Phylogenetic analysis of variable and conserved genomic regions in severe acute respiratory syndrome coronavirus 2 (COVID-19)

Abeer F. El Nahas (abeer.elnahas@alexu.edu.eg)

Genetics Laboratory, Animal Husbandry and Animal Wealth Development Department. Alexandria University. Egypt https://orcid.org/0000-0001-8452-5557

Nasema M. Elkatatny

Biotechnology Department, Animal Health Research Institute, Agriculture Research Center, Egypt https://orcid.org/0000-0002-7459-6420

Haitham G. Abo-Al-Ela

Department of Aquaculture, Faculty of Fish Resources, Suez University, Suez, Egypt. https://orcid.org/0000-0003-4157-5372

Short Report

Keywords: COVID-19, phylogenetic analysis, ORF1ab, Orf3a, N gene, E gene, M gene

Posted Date: October 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-88200/v1

Abstract

SARS-CoV-2 has rapidly spread around the world. Several mutations have been detected in its genome, but they do not seem to affect the abilities of the virus to spread or infect. We aimed to explore the conserved genomic regions in coronavirus that could contain the key strengths of the virus. SARS-CoV-2 sequence data were retrieved from Genbank from the period of December 2019 to March 2020. Phylogenetic analyses were conducted for 207 sequences using MEGAX compared with the reference sequence (MN908947.3- CHN-Wuhan Dec-2019). The analysis included seven important genomic regions, the *ORF1ab* gene (21,290 bp), S gene (3,822 bp), *Orf3a* gene (827 bp), E gene (227 bp), M gene (669 bp), and N gene (1,259 bp), which play critical roles in virus invasion and replication. Furthermore, the variant nucleotides and amino acids were detected by MEGAX and BLAST. Through the phylogenetic analysis and amino acid substitution, the *ORF1ab* gene showed 11 conserved regions and also several variable sites. The E and M genes were mainly conserved, and all sequences were included in one clade, with one or two amino acid variants. *Orf3a* and the N gene have four conserved sites distributed along the genes. The S gene has 12 mutations and four main large conserved regions

We conclude that the favored occurrence of mutations at the *ORFab* and *Orf3a* genes during the SARS-CoV epidemic is an important mechanism for virus pathogenesis. The E and M proteins have an almost conserved structure, whereas the S and N genes have many conserved regions, which could serve as possible targets for vaccine design for SARS-CoV.

Introduction

Coronaviruses are a large family of RNA viruses that cause different coronavirus diseases, including severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and the common cold. In late 2019, a new member was detected in the family, named SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), causing COVID-19 disease (coronavirus disease 2019) [1]. SARS-CoV-2, a virus that will permanently change the world, has been considered a pandemic. Since its detection in Wuhan, Hubei Province, China, the virus has been widely spreading, and the number of infected patients and deaths are notably rising every day.

Coronaviruses are RNA in nature; therefore, their mutation frequencies are 300-fold higher than those of DNA-based viruses. These viruses show frequent genetic recombination and mutations [2]. In SARS-CoV-2, Tang et al. [3] identified mutations in 149 genomic sites across 103 sequenced strains.

The question is, “Do these mutations affect or prevent an effective vaccine against SARS-CoV-2 from being developed?” Many previous coronavirus vaccine formulations have failed, raising a need for developing rapid response vaccine platforms for coronaviruses [4].

The coronavirus genome can be divided into (i) the first two-thirds, which encodes the replicase genes and is processed into 15 or 16 non-structural proteins [5], and (ii) the remaining one-third, which encodes open reading frame (ORFs) for the structural proteins and the spike (S), envelope (E), membrane (M), and

nucleocapsid (N) proteins [6]. Of these, the envelope-embedded surface-located spike (S) glycoprotein mediates the entry process [7]. S proteins that are expressed on the virus surface can stimulate host antibodies, which can neutralize the virus [8]. The functional domains in the S protein of SARS-CoV-2 comprise a signal peptide, N-terminal domain, receptor-binding domain (RBD), fusion peptide, heptad repeat 1, heptad repeat 2, transmembrane domain, and cytoplasmic domain. RBD is a target for the development of vaccines and antibodies, while heptad repeat 1 is a target for fusion/entry inhibitors [9].

According to GenBank accession no. MN908947.3, SARS-CoV-2 comprises a number of open-reading frames that encode six accessory proteins, namely Orf1ab, Orf3a, Orf6, Orf7a, Orf8, and Orf10. Orf1ab and Orf3a are efficiently expressed on the host cell surface and also present intracellularly in patients infected with SARS-CoV.6 The other SARS-CoV viral particles (Orf6, 7a, 7b, 8a, 8b, and 9) are not essential for the viral cycle [10].

We aimed to explore both the variable and conserved genomic regions in SARS-CoV-2 using a phylogenetic analysis. The analysis was conducted on seven important genomic regions that play essential roles during virus invasion and replication. The analysis included the *ORF1ab* gene, *S* gene (spike glycoprotein), *Orf3a* gene, *E* gene (envelope protein), *M* gene (membrane glycoprotein), and *N* gene (nucleocapsid phosphoprotein).

Methods

Full genome sequences of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2, 2019-nCoV) were obtained from the gene bank database. A total of 207 full-genome sequences were used, covering the samples collected from the period of December 2019 to March 2020. A phylogenetic analysis was performed on seven important regions in the virus genome: the *ORF1ab* gene (21,290 bp), *S* gene (spike glycoprotein) (3,822 bp), *Orf3a* gene (827 bp), *E* gene (envelope protein) (227 bp), *M* gene (membrane glycoprotein) (669 bp), and *N* gene (1,259 bp). The accession numbers used in this study are provided in the supplementary data. Each sequence was cleaved into the seven studied regions, and the number of the accession numbers used in each region were variable according to the quality of the sequence of each segment. All sequences were compared with the reference sequence (MN908947.3- CHN-Wuhan Dec-2019). The sequences of each segment were aligned using MUSCLE in MEGAX [11], and amino acid alignment and substitution were performed using MEGAX. The neighbor-joining phylogenetic tree included four rate categories and was constructed using MEGAX, and the robustness of the tree topology was assessed with 1,000 bootstrap replicates. All parameters were estimated from the data. The gamma distribution with invariant sites (G+I) was used to model the evolutionary rate differences among sites. The variant nucleotides and amino acids were detected by MEGAX, and their corresponding numbers were determined using BLAST for nucleotides and proteins (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Results

The phylogenetic analysis of 155 sequences of exon 1 (*ORF1ab* gene (21290 bp) based on their nucleotide sequences revealed the presence of several clades and clusters of the virus as an indication of the large number of mutations (Fig. 1A). The analysis of the amino acid variants revealed the presence of multiple variable and conserved regions. The variable regions contained more than one amino acid substitution (data not shown). Additionally, 11 conserved regions without amino acid substitutions were detected along exon 1 (Fig. 1B. Table 1).

Regarding the output analysis of 197 different sequences of the *S* gene (surface glycoprotein), one clade and two clusters of the virus were recorded (Fig. 2 A). A comparison of amino acid variants revealed the presence of 13 mutations separating many conserved regions. Four main conserved regions presented 124, 260, 1015, and 105 amino acids distributed along the gene (Fig. 2B, Table 1). Four conserved regions were detected on the *Orf3a* gene (1-43, 141-195, 197-250, and 255-275), and non-synonymous mutations (10) were detected along its amino acid residues (275) (Fig. 3B, Table 1). Additionally, the presence of several clusters of the virus was recorded in the phylogenetic tree (Fig. 3A; Table 1).

A conserved structure of the *E* gene of envelope protein was observed in the phylogenetic tree (one cluster of all 179 sequences) (Fig. 4 A), and one amino acid variant was detected in only one sequence (Fig. 4B, Table 1). Similarly, the *M* gene (membrane glycoprotein) likely tends to be conserved, as it has only one cluster for 197 sequences in the phylogenetic tree (Fig. 5A.), and two amino acid substitutions were detected (Fig. 5B, Table 1). Regarding the *N* gene of nucleocapsid phosphoprotein, many variable sites were recorded at this gene; the phylogenetic tree showed three clusters of the virus (Fig. 6 A), and 14 amino acids substitutions were present (Fig. 6B, Table 1). However, four conserved sites were detected at this gene at 13-193, 212-271, 288-327 and 344-419 (Fig. 6B, Table 1).

Discussion

Many studies have been released on the role of structural and accessory proteins in the pathogenesis of severe acute respiratory syndrome coronavirus (SARS-CoV) infections, yet a proper vaccine is still not available. The accessory proteins encoded by coronaviruses help the virus infect the host and enhance virus virulence [12]. Viruses mutate all the time. The mutation of COVID-19 varies across different parts of the world. A genetic tracking and network analysis can provide a better understanding of antigenic drift and improve the detection and the control of novel emerging strains [13].

ORF1a and ORF1b (ORFab) are SARS-CoV accessory proteins, known as the replicase/transcriptase genes; they are translated to proteins that are responsible for viral RNA replication and transcription, and they are important during viral pathogenesis [14, 15]. We have reported many mutations along the largest SARS-CoV exon (21555 bp). Evidence for alteration in the ORF1ab coding sequence during the coronaviruses epidemic indicates that the ORF1ab proteins play roles in virus pathogenesis in addition to viral replication [14]. Additionally, Ketteler revealed the presence of a frameshifting stimulation element and a conserved RNA sequence forming a stem-loop that allows ribosomal frameshifting, a mechanism in which open-reading frame 1b (orf1b) is expressed [16].

Several mutations were recorded in the *S* protein between 4 and 613 a.a. Similarly, Kim et al. [17] recorded four non-synonymous mutations in the MERS-CoV *S* gene from strains isolated in South Korea distributed from 137 to 629 a.a; the mutations were located at the site that does not interfere with the host receptor. Kleine-Weber et al. [18] reported that D510G and I529T mutations in RBD of the S protein resulted in a decrease in the binding affinity to DPP4 and reduced viral entry into target cells. In addition, these mutations increased resistance to antibody-mediated neutralization; however, none of these mutations were recorded in all sequences included in this study.

Orf3a is one of the accessory proteins of the SARS-CoV; it is the largest unique open reading frame of the virus genome, and it comprises three transmembrane domains [19]. The *Orf3a* gene encodes for protein 3a; it is expressed on the patient cell surface and can be easily detected in SARS patients, stimulating a humoral and cellular immune response [20]. Yount et al. [21] suggested the importance of this gene through a significant reduction in virus titers following infection with deleted ORF3a recombinant virus. Our data revealed the presence of 10 non-synonymous mutations along the *Orf3a* gene together with four conserved regions. Interestingly, Tan et al. [22] and Wang et al. [23] found the advantage for the occurrence of frameshift mutations in the protein 3a gene, as this mutation encodes for 3a variants. Additionally, Lu et al. [24] induced Cys133 point mutations at the gene, which is important for protein oligomerization and virus pathogenesis in the host cells.

The conserved structure of the E gene of the envelope protein of the coronavirus may be explained by the vital roles of this protein; it is involved in many important aspects of the virus life cycle: pathogenesis, envelope formation, budding, viral assembly, and structural motifs and virus topology [25, 26]. All E proteins have conserved cysteine residues. Lopez et al. [27] proposed the importance of the conserved cysteines of coronavirus envelope (E) for virus production, as the virus with multiple mutations at three cysteine residues at positions 40, 44, and 47 exhibited an increased rate of its degradation. Additionally, DeDiego et al. [28] proposed that a lack of the E gene caused in vivo and in vitro attenuation of SARS-CoV; this could be used for the development of a live attenuated SARS-CoV vaccine.

The coronavirus M protein plays a major role in virus assembly, when the virus and host factors come together to make new virus particles; this protein is also involved in virus spike density, and its interaction with genomic RNA and S and N proteins regulates virions [29]. Only two mutations have been detected in M protein in the phylogenetic analysis of 197 sequences; this is coincident with the observation by den Boon et al. [30], who found that M protein is moderately well conserved within each coronavirus group. However, Hu et al. [31] demonstrated the highest substitution rate of SARS-CoV-M protein compared with other proteins among 12 coronaviruses; they related these variations to the selection regarding the host range or the ability to escape from host immuno-surveillance.

M protein is one of the proteins that attaches to the envelope membrane surface of the SARS-CoV particles. It has dominant cellular immunogenicity; it potentiates strong humoral response in infected patients; and together with its most conserved structure, it serves as a possible target for vaccine design for SARS-CoV [26, 32, 33]. The nucleocapsid (N) of coronavirus is a structural protein; it plays an

important role during assembly of the virion and also during virus transcription [34]. In this study, the phylogenetic analysis of N protein showed the presence of four conserved sites at the gene; interestingly, McBride et al. [34] proposed that CoV-N proteins have three distinct and highly conserved domains: an N-terminal domain, a C-terminal domain (CTD/domain 3), and a central region (RNA-binding domain); the location of these domains matches with the conserved regions detected in this study. Huang et al. [35] found that the structure of the N-terminal RNA-binding domain (NTD) of the SARS-CoV N protein is 45–181 amino acids. Additionally, they demonstrated that the Arg-94 and Tyr-122 residues in the IBV N protein are well conserved across the whole CoV family, and they are critical for SARS N-RNA binding.

Mutation rates are variable in the different regions of COVID-19; some regions have a high mutation rate, and other regions tend to be conserved. Koyama et al. [36] demonstrated that *ORF1ab* contains more variants amino acids in the NSP3 domain than in other domains.

The protective efficacy of vaccine-induced immunity to viral infection depends mainly on adaptive immune responses. The success of vaccination depends on the properties of the recognized antigen; its ability to activate, expand and memorize a multitude of specialist functions of lymphocytes; and its ability to control the spread and maintain the viral pathogen within a population [37]. We suggest that with the use of recombinant vaccines targeting wide ranges of strategies by using the conserved regions of COVID-19, intervention for this virus may become possible.

Based on the sequence data and the previous publications, we conclude that the favored occurrence of mutations at the *ORFab* and *Orf3a* genes during the SARS-CoV epidemic is an important mechanism in host cells for virus pathogenesis. E and M proteins have an almost conserved structure; the S and N genes have many conserved regions, and they could serve as possible targets for vaccine design for SARS-CoV.

Declarations

Funding:

The authors received no specific funding for this work.

Conflict of interest: The authors declare that they have no conflict of interest.

Author contribution: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [Abeer F. El Nahas], [Nasema M. Elkatatny]. The first draft of the manuscript was written by [Abeer F. El Nahas, Haitham G. Abo-Al-Ela] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Availability of data and material: on request

References

1. Naming the coronavirus disease (COVID-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. World Health Organization. 2020.
2. Wang Y, Sun J, Zhu A, Zhao J, Zhao J. Current understanding of middle east respiratory syndrome coronavirus infection in human and animal models. J. Thorac. Dis. 2020;10:S2260–S2271.
3. Tang X, Wu C, Li X, Song Y, Yao X et al. On the origin and continuing evolution of SARS-CoV-2. Nat Sci Rev. 2020; nwaa036.
4. Menachery VD, Gralinski LE, Mitchell HD, Dinnon KH, Leist SR, Boyd L. Combination attenuation offers strategy for live attenuated coronavirus vaccines. J Virol. 2018;92:e00710-00718.
5. Thiel V, Ivanov KA, Putics Á, Hertzig, T., Schelle, B. Mechanisms and enzymes involved in SARS coronavirus genome expression. J Gen Virol. 2003; 84: 2305–15.
6. Liu DX, FungTS, Chong KK-L, Shukla A, Hilgenfeld R. Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res. 2014;109: 97–109.
7. Lu G, Wang Q, Gao GF. Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends Microbiol. 2014;23: 468–78.
8. Song F, Fux R, Provacia LB, Volz A, Eickmann M, et al. Middle East respiratory syndrome coronavirus spike protein delivered by modified vaccinia virus ankara efficiently induces virus-neutralizing antibodies. J. Virol. 2013; 87: 11950–4.
9. Jiang S, Du L, Shi Z. An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies. Emerg Microbes Infect. 2020; 9: 275–7.
10. DeDiego ML, Pewe L, Alvarez E, Rejas MT, Perlman S, Enjuanes L. Pathogenicity of severe acute respiratory coronavirus deletion mutants in hACE-2 transgenic mice. Virology (2008;376:379–9.
11. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 2018;35(6): 1547–9.
12. LuW, Xu K, and Sun B. SARS Accessory Proteins ORF3a and 9b and Their Functional Analysis. In: Lal SK (ed) Molecular Biology of the SARS-Coronavirus. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010; pp 167–75.
13. Woo PC, HuangY, Lau SK,Yuen KY. Coronavirus genomics and bioinformatics analysis. Viruses 2010;2:1804–20.
14. Baranov PV, Henderson CM, Anderson CB, Gesteland RF, Atkins JF, Howard "Programmed ribosomal frameshifting in decoding the SARS-CoV genome". Virology 2005;332(2):498-510.
15. FehrAR, Perlman S. Coronaviruses: An overview of their replication and pathogenesis. Method Mol 2015;1282: 1-23.
16. Ketteler R. On programmed ribosomal frameshifting: the alternative proteomes. Front Genet. 2012;3: 242.
17. Kim D-W, KimY-J, Park SH, Yun MR, Yang J-S et al. Variations in spike glycoprotein gene of MERS-CoV, South Korea, 2015. Emerg. Infect. Dis. 2016;22:100–4.

18. Kleine-Weber H, Elzayat MT, Wang L, Graham BS, Müller MA, et al. Mutations in the spike protein of Middle East respiratory syndrome coronavirus transmitted in Korea increase resistance to antibody-mediated neutralization. J Virol. 2019;93:e01381-01318.
19. Liua DX, Funga S, Chonga LK, Shuklab A, Hilgenfeld Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res. 2014;*109*, 97-109.
20. Lu B, Tao L, Wang T, Zheng Z, Li B et al. Humoral and cellular immune responses induced by 3a DNA vaccines against severe acute respiratory syndrome (SARS) or SARS-like coronavirus in mice. ClinVaccine Immunol. 2009;16:73–7.
21. Yount B, Roberts RS, Sims AC, Deming D, Frieman MB, et al. Severe acute respiratory syndrome coronavirus group-specific open reading frames encode nonessential functions for replication in cell cultures and mice. J Virol. 2005;79: 14909–22.
22. TanT, Barkham T, Fielding BC, Chou C-F, Shen S. et al. Genetic lesions within the 3a gene of SARS-CoV. Virol J. 2005; 2: 51.
23. Wang X, Wong S-M, Liu D. Identification of hepta-and octo-uridine stretches as sole signals for programmed +1 and _1 ribosomal frameshifting during translation of SARS-CoV *ORF 3a* Nucleic Acids Res. 2006;34:1250–60.
24. Lu W, Zheng B, Xu K, Schwarz W, Du L, Charlotte KL. Severe acute respiratory syndrome-associated coronavirus 3a protein forms an ion channel and modulates virus release. PNAS. 2006;103(33):12540-12545.
25. Ye Y, Hogue BG. Role of the coronavirus E viroporin protein transmembrane domain in virus assembly. J Virol. 2007; 81(7):3597–607.
26. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. Virol J. 2019;16:69.
27. Lopez LA, Riffle AJ, Pike SL, Gardner D, Hogue BG. Importance of Conserved Cysteine Residues in the Coronavirus Envelope Protein. J. virol. 2008; 82:3000–3010.
28. DeDiego ML, Álvarez E, Almazán F, Rejas MT, Lamirande E, et al. A Severe Acute Respiratory Syndrome Coronavirus That Lacks the E Gene Is Attenuated In Vitro and In Vivo. J. Virol. 2007;81(4):1701–1713.
29. Neuman BW, Kiss G, Kunding AH, Bhella D, Baksh MF, et al. A structural analysis of M protein in coronavirus assembly and morphology. J Struct Biol. 2011;174(1): 11–22.
30. den Boon JA, Snijder EJ, Locker JK, Horzinek MC, Rottier PJM. Another triple-spanning envelope protein among intracellularly budding RNAviruses:The torovirus E protein. Virology 1991; 182:655–663.
31. HuY, Wen J, Tang L, Zhang H, Zhang X, Li et al. The M protein of SARS-CoV: basic structural and immunological properties. Genom Proteom Bioinformat. 2003; 1:118–130.
32. Kilianski A, Mielech A, Deng X, Baker SC. Assessing activity and inhibition of MERS-CoV papain-like and 3C-like proteases using luciferase-based biosensors. Virol. 2003; 66: JVI. 02105–02113.

33. Rest JS, Mindell DP. SARS associated coronavirus has a recombinant polymerase and coronaviruses have a history of host-shifting. Infect. Genet. Evol 2003;3;219–225.

34. McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional Protein. Viruses. 2014;6(8): 2991–3018.

35. Huang Q, Yu L, Petros AM, Gunasekera, A., Liu Z, et al. (2004) Structure of the N-terminal RNA-binding domain of the SARS CoV nucleocapsid protein. *Biochemistry 43(20),* 6059-6063.

36. Koyama T, Platt D, Parida L. Variant analysis of COVID-19 genomes. [Submitted]. Bull World Health Organ. E-pub: 24 February. 2020. doi: http://dx.doi.org/10.2471/BLT.20.253591

37. Afrough B, Dowall S, Hewson R. Emerging viruses and current strategies for vaccine intervention. Clin Exp Immunol. 2019;196(2): 157–166.

Table

Table 1. Conserved and variable sites of different regions of severe acute respiratory syndrome coronavirus 2

Gene name	Number of nucleotide	Number of amino acids	Conserved site (based on a.a)	Variable site (based on a.a)
ORF1ab (Exon 1)	21290 bp	7096	392-543 600-902 1116-1323 1352-1473 1600-1820 1841-2125 2274-2418 2709-2890 3100-3346 3697-3827 4177-4414	1-391 544-599 903-1115 1324-1351 1474-1599 1821-1840 2126-2273 2419-2709 2891-3099 3347-3696 3828-4176
S gene (spike glycoprotein)	3822	1273	98-220 222-481 614-1077 1196-1273	13 mutations
Orf3a gene	828	275	1-43 141-195 197-250 255-275	44-140 251-254
E gene (envelope protein)	228	75	All conserved except one mutation	One non-synonymous mutation at 36 L>H
M gene (membrane glycoprotein)	669	222	All conserved except two mutations	Two non-synonymous mutations 1 A>S 70 V>I
N gene	1259 bp	419	13-193 212-271 288-327 344-419	14 non-synonymous mutations (Fig. 5)

Figures

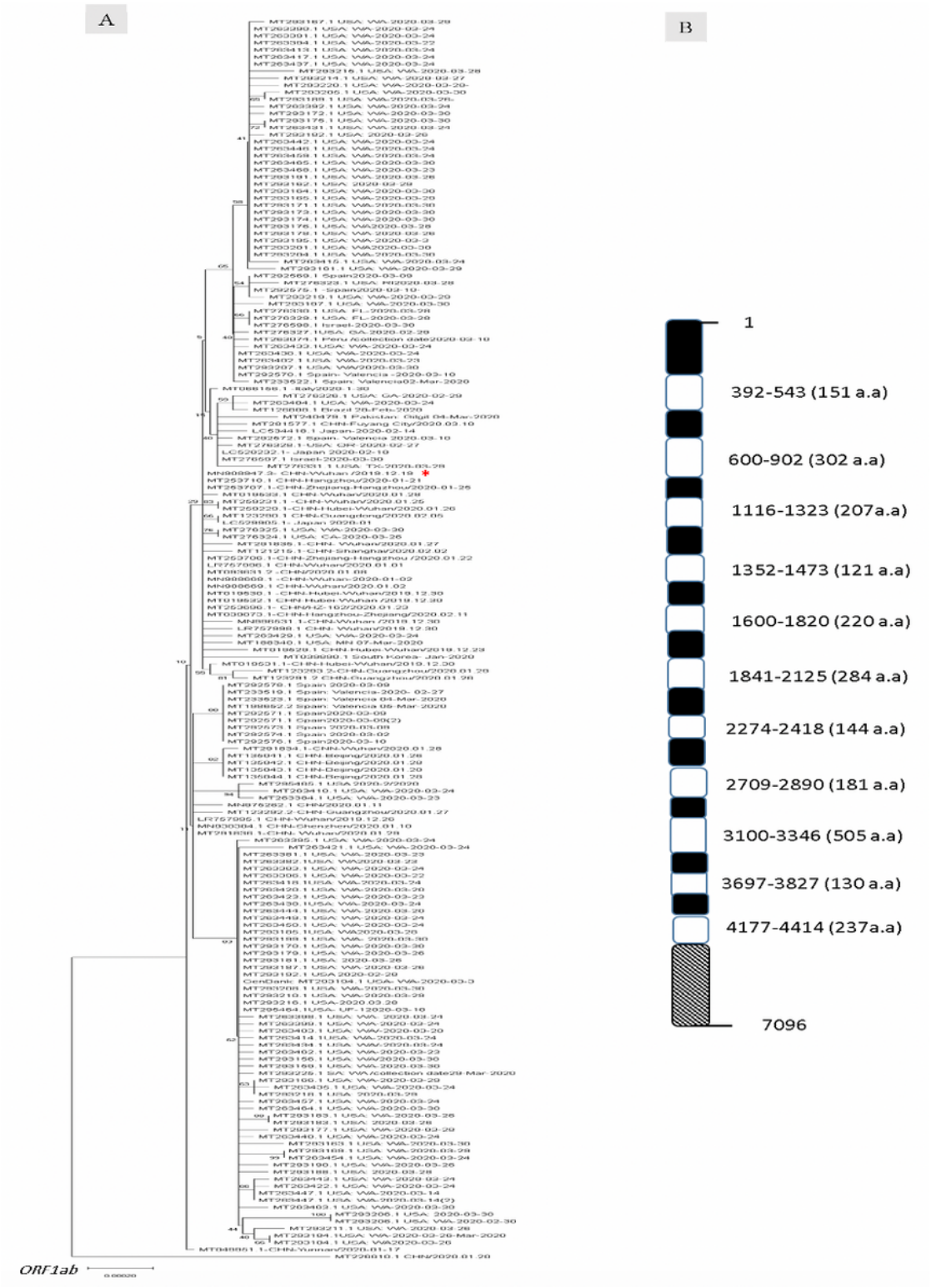

Figure 1

A) The neighbor-joining tree of 155 different nucleotide sequences of exon 1 (ORF1ab gene) (21,290 bp) of severe acute respiratory syndrome coronavirus 2 covering the region of nucleotides from 226 to 21,290 bp. B) The corresponding amino acid substitutions residues (1 to 7096); the white parts of the column

represent the conserved parts among all the virus isolates (without amino acids substitution) while the variable sites are included at the black parts. The linear end of the column has no comparisons as it contains gaps. * indicate the reference sequence (MN908947.3- CHN-Wuhan Dec-2019).

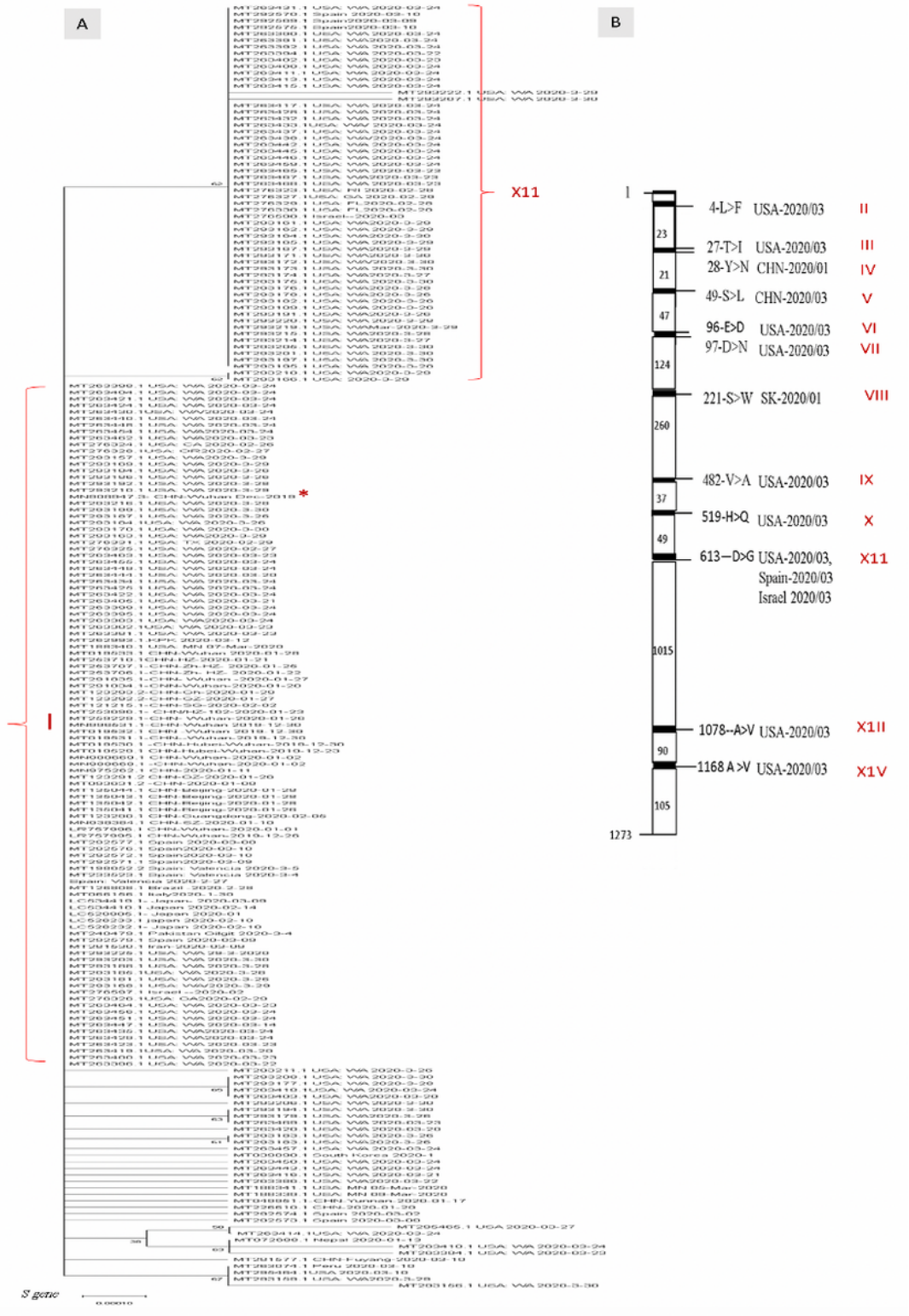

Figure 2

A) The neighbor-joining tree of 197 different nucleotide sequences of S gene (3,822 bp) of severe acute respiratory syndrome coronavirus 2 covering the region of nucleotides 21563 to 25384. B) The

corresponding amino acid substitution residues (1 to 1273). * indicates the reference sequence (MN908947.3- CHN-Wuhan Dec-2019), where I is the basic genotype (similar to the reference sequence) and II-XIV are the different genotypes with amino acid substitutions within the S gene. Genotype XII was recorded among 41 sequences, and the other genotypes were recorded in only one sequence. The different numbers listed are the a.a numbers of the conserved regions.

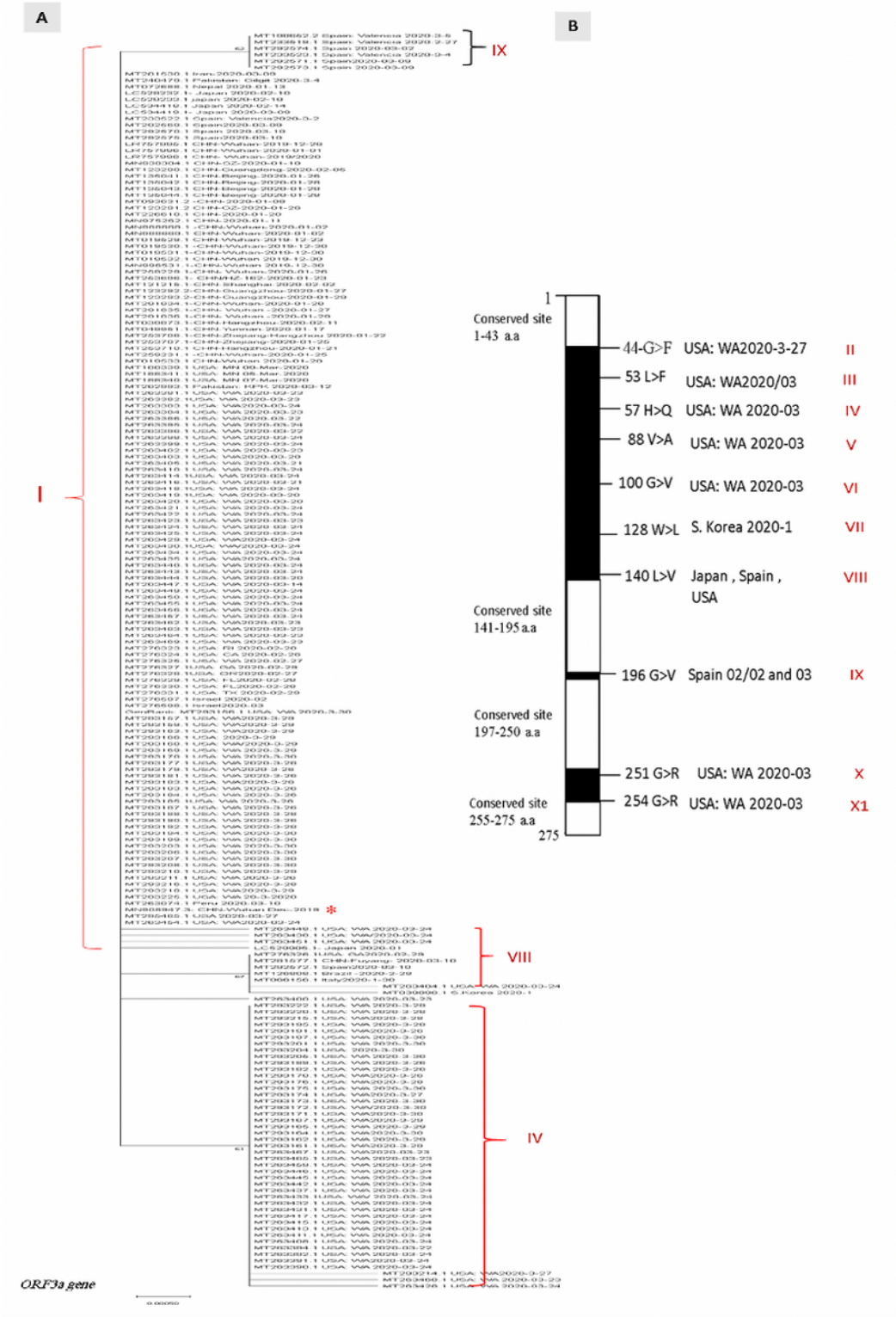

Figure 3

A) The neighbor-joining tree of 197 different nucleotide sequences of the ORF3a gene (827 bp) of severe acute respiratory syndrome coronavirus 2 covering the region of nucleotides from 25393 to 26220. B) The corresponding amino acid substitution residues (1 to 275). * indicates the reference sequence (MN908947.3- CHN-Wuhan Dec-2019), where I is the basic genotype (similar to the reference sequence) and II-XI are the different genotypes with amino substitution within 1-275 of the ORF3a gene. Genotype XII was recorded among 41 sequences, and the other genotype were recorded in only one sequence.

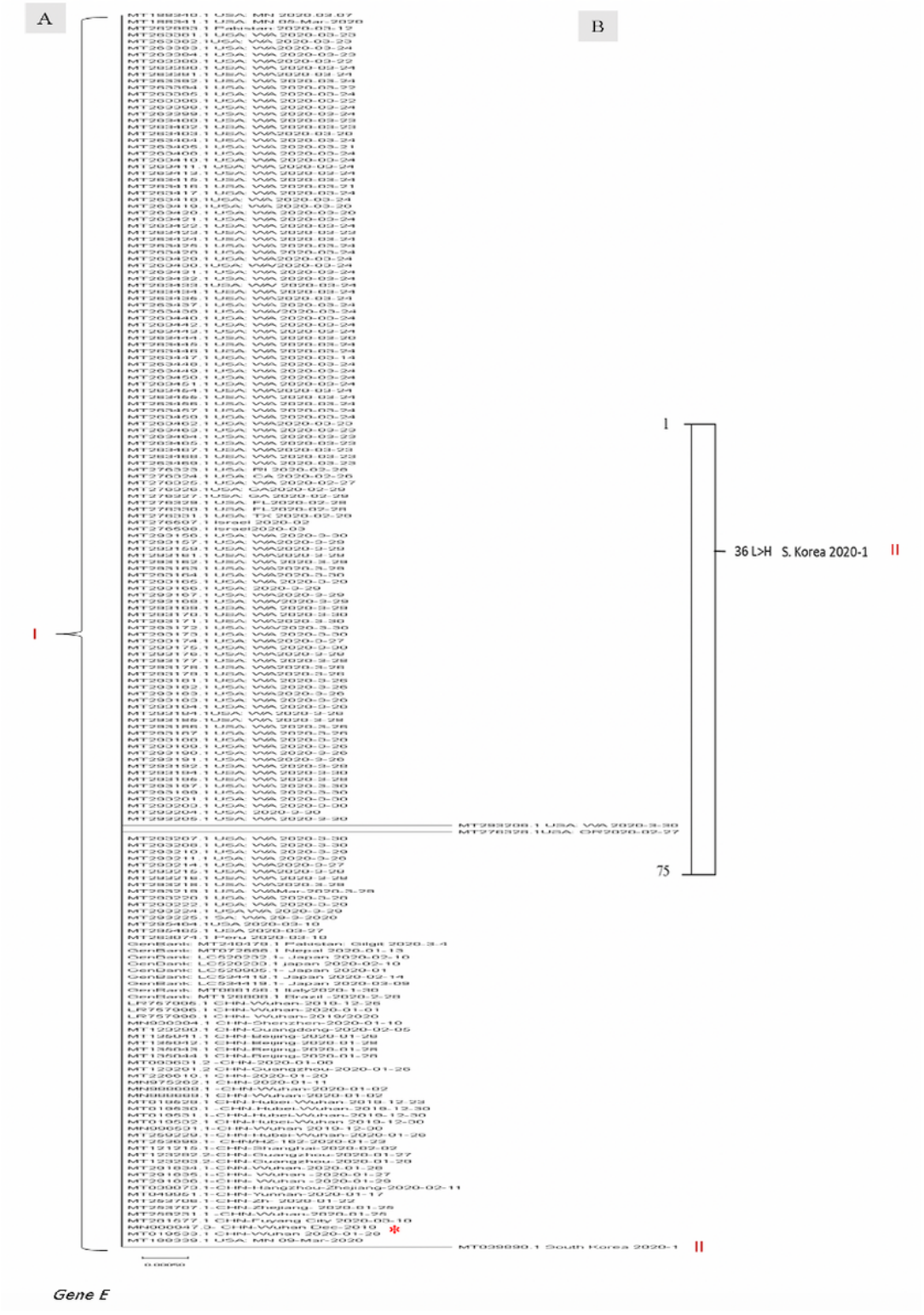

Figure 4

A) The neighbor-joining tree of 197 different nucleotide sequences of the E gene, which encodes for the envelope protein (227 bp) of severe acute respiratory syndrome coronavirus 2, covering the region of nucleotides from 26238 to 26465. B) The corresponding amino acid substitutions residue (1 to 75). * indicates the reference sequence (MN908947.3- CHN-Wuhan Dec-2019), where I is the basic genotype (similar to the reference sequence) and II is a second genotype recorded only once in a South Korea isolate.

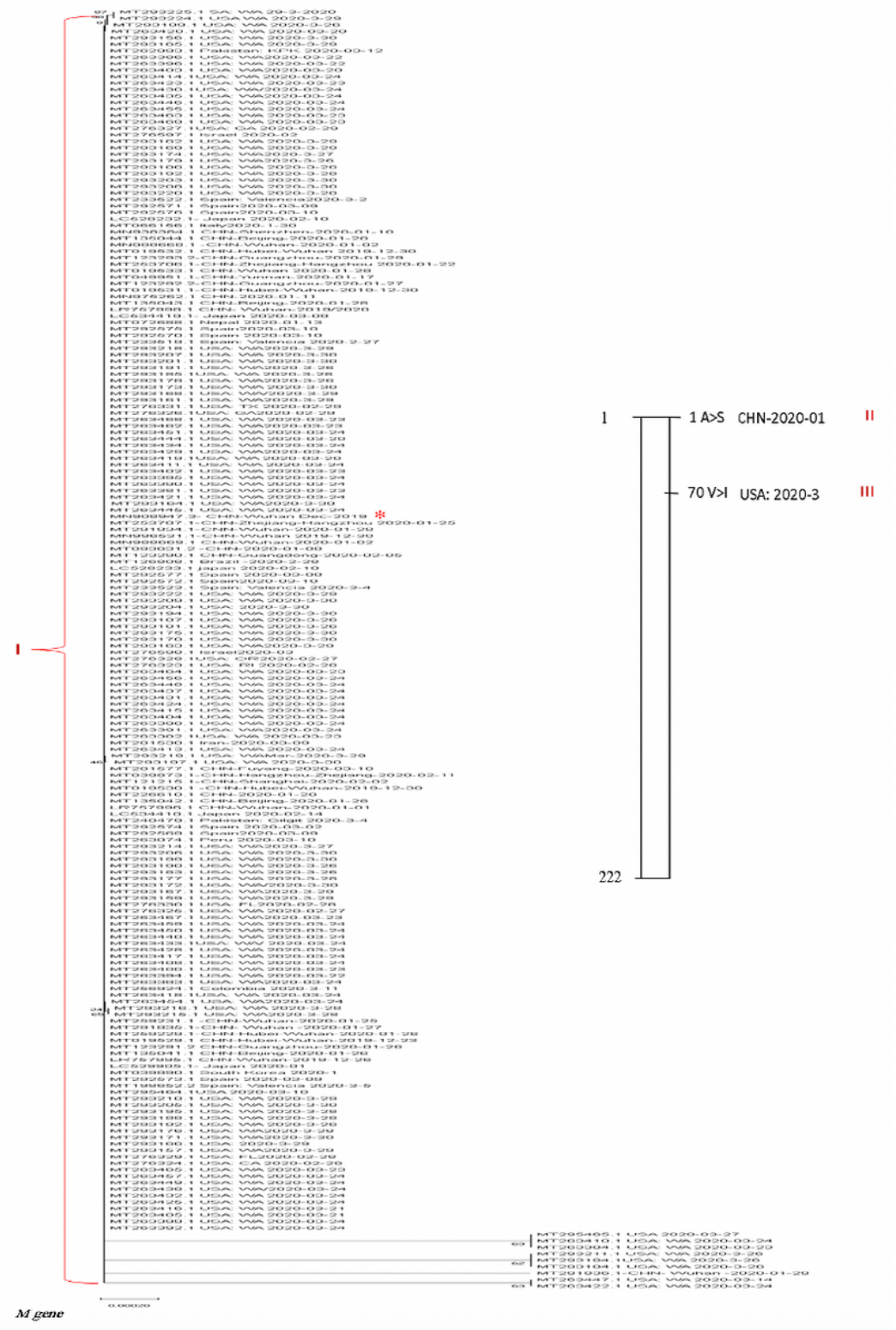

Figure 5

A) The neighbor-joining tree of 197 different nucleotide sequences of the M gene, which encodes for the membrane glycoprotein (669 bp) of severe acute respiratory syndrome coronavirus 2, covering the region of nucleotides from 26523 to 27191. B) The corresponding amino acid substitution residues (1 to 222). * indicates the reference sequence (MN908947.3- CHN-Wuhan Dec-2019), where I is the basic genotype (similar to the reference sequence) and II and III are two different genotypes with amino substitutions within the N gene.

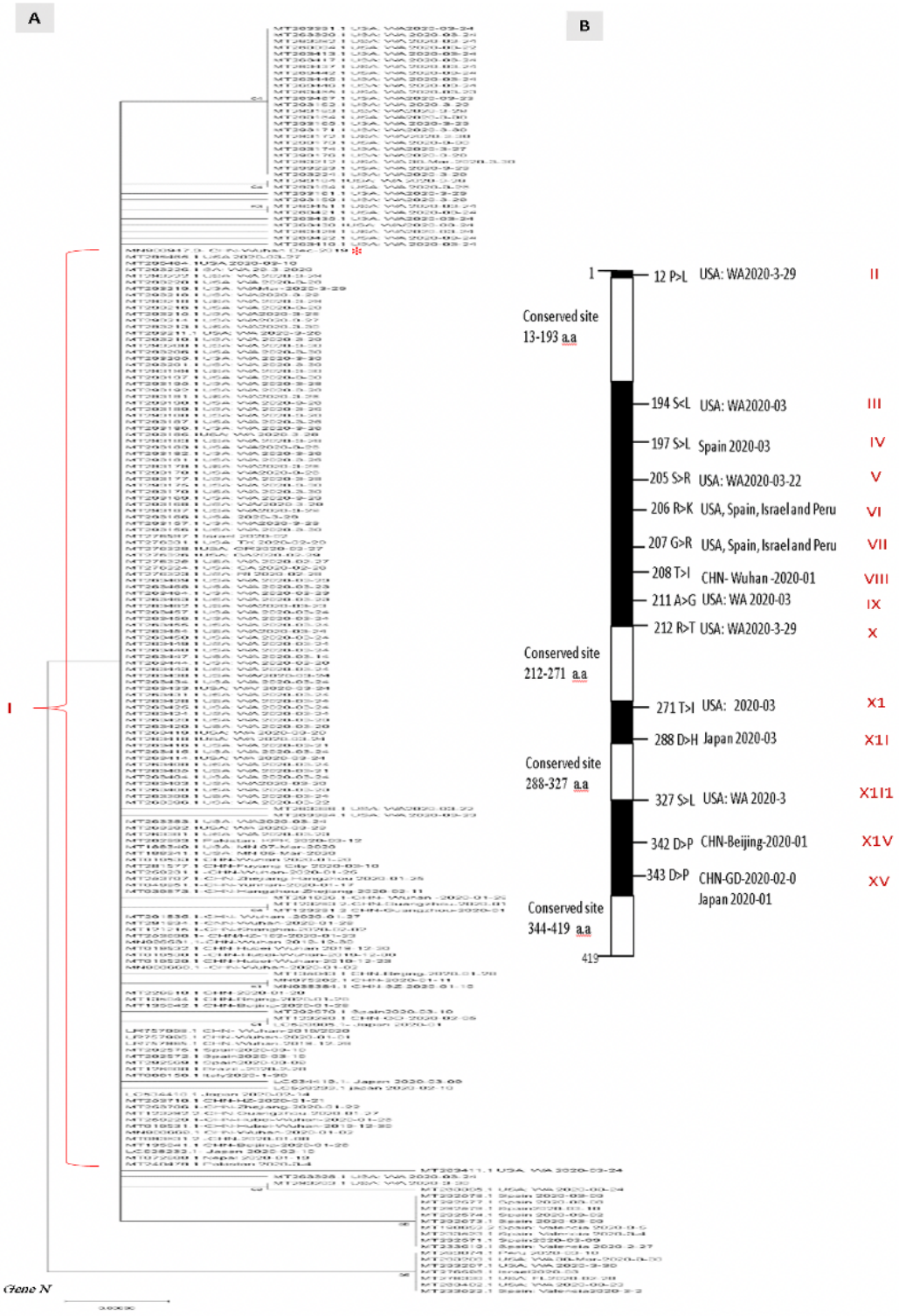

Figure 6

A) The neighbor-joining tree of 197 different nucleotide sequences of the N gene, which encodes for nucleocapsid phosphoprotein (1,259 bp) of severe acute respiratory syndrome coronavirus 2, covering the region of nucleotides from 28274 to 29533. B) The corresponding amino acid substitution residues (1 to 419). * indicates the reference sequence (MN908947.3- CHN-Wuhan Dec-2019), where I is the basic genotype (similar to the reference sequence) and II-XV are different genotypes with amino substitutions within 1-419 of the N gene. Four conserved sites were recorded without amino acid substitutions (marked in white).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Suppelmentarydata.docx