	2.32(1.38-3.91)	<0.0001*
	Deceased	4.4(2.66-7.4)	
Positive (1)	Hospitalized	0.23(0.14-0.38)	
	Deceased	0.43(0.25-0.73)	
Mutation Load	Outcome, Alive = 0		
Low (mutations ≤ 8)	Hospitalized	2.45(1.46-4.16)	0.0001*
	Deceased	0.79(0.50-1.25)	
High (mutations > 8)	Hospitalized	0.41(0.24-0.69)	
	Deceased	1.26(0.79-1.99)	

*p < 0.05 was considered statistically significant.

the Wuhan strain was 3.8. The 20A node gave rise to multiple sub-nodes including viruses from Jeddah, Madinah and Makkah cities. This node is similar to the Asian and American viruses which is explained by the first reported cases to have been from patients with travel history from the USA. Subsequently, this node gave rise to a sub-node which has a mutation within the ORF1a polyprotein (Q57H), which is a common mutation in the population. The third major node was 20C, which is characterized by a mutation in ORF1a polyprotein (T65I). this node has a divergence rate of 5.7 and was mostly from Jeddah and very close to the American strains. The fourth main node was 20B, with a divergence rate of 6.8 and characterized with mutations located at the N gene and ORF14 gene. This clade was also mostly found in Jeddah.

Distribution and evolutionary timeline of common SAR-CoV-2 variants

One of the most globally reported SAR-CoV-2 variants is that with the D614G spike mutation. This mutation has served as a biomarker to monitor changes in the SARS-CoV-2 virus. The SAR-CoV-2 virus was introduced to Saudi Arabia in February 2020 with the D614G spike mutation present. However, between February and August 2020, increasing numbers of patients infected with the wild-type virus were also reported (Figure 3A). The D614G spike variant was most common in samples collected from Jeddah and Madinah cities (Figure 3B). Outcomes for patients stratified by the presence of the mutation indicated that although more patients with this mutation were hospitalized (32.1% *vs*. 20%), they showed a lower death rate (23.2% *vs*. 35.7%) compared with patients infected with the wild-type virus (Figure 3C).

A second SAR-CoV-2 variant detected in our study population had a mutation located in the ORF3a polyprotein, NS3_Q57H. The timeline between February and August 2020 in Figure 4A shows that the virus was introduced to Saudi Arabia with this mutation present, but that over time, the dominant virus was the wild type. This finding indicates that this mutation is not favorable for virus transmission. The number of SAR-CoV-2 variants with the NS3_Q57H mutation stratified by region is shown in Figure 4B and indicates that its presence was highest in Jeddah and Madinah. Outcomes for patients with this mutation indicated a lower death rate (20.2%) compared with patients with the wild-type virus (43.6%) Figure 4C.

The third mutation of interest was located in the ORF1b gene, the NSP12_P323L mutation. The timeline between February and August 2020 shows that the SARS-CoV-2 viruses introduced to Saudi Arabia contained this mutation and that this mutation remained dominant over the wild-type virus (Figure 5A).

Figure 2. SARS-CoV-2 phylogenetic analysis of the included Saudi viruses conducted using the Nextclade tool, comparing the Saudi viruses with global strains. The first virus sequenced in March is represented by a star and is related to a patient with a travel history. Most of the viruses (n = 325) fall under the 20A node, characterized by a mutation located within the ORF3a gene (Q57H). Gray represents the Saudi viruses; the remaining colors, other global regions.

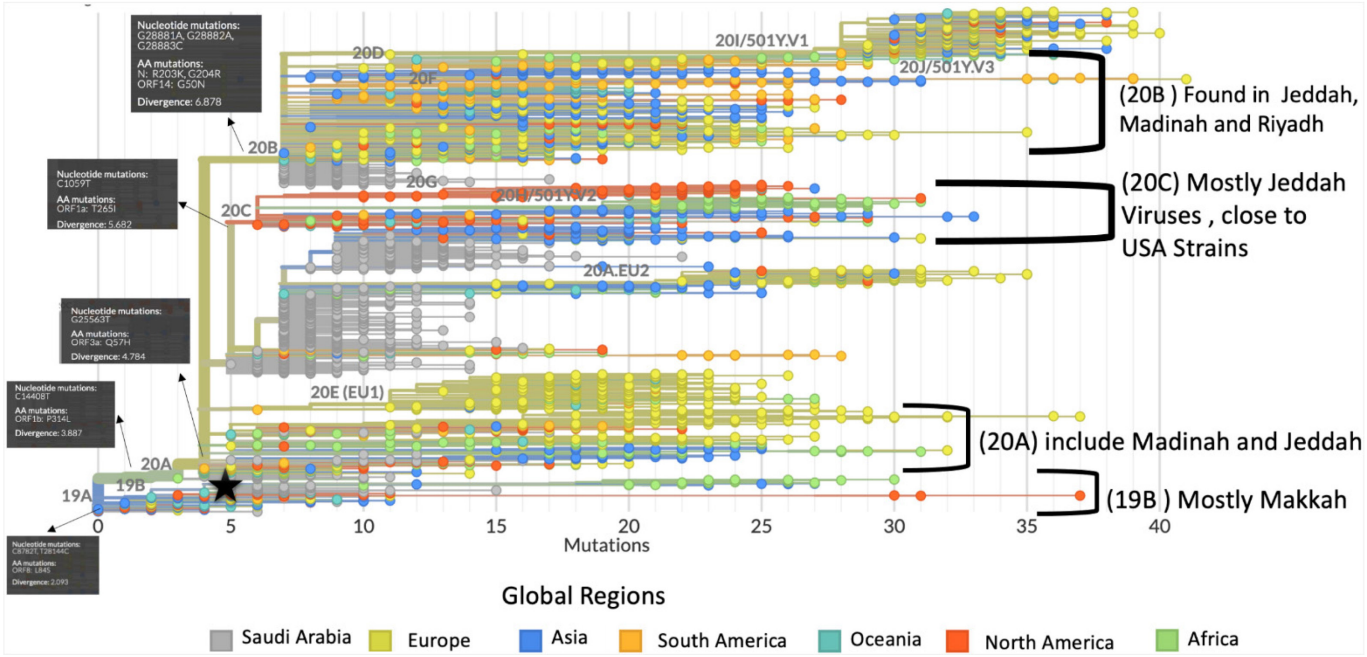


Figure 3. Distribution of D614G mutation in SARS-CoV-2 detected in Saudi Arabia. A) Timeline of collected samples with D614G mutation. The first viruses detected were mutated, blue represents samples harboring the mutation, and red represents samples with the wild type strain. B) The number of (D614G) mutated viruses based on region where the viruses were isolated, the highest number of mutations was reported in Jeddah, Madinah and Makkah. C) The effect of the mutation on patient's outcome.

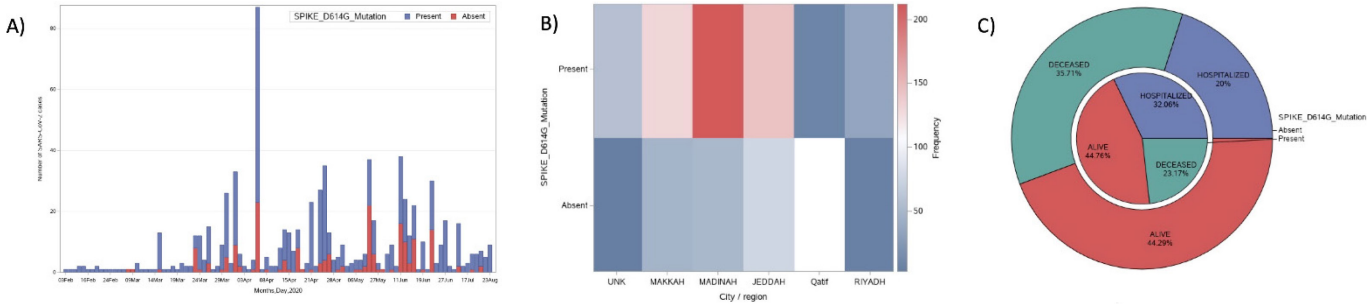


Figure 4. Distribution of NS3_Q57H mutation in SARS-CoV-2 detected in Saudi Arabia. A) Timeline of collected samples with NS3_Q57H mutation. The first viruses detected were mutated, blue represents samples harboring the mutation, and red represents samples with the wild type strain. B) The number of NS3_Q57H mutated viruses based on region where the viruses were isolated, the highest number of mutations was reported in Jeddah and Madinah. C) The effect of the mutation on patient's outcome.

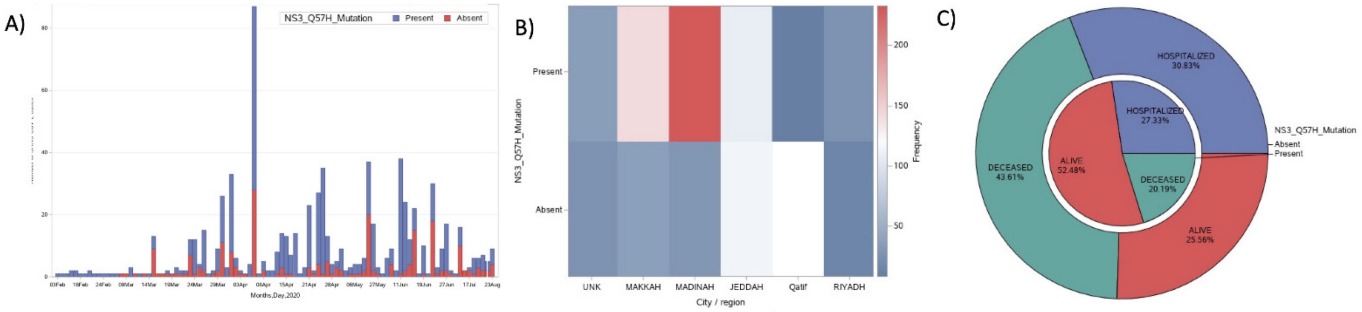
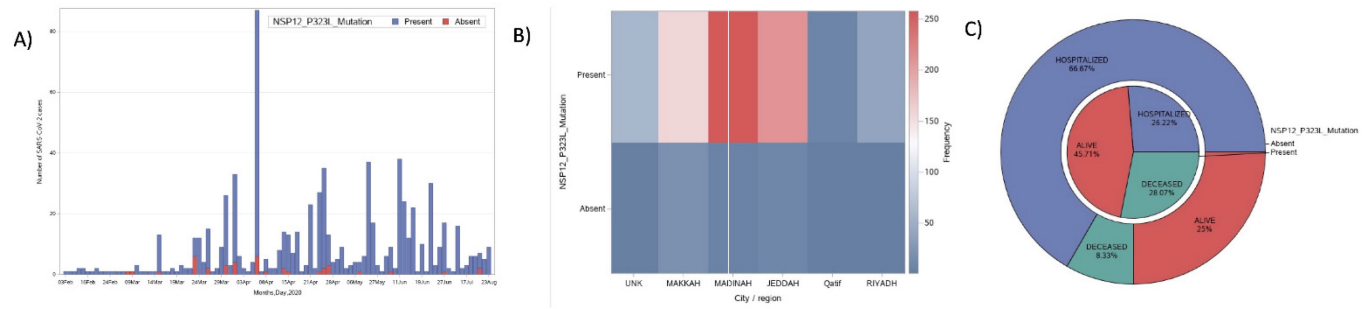


Figure 5. Distribution of NSP12_P323L mutation in SARS-CoV-2 detected in Saudi Arabia. A) Timeline of collected samples with NSP12_P323L mutation. The first viruses detected were mutated, blue represents samples harboring the mutation, and red represents samples with the wild type strain. B) The number of NSP12_P323L mutated viruses based on region where the viruses were isolated, the highest number of mutations was reported in Jeddah and Madinah. C) The effect of the mutation on patient's outcome.



The number of SAR-CoV-2 variants with the NSP12_P323L mutation stratified by region indicates that the mutation was highest in Jeddah and Madinah (Figure 5B). Outcomes for patients indicated that although the hospitalization rate was lower for those with vs. without this mutation (26.22% *vs*. 66.67%), the survival rate was much higher (45.7% *vs*. 25.00%) for those with this mutation compare with patients with the wild-type virus (Figure 5C).

Prediction of patient outcomes stratified by SAR-CoV-2 variant

To assess the SAR-CoV-2 variant and its mutation burden with the prediction of patient outcomes, we used a univariable logistic model (the Wald test) and calculated the confidence intervals for three outcomes of interest (alive, deceased, and hospitalized). The model results are shown in Table 2. The highest risk of death was associated with the NSP12_P323L mutation (HR = 1.84; 95% confidence interval [CI] = 0.37-9.30), and the lowest risk of death was associated with the NS3_Q57H mutation (HR = 0.43; 95% CI = 0.25-0.73). For hospital admissions, the highest risk was found with patients who harbored the wild type of virus for the NSP12_P323L mutation (HR = 4.69; 95% CI = 1.85-13.26). Interestingly, the lower mutation load was linked to a higher risk of hospital admissions (HR = 2.45; 95% CI = 1.46-4.23) Conversely, a high mutation rate was associated with a higher risk of death (HR = 0.79; 95% CI = 0.5-1.25), and lower risk of hospitalization (HR = 0.41; 95% CI = 0.24-0.69).

Discussion

The genetic diversity as well as the host-virus interaction and transmission of SARS-CoV-2 in many geographical locations are uncertain. In the present study, we evaluated the genetic variations of SARS-CoV-2 in Saudi Arabia. Sequencing of the first samples in February revealed that the earliest SARS-CoV-2 virus belonged to clade GH/20C. The present study found that the most common clade in Saudi Arabia was the GH clade, followed by the O, GR, G, and S clades. By region, the most common clade in Jeddah was O/19A, whereas the most common clade in Makkah and Madinah was GH/20C. The O clade was associated with the highest death rate, whereas the GH clade was associated with the most hospitalized patients but the highest rate of patient survival. We found significant associations by patient outcomes in this population. Evaluating age as a host factor showed that age was significantly associated with many variables, including geographical location, clade, and patient outcome. Patients were older in Jeddah than in Madinah or Makkah. Clades O and G were mostly associated with patients older than those in the other clades, clade S, with the youngest. The youngest patients were infected with viruses from the earlier lineages and G clade. Patient outcome and mutation burden were both significant predictors of age: high burden of mutation and deceased patients were both strongly associated with older age. Our phylogeny analysis of this population revealed that the earliest divergence (19B) of the virus occurred mostly in Makkah. The virus then diverged at the 20A node and was distributed to different geographical locations. The third major node was 20C and mostly involved Jeddah. Many of the SARS-COV-2 mutations assessed in our study were introduced to Saudi Arabia as SARS-COV-2 variants, with some of these reverting to the wild-type virus over time.

Unlike some studies in other countries, our study population was composed mainly of older males. However, this finding is consistent with an Icelandic study that found that children under 10 years of age and females had lower SARS-COV-2 infection rates compared with adults, adolescents, or males [18].

The distribution of clades across our population was dominated by the GH/C20 clade. In a European review of clade distribution, the GH clade only dominated in a few countries, including Norway, Denmark, Finland, and Georgia. Over time, the GR/20B clade became increasingly significant in the European Union, whereas in Saudi Arabia, only the GH/20B and O/19A clades were predominating [19]. One review reported four main clades distributed in Asia within the first 6 months of the pandemic: G (44%), S (14%), V (3%), and I (3%), with 36% unassigned genomes [20]. This result likely differs from ours because most of the East Asian countries were infected with the earliest forms of the virus [21]. A recent study analyzed 553 sequences from the Middle East and North African region and found that the most frequent S gene mutation includes D614G (n = 435), Q677H (n = 8), and V6F (n = 5), with a significant increase in the appearance of the D614G mutation from 63% in February to 98.5% in June 2020 [22]. Similar results were found in a recent Moroccan study, with the most detected mutations being D614G and NSP12_P323L [23]. Although, Saudi Arabia is located in Asia, the distribution of introduced virus variants differed from that in eastern Asian countries as well as in European countries. This difference may be linked to virus-host interaction factors such as host factors that interact directly with viral proteins or are involved in signaling pathways, and host immunity.

The average synonymous and nonsynonymous mutation rate of the SARS-CoV-2 genome in our population was 8.5 per sample. This estimation is close to a worldwide estimate of 0.80-2.38 × 10^{-3} nucleotide substitutions per site per year [24]. The most common mutation in Saudi Arabia, NSP12_P323L, was detected in nearly ~95% of the samples in our study, followed by D614G and NS3_Q57H with 76% and 71% occurrence respectfully. The NSP12-P323L mutation is located within the RNA dependent RNA polymerase (RdRp) gene, this gene is responsible for RNA synthesis, a crucial process required during viral replications. Higher mutations in the RdRp gene has been have been reported in samples harboring this special mutation [25,26]. Expectedly, mutation in the RdRp protein will result in a dysfunctional enzyme which will inadvertently result in mistakes during RNA synthesis, increasing the chances of mutation to occur. The spike mutation D614G was found in 76% of the samples in our study and was strongly associated with higher morbidities, including higher rates of death and hospitalization. A global tracking study has shown that the G614 variant in the spike protein has spread faster than D614 [1]. Therefore, G614 has been linked to the infectious ability of the virus as well as to higher viral load; however, that study found no link to disease severity. It is possible that this mutation enhances the cell surface receptor binding of the virus to promote intracellular translocation of surface bound virus.

The other mutations detected in our population (NSP12_P323L, N_S194L, NS8_L84S, N_R203K, and N_G204R) were also associated with severe morbidities in the Saudi population. These mutations mainly effect the RNA-synthesis complex by catalyzing the function of which enhance replication, thus improve overall viral fitness. The NS3_Q57H mutation was the only one associated with better outcomes for patients in our population. For many of the mutations, including D614G, NS3_Q57H, and NSP12_P323L, patient survival rates were associated with the mutated virus rather than the wildtype. Patients who were negative for the NS3_Q57H mutation had the highest rate of hospitalization. The concerned variants B.1.1.7, B.1.351, and P.1 were not detected in our population from Feb to July 2020.

Many SARS-CoV-2 variants have been linked to host-virus interaction as the virus evolves independently in each geographical location around the globe. For example, during the early evolution of SARS-CoV-2, a novel deletion in ORF8 was detected in Taiwan and Singapore, resulting in the removal of the ORF8 transcription regulatory sequence and the elimination of the ORF8 transcription which is a hot spot for SARS-CoV-2 virus host evolution. This deletion has been associated with a reduction in the replicative ability of the virus [21,27]. SARS-CoV-2 in South American countries has shown massive nucleotide variations in the N and RNA-dependent RNA polymerases genes, in contrast to the E gene, which shows no variation and is considered the most conserved and reliable target regarding single gene target testing [28]. One study investigating the human leukocyte antigen (HLA) system as a host factor and the effects of HLA genotypes on SARS-CoV-2 mutations found that the susceptibility to SARS-CoV-2 infection or severity of COVID-19 is highly correlated to HLA genotypes [11]. These results indicate the importance of understanding human host factors linked to geographical locations and their effects on SARS-CoV-2 transmission and severity. In addition, the genetic variations in a single location of the virus will cause less sensitive polymerase chain reaction testing, and thus may also indicate the wrong antiviral medicine, because some of these variations can make the virus different within the targeted region.

Although this study was the first, to our knowledge, to evaluate a substantial number of SAR-CoV-2 genomes isolated from Saudi Arabia, it had a few shortcomings. Variable validity, such as the collection date and source of transmission, was not controlled owing to the difficulty in obtaining such information from each source of the GISAID submitters. Data necessary to evaluate host factor interaction, such as nationality, length of isolation, and travel history, were not available for our population. Many regions of Saudi Arabia were not included in our analysis because no data had been submitted to the GISAID portal or the data had not been defined.

Conclusions

In conclusion, we examined the SAR-CoV-2 genome diversity in samples from patients in Saudi Arabia, exploring the geographical impact and its evolution in this population. Although Saudi Arabia is home to two holy Muslim mosques and has many foreigner workers, both of which could play important roles in SAR-CoV-2 genetic diversity, numerous acts by the Saudi government to prevent the spread of the virus—including the earliest lockdown worldwide outside of China, may have resulted in an evolution of the virus unique and independent from other countries. Our findings provide evidence to begin linking the evolutionary implications to host factors and their effects on the virus severity and transmission.

Funding

This work was supported by King Faisal Specialist Hospital and Research Centre (RAC # 2200009).

Authors Contribution

D.O and F.A: Conceptualization, All authors were involved in manuscript writing and methodology, Formal analysis was done by D.O. F.A and A.Q edited the final manuscript. All authors read and approved the contents of the manuscript.

References

1. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, Angyal A, Brown RL, Carrilero L, Green LR, Groves DC, Johnson KJ, Keeley AJ, Lindsey BB, Parsons PJ, Raza M, Rowland-Jones S, Smith N, Tucker RM, Wang D, Wyles MD, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC (2020) Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182: 812-827.e19.
2. Koyama T, Platt D, Parida L (2020) Variant analysis of SARS-cov-2 genomes. Bull World Health Organ 98: 495–504.
3. 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 182: 1284-1294.e9.
4. Petrosillo N, Viceconte G, Ergonul O, Ippolito G, Petersen E (2020) COVID-19, SARS and MERS: are they closely related? Clin Microbiol Infect 26: 729–734.
5. Abdullahi IN, Emeribe AU, Ajayi OA, Oderinde BS, Amadu DO, Osuji AI (2020) Implications of SARS-CoV-2 genetic diversity and mutations on pathogenicity of the COVID-19 and biomedical interventions. J Taibah Univ Med Sci 15: 258–264.
6. Eaaswarkhanth M, Al Madhoun A, Al-Mulla F (2020) Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? Int J Infect Dis 96: 459–460.
7. Grubaugh ND, Hanage WP, Rasmussen AL (2020) Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell 182: 794–795.
8. Becerra-Flores M, Cardozo T (2020) SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int J Clin Pract 74: 4–7.
9. Islam OK, Al-Emran HM, Hasan MS, Anwar A, Jahid MIK, Hossain MA (2020) Emergence of European and North American mutant variants of SARS-CoV-2 in South-East Asia. Transbound Emerg Dis 68: 824-832.
10. Dawood AA (2020) Mutated COVID-19 may foretell a great risk for mankind in the future. New Microbes New Infect 35: 100673.
11. Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K (2020) SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J Hum Genet 65: 1075-1082.
12. Public Health of England (2021) Investigation of novel SARS-CoV-2 variant. Variant of concern 202012/01. Technical briefing 3. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959360/Variant_of_Concern_VOC_202012_01_Technical_Briefing_3.pdf. Accessed 10 December 2021.
13. Wu K, Werner AP, Moliva JI, Koch M, Choi A, Stewart-Jones GBE, Bennett H, Boyoglu-Barnum S, Shi W, Graham BS, Carfi A, Corbett KS, Seder RA, Edwards DK (2021) mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. bioRxiv Preprint 2021.01.25.427948.
14. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N, Mlisana K, von Gottberg A, Walaza S, Allam M, Ismail A, Mohale T, Glass AJ, Engelbrecht S, Van Zyl G, Preiser W, Petruccione F, Sigal A, Hardie D, Marais G, Hsiao M, Korsman S, Davies M-A, Tyers L, Mudau I, York D, Maslo C, Goedhals D, Abrahams S, Laguda-Akingba O, Alisoltani-Dehkordi A, Godzik A, Wibmer CK, Sewell BT, Lourenço J, Alcantara LCJ, Pond SLK, Weaver S, Martin D, Lessells RJ, Bhiman JN, Williamson C, de Oliveira T (2020) Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. MedRxiv Preprint 2020.12.21.20248640.
15. Ministry of Health, Saudi Arabia (2020) Daily press release. Available: https://twitter.com/SaudiMOH. Accessed: 15 May 2021. [Available in Arabic]
16. Shu Y, McCauley J (2017) GISAID: global initiative on sharing all influenza data - from vision to reality. Euro Surveill 22: 30494.
17. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34: 4121–4123.
18. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, Saemundsdottir J, Sigurdsson A, Sulem P, Agustsdottir AB, Eiriksdottir B, Fridriksdottir R, Gardarsdottir EE, Georgsson G, Gretarsdottir OS, Gudmundsson KR, Gunnarsdottir TR, Gylfason A, Holm H, Jensson BO, Jonasdottir A, Jonsson F, Josefsdottir KS, Kristjansson T, Magnusdottir DN, le Roux L, Sigmundsdottir G, Sveinbjornsson G, Sveinsdottir KE, Sveinsdottir M, Thorarensen EA, Thorbjornsson B, Löve A, Masson G, Jonsdottir I, Möller AD, Gudnason T, Kristinsson KG, Thorsteinsdottir U, Stefansson K (2020) Spread of SARS-CoV-2 in the Icelandic population. N Engl J Med 382: 2302–2315.
19. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, Melidou A, Neher RA, O'Toole Á, Pereyaslov D, group WHOER sequencing laboratories and GE, group* WHOER sequencing laboratories and GE (2020) Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Euro Surveill 25: 2001410.
20. Guan Q, Sadykov M, Mfarrej S, Hala S, Naeem R, Nugmanova R, Al-Omari A, Salih S, Al Mutair A, Carr MJ, Hall WW, Arold ST, Pain A (2020) A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. Int J Infect Dis 100: 216-223.
21. Gong Y-N, Tsao K-C, Hsiao M-J, Huang C-G, Huang P-N, Huang P-W, Lee K-M, Liu Y-C, Yang S-L, Kuo R-L, Chen K-F, Liu Y-C, Huang S-Y, Huang H-I, Liu M-T, Yang J-R, Chiu C-H, Yang C-T, Chen G-W, Shih S-R (2020) SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion

mutant and clade possibly associated with infections in Middle East. Emerg Microbes Infect 9: 1457–1466.
22. Sallam M, Ababneh N, Dababseh D, Bakri F, Mahafzah A (2020) Temporal increase in D614G mutation of SARS-CoV-2 in the Middle East and North Africa: phylogenetic and mutation analysis study. medRxiv Preprint 2020.08.24.20176792.
23. Badaoui B, Sadki K, Talbi C, Driss S, Tazi L (2020) Genetic diversity and genomic epidemiology of SARS-COV-2 in Morocco. bioRxiv Preprint 2020.06.23.165902.
24. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, Boerwinkle E, Fu Y-X (2004) Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol 4: 21.
25. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, Zella D, Ippodrino R (2020) Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 18: 179.
26. Eskier D, Karakülah G, Suner A, Oktay Y (2020) RdRp mutations are associated with SARS-CoV-2 genome evolution. PeerJ 8: e9587–e9587.
27. Su YCF, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J, Zhuang Y, Kalimuddin S, Low JGH, Tan CW, Chia WN, Mak TM, Octavia S, Chavatte J-M, Lee RTC, Pada S, Tan SY, Sun L, Yan GZ, Maurer-Stroh S, Mendenhall IH, Leo Y-S, Lye DC, Wang L-F, Smith GJD (2020) Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. MBio 11: e01610-20.
28. Ramírez JD, Muñoz M, Hernández C, Flórez C, Gomez S, Rico A, Pardo L, Barros EC, Paniz-Mondolfi AE (2020) Genetic diversity among SARS-CoV2 strains in South America may impact performance of molecular detection. Pathogens 9: 580.

Corresponding author

Fatimah Saeed Alhamlan, PhD
Department of Infection and Immunity, King Faisal Specialist Hospital and Research Center, PO Box 3354 (MBC-03), Riyadh 11211, Saudi Arabia
Phone: +966 11 442 4365
Fax: 966 11 442 4519
Email: falhamlan@kfshrc.edu.sa

Conflict of interests: No conflict of interests is declared.

Annex – Supplementary Items

Supplementary Table 1. Patients with SARS-CoV-2 stratified by age and study variable.

Variable[1]	Age (mean, standard deviation)	*p*-value
Sex		
Male (n = 368)	47.9 (14.6)	0.97
Female (n = 87)	47.9 (17.9)	
Location/City		
Jeddah (n = 197)	50.7 (14.6)	0.0004*
Madinah (n = 172)	46.4 (15.3)	
Makkah (n = 96)	44.3 (16.2)	
Clade, GISAID Classification		
G (n = 14)	44.4 (16.4)	
GH (n = 253)	46 (15.6)	0.0033*
GR (n = 35)	51.7 (13.7)	
O (n = 142)	51.1 (14.7)	
S (n = 24)	41.5 (12.7)	
Patient Outcome		
Alive (n = 203)	44.1 (13.8)	<.0001*
Deceased (n = 123)	57.8 (12.9)	
Hospitalized (n = 129)	44.5 (15.5)	
Number of Gene Mutations		
High (n > 8) (n = 153)	49.9 (14.7)	0.057
Low (n ≤ 8) (n = 315)	47.1 (15.5)	

*Indicates statistically significant with alpha equal to 0.05; [1] Unknow patient's data was excluded from the analysis.

Supplementary Table 2. The least common mutations detected in SARS-CoV-2, stratified by patient outcome.

Mutation	Patient Outcome, No.				
	Alive	Deceased	Hospitalized	Unknown	Total
Spike A27S	0	0	0	3	3
NSP_L37F	0	1	0	3	4
N_P13L	0	0	0	35	35
NS3_G251V	0	1	0	0	1
NSP2_T85I	1	0	0	2	3
NSP3_T1108K_1	0	0	0	2	2
NSP6_L37F_1	0	1	0	2	3
NS8 V32A	0	0	0	2	2
NSP8 A16V	0	0	0	1	1
NSP12_A97V	0	0	0	2	2
NSP2_T85I	0	0	0	3	3
NSP13_Y541C	1	0	0	0	1