*American Journal of Biochemistry and Biotechnology*

Original Research Paper

# Variants Within over a Hundred Complete COVID-19 Genomes and the Impact on Health Security

**Sayed A.M. Amer**

*Department of Forensic Sciences, College of Criminal Justice, Naif Arab University for Security Sciences, Saudi Arabia*

*Article history*
Received: 04-10-2020
Revised: 21-01-2021
Accepted: 22-01-2021

Email: samer@nauss.edu.sa

**Abstract:** 102 complete COVID-19 genomes have been collected from the viral genomes database to track and characterize novel variants. The data were treated bioinformatically so that 172 variants, with 127 unique and 45 polymorphic variants were found. The 127 unique variants consist of 76 missense, 39 synonymous, 6 non-coding, 5 deletions and 1 insertion. The 45 polymorphic variants consist of 25 missense, 15 synonymous, 4 non-coding and 1 *in-frame-deletion*. Most common variants are 28144T>C (33 missense), 8782C>T (31 synonymous), followed by missense 11083G>T (11 samples), 18060C>T (9 samples) and 26144G>T (7 samples). L3606F, S5932F and L84S are the amino acid changes in the last three common variants. Most variants were found in ORF1ab gene within the region encoded for domains (nsp4 and nsp6) and in the coding ORF8 gene. The variant 28144T>C could be among the main enhancers of viral transmission. There is a tendency for a national specificity of the most recorded variants. The virus outbreak could be between countries or dependent on the place of origin. Reasonable evidence of Chinese origin of the virus could be possible and thus more genomes should be collected and analyzed to understand the origin and the reason for its outbreak. This could support human health security by either finding out suitable vaccines or managing health precautionary measures.

**Keywords:** Coronavirus, Genome, Variants, Mutation, Health Security

## Introduction

Since the virus is like SARS-CoVs, the International Committee on Taxonomy of Viruses (ICTV) termed the newly discovered Coronavirus (2019-nCoV) SARS-CoV-2 (Cascella *et al*., 2020). It was first reported in Wuhan, the largest metropolitan area in China on December 31st, 2019. Therefore, the (WHO, 2020) acronymized this 2019 disease as "COVID-19". CoVs can infect different species including birds, mice, bats, livestock and humans (Wang *et al*., 2006; Ge *et al*., 2013; Chen and Guo, 2016), with a serious infection in humans, causing pneumonia.

Systematically, CoVs belong to the subfamily Orthocoronavirinae of the family Coronaviridae and the order Nidovirales. Coronavirinae contains four genera: Alphacoronavirus, Betacoronavirus (4 subgenera) (Chan *et al*., 2013), Gammacoronavirus and Deltacoronavirus (Chen *et al*., 2020). Bats and rodents are the probable source of alpha and Betacoronaviruses, whereas birds could be the source of Gamma and delta viruses (Cascella *et al*., 2020). Over 210,000 COVID-19 complete genomes have so far been sequenced with an approximate genome length of 30000 bases. As fast as the virus spreads, more variants accumulate with the possibility of further emerging virulent strains. The virus genome is found in a single-stranded positive-sense RNA (+ssRNA) (Perlman and Netland, 2009) acquiring 5′-cap structure and 3′-poly-A tail (Chen *et al*., 2020). As a severe global health threat, CoVs outbreaks are, most probably, unavoidable in the future. There is an urgent need, thus, to identify the possible variants in the 102 sequenced genomes that might support in understanding the reasons for the virus outbreak and in producing an effective therapy and vaccine against the virus.

## Materials and Methods

The COVID-19 page (https://bigd.big.ac.cn/ncov) available in China's National Genomics Data Center (NGDC) was accessed on the 27th of September 2020.

From 2019-nCoV genomes available on this page and in three Genbank databases, 102 publicly complete genomes have been collected and are listed in Table 1. As has been previously used (Koyama *et al*., 2020; Matsuda *et al*., 2020), the two longest and identical NC_045512 and MN908947 sequences (29903 bp) were used as reference genomes to coordinate the differences in the ribosomal slippage of the 102 used genomes. The genome data were checked for the accuracy of the alignments and the aligned data could be obtained from the author upon request. Since genomes acquire differences in their start and end points, their lengths have been adjusted to the lengths of NC_045512 and MN908947 and their variants were numbered according to the positions of these two genomes. All genomes were first aligned to each other using BioEdit Sequence Alignment Editor (Hall *et al*., 2011) and the FASTA file of the aligned data was executed to MacClade v.4.10 program (Maddison and Maddison, 2002) and manually aligned to the reference genomes. Using MacClade v.4.10, mutations were recorded and data were transferred to the Paup v. 4.0b10 (Swofford, 2002) program for phylogenetic analysis.

After deleting ambiguous, gap-containing sites and N or mixed bases sites, the remaining 29532 sites were analyzed by Maximum-Parsimony (MP), Neighbor-Joining (NJ) and Maximum-Likelihood (ML) methods in Paup v. 4.0b10. For MP, heuristic searches of 10 random stepwise additions were conducted by Tree Bisection Reconnection (TBR) branch swapping and 1000 bootstrap replications. For NJ analysis (Saitou and Nei, 1987; Tamura and Nei, 1993) distance option and 1000 bootstrap replications were used. As the MT226610 sequence was rapidly evolved, it was used as an outgroup. For ML, heuristic searches by axis additions and Nearest-Neighbor Interchange (NNI) branch-swapping were adjusted. Other conditions for the ML analysis like gamma shape parameter of 1.5214 and 4 rate categories were also adjusted. The substitution rate matrix of the data was as follow: R(a) = 0.1571, R(b) = 0.7500, R(c) = 0.2121, R(d) = 0.6049, R(e) = 2.6308 and R(f) = 1.00. Likelihood settings from best-fit model (GTR+I+G) were selected by AIC in Modeltest Version 3.06.

**Table 1:** Organization of the reference + ssRNA genome used in this study. Nsp = non structural protein; ORF = Open Reading Frame

| Gene | Position | Product | Notes |
|---|---|---|---|
| 5'UTR | 1-265 | 5'UTR | cap |
| Orf1ab | 266-805 | nsp1 | Leader protein |
| | 806-2719 | nsp2 | |
| | 2720-8554 | nsp3 | |
| | 8555-10054 | nsp4 | Transmembrane domain 2 |
| | 10055-10972 | nsp5 | 3C-like proteinase |
| | 10973-11842 | nsp6 | Putative transmembrane domain |
| | 11843-12901 | nsp7 | |
| | 12092-12685 | nsp8 | |
| | 12686-13042 | nsp9 | ssRNA-binding protein |
| | 13025-13441 | nsp10 | Growth-factor-like protein |
| | 13442-13480 | nsp11 | |
| | 13476-13503 | Stem_loop | Frameshifting stimulation element |
| | 13488-13542 | Stem_loop | Frameshifting stimulation element |
| | 13442-13468, 13468-16235 | nsp12 | RNA-dependent RNA polymerase |
| | 16237-18093 | nsp13 | Helicase |
| | 18040-19620 | nsp14 | 3'-to-5' exonuclease |
| | 19621-20658 | nsp15 | endoRNAse |
| | 20659-21552 | nsp16 | O-ribose methyltransferase |
| S | 21563-25384 | spike glycoprotein | Surface glycoprotein |
| ORF3a | 25393-26220 | ORF3a protein | |
| E | 26245-26472 | envelope protein | |
| M | 26523-27191 | membrane glycoprotein | |
| ORF6 | 27202-27387 | ORF6 protein | |
| ORF7a | 27394-27759 | ORF7a protein | |
| ORF7b | 27756-27887 | ORF7b | |
| ORF8 | 27894-28259 | ORF8 protein | |
| N | 28274-29533 | nucleocapsid phosphoprotein | ORF9 protein |
| ORF10 | 29558-29674 | ORF10 protein | |
| ORF10 | 29609-29644 | -- | 3'UTR pseudoknot stem-loop 1 |
| ORF10 | 29629-29657 | -- | 3' UTR pseudoknot stem-loop 2 |
| 3'UTR | 29675-29903 | 3'UTR | tail |

## Results and Discussion

The COVID-19 genome is approximately 30,000 bp. It contains within its 5'-proximal two-thirds, nonstructural protein coding regions (nsp1-nsp16; replicase genes) encoded on orf1ab gene. Within the 3'-proximal one-third of the genome, structural proteins and nonessential accessory protein coding regions are encoded (ORF3-ORF10, S, E, M, N) (Paul, 2006). The 5' UTR (265 b) comprises the genome cap and the 3' UTR (~300 b) forms the tail (Table 1). The recorded variants were counted between positions 33 and 29830 to avoid base bias. The mixed bases were not considered in the analysis. 172 total variants are found, 127 unique and 45 polymorphic (Table 2). The unique variants (Table 3) are 121 substitutions (76 missense, 39 synonymous and 6 non-coding), 5 deletions (2 *in-frame-deletions*, 2 *out-of-frame-deletions* and 1 non-coding) and 1 insertion. The polymorphic variants (Table 4) are 45 substitutions (25 missense, 15 synonymous, 4 non-coding and 1 *in-frame-deletion*) (Fig. 1).

In agreement with previous studies (Koyama *et al*., 2020; Matsuda *et al*., 2020), most common variants were 8782C>T (ORF1ab) and 28144T>C (ORF8) in 31 samples followed by 11083G>T (nsp6) in 11 samples and 18060C>T (nsp14) in 9 samples. The occurrences of 8782C>T and 28144T>C concur and most of the other common variants are subsets of these most common ones. 8782C>T is synonymous; however, 11083G>T (L3606F) and 18060C>T (S5932F) and 28144T>C (L84S) are missense causing amino acid changes in nsp6, nsp14 and ORF8, respectively. The variant 28144 T>C which exhibited an amino acid change in ORF8 protein from Leucine to Sereine, was indicated in the polypeptide involved in enhancing virus transition from bat to human (Chan *et al*., 2020; Nguyen *et al*., 2020). As this variant was recorded in thirteen North Americans, ten central Asians (Chinese including Wuhan residents), three Japanese, one Taiwanese, one Indian and one Spanish, it could be among the main enhancers of viral transmission.

**Table 2:** Statistics of the 172 recorded mutations in the 102 COVID-19 genomes

| | Substitution | | | Deletion | | | |
|---|---|---|---|---|---|---|---|
| Variants | Missense | Synonymous | Non-coding | In-frame-deletion | Out-of-frame deletion | Non-coding | Insertion |
| Unique | 76 | 39 | 6 | 2 | 2 | 1 | 1 |
| Polymorphic | 25 | 15 | 4 | 1 | - | - | - |

**Table 3:** Unique monomorphic variants (substitution, deletion and insertion), their positions, number of cases and country distribution found in the 102 CoVs genomes

| # | Variant | Type | Amino acid change | Accession number, country |
|---|---|---|---|---|
| 1 | 35A>T | Non-coding | | MT163716, USA |
| 2 | 36C>T | Non-coding | | MT163716, USA |
| 3 | 75C>A | Non-coding | | MT049951, CHN |
| 4 | 186C>T | Non-coding | | MT093631, CHN |
| 5 | 382A>T | Synonymous | | LC521925, JP |
| 6 | 359-382 | In-frame-deletion | G32_V39 | LC521925, JP |
| 7 | 490T>A | Synonymous | | MT044257, USA |
| 8 | 514T>C | Synonymous | | MT188340, USA |
| 9 | 654G>A | Synonymous | | MT123293, CHN |
| 10 | 686-694 | In-frame-deletion | K141_F143 | MT044258, USA |
| 11 | 1102C>T | Synonymous | | MT188339, CHN |
| 12 | 1348C>T | Synonymous | | GWHACDD01000001, PAK |
| 13 | 1385C>T | Missense | H374Y | MT159720, USA |
| 14 | 1397G>A | Synonymous | | GWHACDD01000001, PAK |
| 15 | 1548G>A | Missense | S428N | MN994467, USA |
| 16 | 1912C>T | Synonymous | | LC521925, JP |
| 17 | 2091C>T | Missense | T609I | MT027064, USA |
| 18 | 2269A>T | Synonymous | | MT066156, ITA |
| 19 | 2277T>C | Missense | I671T | MT012098, IND |
| 20 | 2446T>C | Synonymous | | MT163716, USA |
| 21 | 2717G>A | Missense | G818S | MT093571, SWE |
| 22 | 2971G>T | Missense | M902I | MT049951, CHN |
| 23 | 3037C>T | Synonymous | | MT192765, USA |
| 24 | 3177C>T | Missense | P971L | MT044257, USA |
| 25 | 3259G>T | Missense | T998G | MT159707, USA |
| 26 | 3411C>T | Missense | E1049V | MT163716, USA |
| 27 | 3738C>T | Missense | P1158L | MT159705, USA |
| 28 | 3792C>T | Missense | E1176V | LC522973, JP |

**Table 3:** Continue

| | | | | |
|---|---|---|---|---|
| 29 | 4288G>T | Synonymous | | MT226610, CHN |
| 30 | 4307A>C | Missense | G1348T | MT226610, CHN |
| 31 | 5572G>T | Synonymous | | MT163716, USA |
| 32 | 5784C>T | Synonymous | | MT152824, USA |
| 33 | 5845A>T | Missense | K1860N | MT159715, USA |
| 34 | 6031C>T | Synonymous | | MT039890, SKorea |
| 35 | 6035A>G | Missense | S1924K | MT188341, USA |
| 36 | 6636C>T | Missense | T2124I | MT159712, USA |
| 37 | 6695C>T | Missense | P2144S | MT012098, IND |
| 38 | 6968C>A | Missense | L2235M | CNA0007332, Wuhan |
| 39 | 6996T>C | Missense | I2244T | MT123293, CHN |
| 40 | 7479A>G | Missense | N2405S | MT226610, CHN |
| 41 | 9034A>G | Synonymous | | MT066176, TWN |
| 42 | 9157T>C | Missense | P2965L | MT184910, USA |
| 43 | 9274A>G | Synonymous | | MT093571, SWE |
| 44 | 9474C>T | Missense | A3070V | MT159706, USA |
| 45 | 9491C>T | Missense | H3076Y | MT066176, TWN |
| 46 | 9561C>T | Missense | S3099L | MN975262, CHN |
| 47 | 9924C>T | Missense | A3220V | MT118835, USA |
| 48 | 10036C>T | Synonymous | | MT159708, USA |
| 49 | 10507C>T | Synonymous | | MT184911, USA |
| 50 | 11207G>C | Missense | F3648V | MT226610, CHN |
| 51 | 11233T>G | Synonymous | | MT226610, CHN |
| 52 | 11557G>T | Missense | E3764D | LC522972, JP |
| 53 | 11750C>T | Missense | L3829F | MT159712, USA |
| 54 | 11764T>G | Missense | N3833K | CNA0007332, Wuhan |
| 55 | 11956C>T | Synonymous | | MT159712, USA |
| 56 | 12041G>C | Missense | D3926H | MT226610, CHN |
| 57 | 12115C>T | Synonymous | | MT039890, SKorea |
| 58 | 12160G>C | Synonymous | | MT226610, CHN |
| 59 | 12202G>C | Synonymous | | MT226610, CHN |
| 60 | 12208G>T | Synonymous | | MT226610, CHN |
| 61 | 12355G>C | Missense | T4030K | MT226610, CHN |
| 62 | 12378G>A | Missense | R4038K | MT226610, CHN |
| 63 | 12464G>T | Missense | A4067S | MT226610, CHN |
| 64 | 12467 G>T | Missense | A4068S | MT226610, CHN |
| 65 | 12491C>T | Synonymous | | MT226610, CHN |
| 66 | 12514G>C | Synonymous | | MT226610, CHN |
| 67 | 12534C>T | Missense | T4090I | MT123292, CHN |
| 68 | 12572G>T | Missense | D4103Y | MT226610, CHN |
| 69 | 12578G>T | Missense | D4015Y | MT226610, CHN |
| 70 | 12582G>T | Missense | S4106I | MT226610, CHN |
| 71 | 12600G>A | Missense | S4112N | MT226610, CHN |
| 72 | 12660G>C | Missense | T4132R | MT226610, CHN |
| 73 | 12685G>C | Missense | T4140K | MT226610, CHN |
| 74 | 12773G>T | Missense | F4170S | MT226610, CHN |
| 75 | 12793G>T | Missense | K4176N | MT226610, CHN |
| 76 | 13072C>T | Synonymous | | MT123292, CHN |
| 77 | 13225C>G | Synonymous | | MT093571, SWE |
| 78 | 13226T>C | Missense | F4321L | MT093571, SWE |
| 79 | 14657C>T | Missense | L4798F | MT012098, IND |
| 80 | 15597T>C | Missense | M5111T | MT039890, SKorea |
| 81 | 15607T>C | Synonymous | | MN975262, CHN |
| 82 | 16467A>G | Missense | H5401R | MT188341, USA |
| 83 | 17000C>T | Missense | H5579Y | MN994468, USA |
| 84 | 17247T>C | Missense | V5661A | MT126808, BRA |
| 85 | 17376A>G | Missense | Q5704R | MT093571, SWE |
| 86 | 18512C>T | Missense | L6083F | LC521925, JP |
| 87 | 18603T>T | Missense | M6113T | MT106054, USA |
| 88 | 18814C>T | Synonymous | | MT192765, USA |

**Table 3:** Continue

| | | | | |
|---|---|---|---|---|
| 89 | 18975T>A | Missense | F6237Y | MT106054, USA |
| 90 | 19065T>C | Missense | L6267P | MT007544, AUSTR |
| 91 | 19175A>C | Missense | M6304L | MT106054, USA |
| 92 | 19610C>T | Missense | T6449R | MT123291, CHN |
| 93 | 20281T>C | Synonymous | | MT163719, USA |
| 94 | 20299-20301 | Out-of-frame deletion | N6678N (nsp15) | MT039887, USA |
| 95 | 20670G>A | Missense | R6802H | NMDC60013002-10, Wuhan |
| 96 | 20679G>A | Missense | R6805Q | NMDC60013002-10, Wuhan |
| 97 | 20936C>T | Missense | R6891C | MT039890, SKorea |
| 98 | 20980G>C | Missense | Q6905M | MT226610, CHN |
| 99 | 21147T>C | Missense | L6961S | MT188339, CHN |
| 100 | 21647T>A | Missense | Y23N | MT049951, CHN |
| 101 | 21707C>T | Missense | H44Y | MT027064, USA |
| 102 | 21784T>A | Missense | F152L | MT226610, CHN |
| 103 | 21386-21388 | Insertion | S7041 | MT188341, USA |
| 104 | 21997-21999 | Out-of-frame deletion | Y144Y (S gene) | MT012098, IND |
| 105 | 22033C>A | Missense | G176V | MT159716, USA |
| 106 | 22104G>T | Synonymous | | MT184910, USA |
| 107 | 22224C>G | Missense | S216F | MT039890, SKorea |
| 108 | 22303T>G | Missense | S242R | MT007544, AUSTRALIA |
| 109 | 22432C>T | Synonymous | | MT049951, CHN |
| 110 | 22785G>T | Missense | R403M | MT012098, IND |
| 111 | 23185C>T | Synonymous | | MT188341, USA |
| 112 | 23403A>G | Synonymous | | MT192765, USA |
| 113 | 23955T>G | Missense | F792Y | MT093571, SWE |
| 114 | 25775G>T | Missense | C120L | MT039890, SKorea |
| 115 | 26354T>A | Missense | L22G | MT039890, SKorea |
| 116 | 27493C>T | Missense | P61S | NMDC60013002-09, Wuhan |
| 117 | 27925C>T | Missense | T18M | MT106054, USA |
| 118 | 28253C>T | Synonymous | | NMDC60013002-09, Wuhan |
| 119 | 28409C>T | Missense | P47S | MT159718, USA |
| 120 | 28792A>T | Synonymous | | MN994467, USA |
| 121 | 28878G>A | Missense | S203N | MT106052, USA |
| 122 | 28916G>A | Missense | G216S | MT188339, USA |
| 123 | 29230C>T | Synonymous | | MT159720, USA |
| 124 | 29301A>T | Missense | D344V | MT135043, CHN |
| 125 | 29705G>T | Non-coding | | LC522973, JP |
| 126 | 29742G>A | Non-coding | | MT106052, USA |
| 127 | 29750-29759 | Non-coding del | | MT007544 (AUSTRALIA) |

**Table 4:** Polymorphic variants (substitutions), their positions, number of cases and country distribution found in the 102 CoVs genomes

| # | variant | Type | aa change | gene | Accession number (country) |
|---|---|---|---|---|---|
| 1 | 241C>T | Non-coding | | 5' UTR | GWHACDD01000001(PAK), MT192765(USA) |
| 2 | 254C>T | Non-coding | | 5' UTR | MT184910, MT184908 (USA) |
| 3 | 508-522 | In-frame-deletion | G82_V86 | orf1ab | MT044258, MT159716 (USA) |
| 4 | 614G>A | Missense | A117T | | MT027062, MT027063 (USA) |
| 5 | 1691A>G | Missense | I476V | | MT027063 (USA), MT050493 (IND) |
| 6 | 2662C>T | Synonymous | | | LC522973-LC522975 (JP) |
| 7 | 3099C>T | Missense | D954I | | MT159717, MT184912 (USA) |
| 8 | 4402T>C | Synonymous | | | MT135041-MT135044 (CHN) |
| 9 | 5062G>T | Missense | L1599A | | MT135041-MT135044 (CHN) |
| 10 | 5084A>G | Missense | I1607V | | MT027062, MT027063 (USA) |
| 11 | 6501C>T | Missense | P2079L | | MT027063 (USA), MT050493 (IND) |
| 12 | 6819G>T | Missense | S2185I | | MT123293, MT123291(CHN) |
| 13 | 8782C>T | Synonymous | | | MN938384, MN975262, MT049951, MT123292, MT226610, MT135041-4, GWHABKI000000004(CHN), MN985325, MN997409, MT020880, MT020881, MT044257, MT106052, MT106054, MT152824, MT163717-19, MT188339, MT188341 (USA), MT066175 (TAW), LC522973-75(JP), MT050493, MT050493 (IND), MT198651, MT198652(ESP) |
| 14 | 9477T>A | Missense | F3071Y | | MT198651, MT198652(ESP) |
| 15 | 10232C>T | Missense | R3323C | | MT192772, MT1927739 (Vietnam) |
| 16 | 11083G>T | Missense | L3606F | | LC528232-33 (JP), MT126808 (BRA), MT163716, MN997409, MT184910-13 (USA), MT226610 (CHN), GWHACDD01000001 (PAK) |

**Table 4:** Continue

| | | | | | |
|---|---|---|---|---|---|
| 17 | 11410G>A | Synonymous | | | MT159722, MT159705 (USA) |
| 18 | 14805C>T | Missense | T4847I | | MT126808(BRA), MT163716(USA), MT198651, MT198652(ESP) |
| 19 | 15324C>T | Missense | T5020M | | LC522972, LC529905 (JP), MT123290 (CHN) |
| 20 | 16877C>T | Missense | T5538R | | MT050943, MT050493 (IND) |
| 21 | 17373C>T | Missense | P5703L | | MT050493(IND), MT123290, MT123293(CHN), MT0398871 (USA) |
| 22 | 17747C>T | Synonymous | | | MT152824, MT163717-19, MT188339-40 (USA) |
| 23 | 17858A>G | Missense | M5865V | | MT152824, MT163717-19, MT188339-40 (USA) |
| 24 | 18060C>T | Missense | S5932F | | MN985325, MT163716, MT152824, MT163717-19, MT188339-40 (USA), MT135041 (CHN) |
| 25 | 21386C>T | Missense | L7041F | | MT188339-40 (USA) |
| 26 | 24034C>T | Synonymous | | S | MT044257, MN994467, (USA), MT066175(TAW) |
| 27 | 24325A>G | Synonymous | | | MT106053 (USA), NMDC60013002-06 (Wuhan) |
| 28 | 24351C>T | Missense | A930V | | MT050493, MT050943 (IND) |
| 29 | 25810C>G | Missense | L140V | ORF3a | LC522972, LC529905 (JP) |
| 30 | 25979G>T | Missense | G196V | | MT198651, MT198652(ESP) |
| 31 | 26144G>T | Missense | G251V | | MN994468, MT163716 (USA), MT007544 (AUSTRALIA), MT039890 (South K), MT093571 (SWE), MT126808 (BRA), MT066156 (ITA) |
| 32 | 26326C>T | Synonymous | | E | MT159722, MT159705 (USA) |
| 33 | 26729T>C | Synonymous | | M | MT044257, MN994467 (USA) |
| 34 | 28077G>C | Missense | V62L | ORF8 | MT044257, MN994467 (USA) |
| 35 | 28144T>C | Missense | L84S | | MN938384, MN975262, MT049951, MT123292, MT226610, MT135041-4, GWHABKI000000004(CHN), MN985325, MN994467, MN997409, MT020880, MT020881, MT044257, MT106052, MT106054, MT152824, MT163717-19, MT188339, MT188341, MN988713 (USA), MT066175 (TAW), LC522973-75(JP), MT050943, MT050493(IND), MT198651, MT198652(ESP) |
| 36 | 28378G>T | Synonymous | | N | MT159717, MT184911 (USA) |
| 37 | 28657C>T | Synonymous | | | MT198651, MT198652(ESP) |
| 38 | 28854C>T | Missense | S198L | | MT027062-63 (USA) |
| 39 | 28863C>T | Synonymous | | | MT198651, MT198652(ESP) |
| 40 | 29095C>T | Synonymous | | | MN938384, MN975262 (CHN), MN997409, MT106054 (USA), LC522973-75 (JP) |
| 41 | 29303C>T | Missense | P344S | | LC522972, LC529905 (JP), MT123290 (CHN) |
| 42 | 29527G>A | Synonymous | | | MT123291, MT123293 (CHN) |
| 43 | 29635C>T | Synonymous | | ORF10 | MT159709, MT159720 (USA), LC528233 (JP) |
| 44 | 29736G>T | Non-coding | | 3' UTR | MT184908, MT184910, MT159718 (USA) |
| 45 | 29751G>C | Non-coding | | | MT184908, MT184910 (USA) |

**Table 5:** Length, country and accession number of COVID-19 genome sequences used in this study

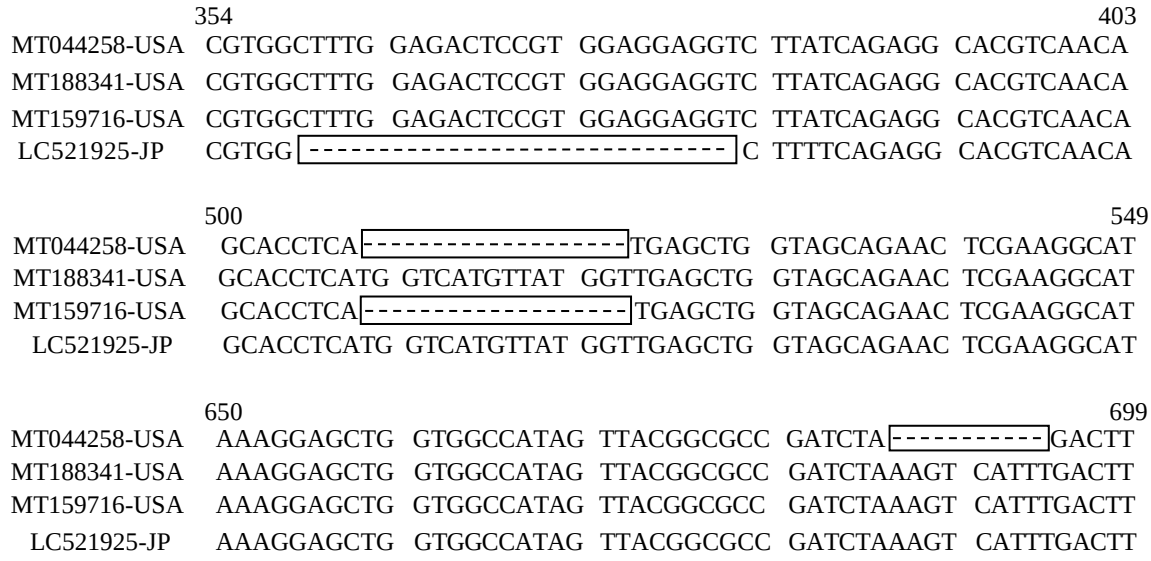
| # | Country | Length | Database | Accession number |
|---|---|---|---|---|
| 1 | Wuhan | 29903 | Genbank | MN908947 |
| 2 | China | 29838 | Genbank | MN938384 |
| 3 | China | 29891 | Genbank | MN975262 |
| 4 | USA | 29882 | Genbank | MN985325 |
| 5 | USA | 29882 | Genbank | MN988713 |
| 6 | USA | 29882 | Genbank | MN994467 |
| 7 | USA | 29883 | Genbank | MN994468 |
| 8 | USA | 29882 | Genbank | MN997409 |
| 9 | Finland | 29806 | Genbank | MT020781 |
| 10 | USA | 29882 | Genbank | MT020880 |
| 11 | USA | 29882 | Genbank | MT020881 |
| 12 | Japan | 29848 | Genbank | LC521925 |
| 13 | USA | 29882 | Genbank | MT027062 |
| 14 | USA | 29882 | Genbank | MT027063 |
| 15 | USA | 29882 | Genbank | MT027064 |
| 16 | Taiwan | 29870 | Genbank | MT066175 |
| 17 | Japan | 29878 | Genbank | LC522973 |
| 18 | Japan | 29878 | Genbank | LC522974 |
| 19 | Japan | 29878 | Genbank | LC522975 |
| 20 | Japan | 29878 | Genbank | LC522972 |
| 21 | USA | 29879 | Genbank | MT039887 |
| 22 | USA | 29858 | Genbank | MT044258 |
| 23 | USA | 29882 | Genbank | MT044257 |
| 24 | South Korea | 29903 | Genbank | MT039890 |
| 25 | China | 29903 | Genbank | MT049951 |
| 26 | Taiwan | 29870 | Genbank | MT066176 |
| 27 | Nepal | 29811 | Genbank | MT072688 |
| 28 | China | 29860 | Genbank | MT093631 |

**Table 5:** Continue

| | | | | |
|---|---|---|---|---|
| 29 | Sweden | 29886 | Genbank | MT093571 |
| 30 | USA | 29882 | Genbank | MT106052 |
| 31 | USA | 29882 | Genbank | MT106053 |
| 32 | USA | 29882 | Genbank | MT106054 |
| 33 | USA | 29882 | Genbank | MT118835 |
| 34 | China | 29882 | Genbank | MT123291 |
| 35 | China | 29891 | Genbank | MT123290 |
| 36 | Japan | 29902 | Genbank | LC528233 |
| 37 | Japan | 29902 | Genbank | LC528232 |
| 38 | USA | 29878 | Genbank | MT152824 |
| 39 | Brazil | 29876 | Genbank | MT126808 |
| 40 | USA | 29903 | Genbank | MT163716 |
| 41 | China | 29903 | Genbank | MT135041 |
| 42 | China | 29903 | Genbank | MT135042 |
| 43 | China | 29903 | Genbank | MT135043 |
| 44 | China | 29903 | Genbank | MT135044 |
| 45 | USA | 29897 | Genbank | MT163717 |
| 46 | USA | 29903 | Genbank | MT163718 |
| 47 | USA | 29903 | Genbank | MT163719 |
| 48 | India | 29851 | Genbank | MT050493 |
| 49 | India | 29854 | Genbank | MT012098 |
| 50 | USA | 29882 | Genbank | MT159717 |
| 51 | USA | 29882 | Genbank | MT159718 |
| 52 | USA | 29882 | Genbank | MT159719 |
| 53 | USA | 29882 | Genbank | MT159720 |
| 54 | USA | 29882 | Genbank | MT159721 |
| 55 | USA | 29882 | Genbank | MT159722 |
| 56 | USA | 29882 | Genbank | MT159705 |
| 57 | USA | 29882 | Genbank | MT159706 |
| 58 | USA | 29882 | Genbank | MT159710 |
| 59 | USA | 29882 | Genbank | MT159707 |
| 60 | USA | 29882 | Genbank | MT159708 |
| 61 | USA | 29882 | Genbank | MT159709 |
| 62 | USA | 29882 | Genbank | MT159711 |
| 63 | USA | 29882 | Genbank | MT159712 |
| 64 | USA | 29882 | Genbank | MT159713 |
| 65 | USA | 29882 | Genbank | MT159714 |
| 66 | USA | 29882 | Genbank | MT159715 |
| 67 | USA | 29882 | Genbank | MT159716 |
| 68 | China | 29923 | Genbank | MT123292 |
| 69 | China | 29871 | Genbank | MT123293 |
| 70 | Italy | 29867 | Genbank | MT066156 |
| 71 | Japan | 29903 | Genbank | LC529905 |
| 72 | USA | 29882 | Genbank | MT184907 |
| 73 | USA | 29880 | Genbank | MT184908 |
| 74 | USA | 29882 | Genbank | MT184909 |
| 75 | USA | 29882 | Genbank | MT184910 |
| 76 | USA | 29882 | Genbank | MT184911 |
| 77 | USA | 29882 | Genbank | MT184912 |
| 78 | USA | 29882 | Genbank | MT184913 |
| 79 | USA | 29783 | Genbank | MT188339 |
| 80 | USA | 29845 | Genbank | MT188340 |
| 81 | USA | 29835 | Genbank | MT188341 |
| 82 | USA | 29829 | Genbank | MT192765 |
| 83 | Taiwan | 29862 | Genbank | MT192759 |
| 84 | Vietnam | 29891 | Genbank | MT192772 |
| 85 | Vietnam | 29891 | Genbank | MT192773 |
| 86 | Spain | 29611 | Genbank | MT198651 |
| 87 | Spain | 29782 | Genbank | MT198652 |
| 88 | China | 29899 | Genbank | MT226610 |
| 89 | Pakistan | 29836 | Genome Warehouse | GWHACDD01000001 |
| 90 | Wuhan | 29899 | Genome Warehouse | GWHABKF0000000001 |

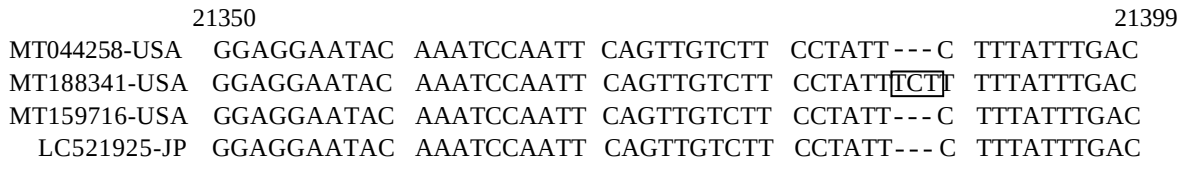
**Table 5:** Continue

| | | | | |
|---|---|---|---|---|
| 91 | Wuhan | 29889 | Genome Warehouse | GWHABKF0000000003 |
| 92 | Wuhan | 29890 | Genome Warehouse | GWHABKF0000000004 |
| 93 | Wuhan | 29891 | NMDC | NMDC60013002-06 |
| 94 | Wuhan | 29890 | NMDC | NMDC60013002-07 |
| 95 | Wuhan | 29891 | NMDC | NMDC60013002-08 |
| 96 | Wuhan | 29896 | NMDC | NMDC60013002-09 |
| 97 | Wuhan | 29891 | NMDC | NMDC60013002-10 |
| 98 | Australia | 29893 | Genbank | MT007544 |
| 99 | Wuhan | 29903 | Genbank | NC_045512 |
| 100 | China | 29871 | Genbank | MN996530 |
| 101 | China | 29894 | Genbank | MN996528 |
| 102 | India | 29874 | Genbank | MT050943 |




(A) 3 *In-frame deletions* at positions: $C_{359}$ - $T_{382}$, $T_{508}$ - $T_{522}$ & $A_{686}$ - $T_{694}$


(B) 2 *Out-of-frame deletions* in MT039887-USA & MT012098-IND




(C) 1 Insertion in MT188341-USA

**Fig. 1:** *In -frame, out-of- frame* deletions and an insertion in USA, Japanese and Indian variants

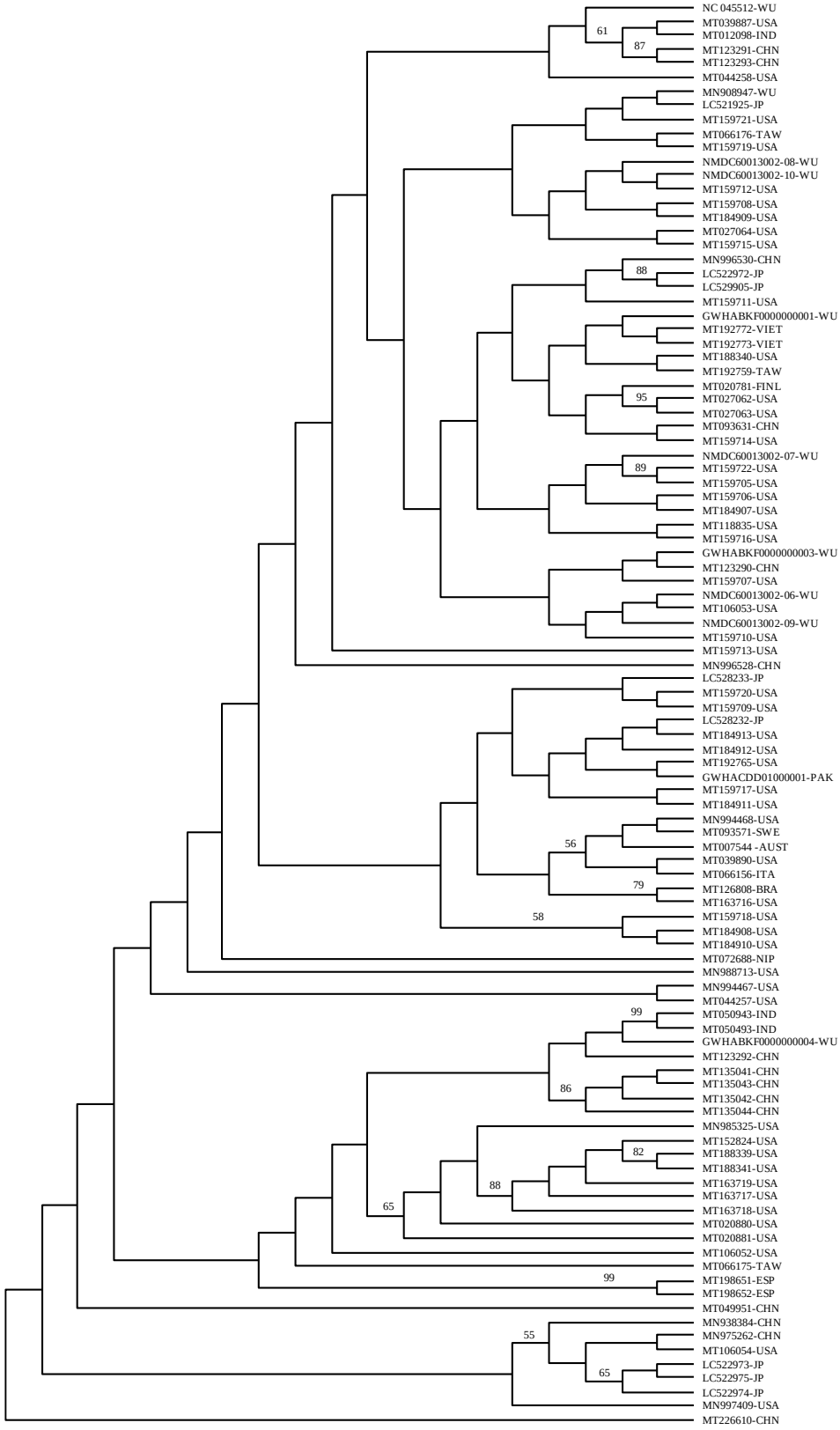



**Fig. 2:** NJ tree constructed by using 29532 sites of the collected genomes. Bootstrap values are shown at nodes whenever they are above 50%


**Fig. 3:** ML phylogenetic tree constructed by the dataset using the modeltest GTR + I + G

For the 101 missense variants, 80 variants are found in the longest ORF1ab gene distributed in the cleaved nonstructural proteins (NSP1-NSP16). However, more variants are found in the structural protein genes (S, ORF3a and N). MT226610-CHN was the fastest evolving substrain as it exhibited 29 substitutions. One of the *out-of-frame-deletions* is found close to 3'end of nsp15 protein of an Indian substrain and the other one is found close to 5'end of S protein of the USA strain (Fig. 1). The first mutation probably did not alter the O-ribose methyltransferase (nsp16) since it is located at the end of the gene, while the second could alter the post-translational spike, glycoprotein. This mutation may increase disease susceptibility (Zimmerman *et al*., 1997) or stop protein function indicating that it is not necessary for efficient viral transmission. It is not known that S deletion enhances virulence or transmission rates of the virus and it is not known whether the strain acquiring this deletion could successfully transmit to a new host (Assiri *et al*., 2016).

Fortunately, this study collected various COVID-19 genomes from the same place of origin as shown in Table 5 (51 genomes from USA, 25 from China (including 10 from Wuhan) and 8 from Japan). This supports that the novel mutations found herein could reflect the diversity of the place of origin rather being acquired during spreading of the infection (Matsuda *et al*., 2020). It is therefore an indication that stopping virus outbreak is possible in the short-term future. However, the constructed tree (Fig. 2) indicated that viruses from the same country did not form a single group, which suggests that CoVs-19 were introduced to each country several times (Koyama *et al*., 2020; Matsuda *et al*., 2020) and it, thus, may be difficult to follow the virus origin. The phylogeny of the maximum-likelihood analysis (Fig. 3) indicated a possible transmission scenario of the virus. The tree referred to Chinese origin of COVID-19 and showed its transmissions to USA, Spain, Japan and India.

Researchers sequenced a lot of SARS-CoV-2 genomes and shared results during the pandemic. The sequenced data allowed public health officials to evaluate the relevant epidemiological parameters such as the reproductive number and virus introduction into new regions. Knowing the possibilities for the outbreak is still managing health precautionary measures which could be conducted in daily life (Hopkins, 2020). Understanding genetic framework of COVID-19 genome enhances WHO's ability to analyze the risk of the virus introduction into countries and define the response actions and prioritization of resources, as well as the possible capacity to manage the virus outbreak. The implementation of action plans for health security is occurring globally with varied progress rates (Samhouri *et al*., 2018) and is actively supported by WHO to enhance operational readiness for the virus in countries (Al-Mandhari *et al*., 2020).

## Conclusion

In conclusion, the virus is still considered a threat to human health security as there is lack of knowledge about the origin and the reasons for its outbreak. Chinese origin could be possible. Two debates about the virus outbreak are either the diversity of the place of origin or spreading the infection through individuals' movements between countries. Emergence of new variants by releasing more genomes could help in clarifying the virus origin, the reasons of its outbreak and the development of vaccines or effective precautions.

## Acknowledgment

The author would like to express his deep thanks to the Vice Presidency for Scientific Research at Naif Arab University for Security Sciences for their kind encouragement of this work.

## Funding Information

This work was supported by Security Research Center at Naif Arab University for Security Sciences.

## Ethics

This article is original and contains unpublished material.

## Conflict of Interest

The author declares that he has no conflict of interest. No ethical approval for this study is needed since it depended on the data deposited in the Genbank database.

## References

Al-Mandhari, A., Samhouri, D., Abubakar, A., & Brennan, R. (2020). Coronavirus Disease 2019 outbreak: preparedness and readiness of countries in the Eastern Mediterranean Region. https://coronavirus. 1science. com/item/6166e6044a2114454e136505bc74865c1c52e8da.

Assiri, A.M., Biggs, H.M., Abedi, G.R., Lu, X., Bin Saeed, A., Abdalla, O., Mohammed, M., Al-Abdely, H.M., Algarni, H.S., Alhakeem, R.F., & Almasri, M.M. (2016), May. Increase in Middle East respiratory syndrome-coronavirus cases in Saudi Arabia linked to hospital outbreak with continued circulation of recombinant virus, July 1-August 31, 2015. In Open forum Infect Diseases (Vol. 3, No. 3), Oxford University Press.

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). Features, evaluation and treatment coronavirus (COVID-19). In Statpearls [internet], StatPearls Publishing.

Chan, J. F. W., Kok, K. H., Zhu, Z., Chu, H., To, K. K. W., Yuan, S., & Yuen, K. Y. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microb & Infect, 9(1), 221-236.

Chan, J. F. W., To, K. K. W., Tse, H., Jin, D. Y., & Yuen, K. Y. (2013). Interspecies transmission and emergence of novel viruses: lessons from bats and birds. Trends In Microbiol, 21(10), 544-555.

Chen, Y., & Guo, D. (2016). Molecular mechanisms of coronavirus RNA capping and methylation. Virol Sin, 31(1), 3-11.

Chen, Y., Liu, Q., & Guo, D. (2020). Emerging coronaviruses: genome structure, replication and pathogenesis. J Med Virol, 92(4), 418-423.

Ge, X.Y., Li, J.L., Yang, X.L., Chmura, A.A., Zhu, G., Epstein, J.H., Mazet, J.K., Hu, B., Zhang, W., Peng, C., & Zhang, Y.J. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature, 503(7477), 535-538.

Hall, T., Biosciences, I., & Carlsbad, C. (2011). BioEdit: an important software for molecular biology. GERF Bull Biosci, 2(1), 60-61.

Hopkins, J. (2020). SARS-CoV-2 Genetics. Center for Health Security. https://www.centerforhealthsecurity.org/resources/COVID-19/COVID-19-fact-sheets/200128-nCoV-whitepaper.pdf

Koyama, T., Platt, D., & Parida, L. (2020). Variant analysis of SARS-CoV-2 genomes. Bull World Health Organization, 98(7), 495.

Maddison, W. P., & Maddison, W. P. (2002). Macclade. sunderland. MA: sinauer Associates.

Matsuda, T., Suzuki, H., & Ogata, N. (2020). Phylogenetic analyses of the severe acute respiratory syndrome coronavirus 2 reflected the several routes of invasion in Taiwan, the United States and Japan. arXiv preprint arXiv:2002.08802.

Nguyen, T. M., Zhang, Y., & Pandolfi, P. P. (2020). Virus against virus: a potential treatment for 2019-nCov (SARS-CoV-2) and other RNA viruses.

Paul, S. M. (2006). The molecular biology of coronavirus. Adv Virus Res, 66(48), 193-292.

Perlman, S., & Netland, J. (2009). Coronaviruses post-SARS: update on replication and pathogenesis. Nature Rev Microbiol, 7(6), 439-450.

Saitou, N., & Nei M (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol, 4, 406-425.

Samhouri, D., Ijaz, K., Rashidian, A., Chungong, S., Flahault, A., Babich, S. M., & Mahjour, J. (2018). Analysis of Joint External Evaluations in the WHO Eastern Mediterranean Region. East Mediterr Health J, 24(5), 477–87.

Swofford, D. L. (2002). PAUP: phylogenetic analysis using parsimony, version 4.0 b10.

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol, 10(3), 512-526.

Wang, L. F., Shi, Z., Zhang, S., Field, H., Daszak, P., & Eaton, B. T. (2006). Review of bats and SARS. Emerg Infect Diseases, 12(12), 1834.

WHO. (2020). Coronavirus disease 2019 (COVID-19): situation report, 82.

Zimmerman, P.A., Buckler-White, A., Alkhatib, G., Spalding, T., Kubofcik, J., Combadiere, C., Weissman, D., Cohen, O., Rubbert, A., Lam, G., & Vaccarezza, M. (1997). Inherited resistance to HIV-1 conferred by an inactivating mutation in CC chemokine receptor 5: studies in populations with contrasting clinical phenotypes, defined racial background and quantified risk. Mol Med, 3(1), 23-36.