Positive selection of ORF3a and ORF8 genes drives the evolution of SARS-CoV-2 during the 2020 COVID-19 pandemic

Lauro Velazquez-Salinas [1,2 *], Selene Zarate [3], Samantha Eberl [4], Douglas P [1]. Gladue, Isabel Novella [5] and Manuel V. Borca [1]

Affiliations

1. Foreign Animal Disease Research Unit, USDA/ARS Plum Island Animal Disease Center, PO Box 848, Greenport NY 11944, USA.

2. College of Veterinary Medicine, Kansas State University, Manhattan, KS 66506, USA.

3. Posgrado en Ciencias Genómicas, Universidad Autónoma de la Ciudad de Mexico, Mexico City, Mexico.

4. Central Connecticut State University, New Britain, CT 06053.

5. Independent researcher

* Corresponding author: Lauro Velazquez-Salinas. Lauro.velazquez@usda.gov

Abstract

In this study, we analyzed full-length SARS-CoV-2 genomes from multiple countries to determine early trends in the evolutionary dynamics of the novel COVID-19 pandemic. Results indicated SARS-CoV-2 evolved early into at least three phylogenetic groups, characterized by positive selection at specific residues of the accessory proteins OFR3a and ORF8a. We also report evidence of epistatic interactions among sites in the genome that may be important in the generation of variants adapted to humans. These observations might impact not only public health, but also suggest more studies are needed to understand the genetic mechanisms that may affect the development of therapeutic and preventive tools, like antivirals and vaccines.

Keywords: Novel coronavirus, pandemic, evolution, epistasis, positive selection, COVID-19, SARS-CoV2, phylogenetics.

Introduction.

The first case of pneumonia confirmed to be caused by the novel virus SARS-CoV-2 was in a 41-year-old man in Wuhan, Hubei province, China on 31/December/2019 (Wu et al., 2020). As of 9/April/2020, the World Health Organization (WHO) has confirmed 1,439,516 cases, 85,711 deaths, and the presence of COVID-19 in 209 countries, areas or territories. Of the confirmed cases, 71% are from seven countries: United States of America (395,030), Spain (146, 690), Italy (139, 422), Germany (108,202), China (83,249), France (81,095), and Iran (66,220) (https://www.who.int/emergencies/diseases/novel-coronavirus-2019). As of the writing of this report, the number of COVID-19 cases continue to increase worldwide, with multiple epicenters.

The International Committee on Taxonomy of Viruses (ICTV) initially named this pathogen 2019-nCoV (also referred to as COVID-19 by WHO) and included it within the *Coronaviridae* viral family (Coronaviridae Study Group of the International Committee on Taxonomy of, 2020). Later, based on the close phylogenetic relationship of COVID-19 with other human and bat severe acute respiratory syndrome coronaviruses (SARS-CoVs), ICTV renamed the virus as SARS-CoV-2 (Coronaviridae Study Group of the International Committee on Taxonomy of, 2020).

The *Coronaviridae* family encompasses a group of single-stranded, positive-sense RNA viruses with a genome length varying between 27 and 32 kb. These are zoonotic viruses with the potential to infect humans and animals. Coronaviruses may cause acute and chronic respiratory, enteric, and central nervous system infections(Phan et al., 2018, Weiss and Navas-Martin, 2005). In the case of SARS-CoV-2, a meta-analysis of 50,466 patients indicate that fever and cough are the most common symptoms (95% CI: 81.8-94.5% and 65.7-78.2%, respectively) (Sun et al., 2020). The disease may worsen, and the percentages of severe cases and fatality rate vary between 12.7-24.3% and 2.7-6.1% (95% CI), respectively (Sun et al., 2020).

The genome organization of SARS-CoV-2 is similar to viruses from the genus *Betacoronavirus*, one of the four genera included in the *Coronaviridae* subfamily Orthocoronavirinae. The ~29,903 nucleotide (nt) genome is organized as follows, 5' to 3': replicase ORF1ab, S (encoding the structural spike glycoprotein), ORF3a (ORF3a protein), E (structural envelope protein), M (structural membrane glycoprotein), ORF6 (ORF6 protein), ORF7a (ORF7a protein), ORF7b (ORF7b protein), ORF8 (ORF8 protein), N (structural nucleocapsid phosphoprotein), and ORF10 (ORF10 protein). ORF1ab (~21,291 nt) encodes sixteen non-structural proteins: leader protein, nsp2, nsp3, nsp4, 3C-like proteinase, nsp6, nsp7, nsp8, nsp9, nsp10, RNA-dependent RNA polymerase, helicase, 3'-to 5' exonuclease, endoRNAse, 2'-o-ribose methyltransferase, and nsp11(Wu et al., 2020) .

Much speculation regarding the origin of SARS-CoV-2 emanates from unfounded theories, such as a man-made laboratory origin; however, a recent study supports the hypothesis that SARS-CoV-2 was the result of cross-species transmission followed by natural selection in the novel human host (Andersen et al., 2020). This hypothesis is strongly supported by studies examining amino acid differences between SARS-CoV-2 and some phylogenetically related betacoronaviruses (e.g., Bat-RatG13 isolate and the human SARS-CoV isolate Urbani) at the receptor-binding domain (RBD) of the spike protein, where such differences seem to increase the ability of SARS-CoV-2 to bind to the human receptor angiotensin-converting enzyme 2 (ACE2) (Andersen et al., 2020). This increased affinity for binding ACE2 might help to explain the infectiousness of SARS-CoV-2 in human populations (Wan et al., 2020).

Considering the extraordinary plasticity shown by other human viral RNA pathogens, for example HIV-1, Influenza viruses, SARS-CoV, and hepatitis C virus, to undergo adaptative changes to evade innate and adaptive immune responses, develop drug resistance or establish an infection in a new host (Frost et al., 2018), multiple questions arise regarding the adaptative changes that SARS-CoV-2 has undergone during the pandemic. SARS-CoV-2 has spread throughout many countries resulting in the infection of people with diverse immunological backgrounds and demographics (age, sex, environmental conditions, etc) that potentially impose significant selective pressures on SARS-CoV-2.

Here, we evaluate the phylogenetic and evolutionary dynamics of SARS-CoV-2 during the early phase of the COVID-19 pandemic. Using different analyses based on a codon-based phylogenetic framework, we identified critical sites in the genome undergoing positive selection, which might favor viral divergence and emergence of multiple viral variants. Our findings are discussed in terms of the potential effects that the rapid evolution of SARS-CoV-2 might have on public health.

Materials and methods

Data collection

Eighty-six full-length SARS-CoV-2 genomes representing early viral isolates from patients living in diverse geographic regions were used for this study. Viral sequences, downloaded from the NCBI GenBank database on 06/03/2020, represent the total number of full-length viral genomes at the time that the analysis was conducted (Figure 1).

Phylogenetic dynamic analysis

Two phylogeographic trees were reconstructed using the programs BEAST2.4.3, BEAUti, and TreeAnnotator (Drummond et al., 2012) , introducing the country/region as a trait. The Markov Chain Monte Carlo was run for 25 million generations, using the HKY85 substitution model and a gamma distribution with four categories as the site heterogeneity model. The resulting file was analyzed with Tracer 1.6 to check for convergence and to determine the Burnin proportion (Rambaut et al., 2018). Finally, TreeAnnotator was used to build the maximum clade credibility tree, which was visualized with FigTree 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree). Additionally, the evolutionary rate of SARS-CoV-2, expressed as substitutions/sites/year, was calculated using the same methodology but using sampling date as a trait.

Pairwise distance calculations

Nucleotide and amino acid pairwise distance calculations among SARS-CoV-2 sequences were conducted using the SSE 1.3 Sequence Distances program (Simmonds, 2012). For this purpose, a sliding window of 50 nucleotides (nt), with a shift of 25 nt, was used to determine pairwise distances. Additionally, p-distances in nucleotide and amino acid sequences between phylogenetic groups were calculated using MEGA 7 (Kumar et al., 2016).

Evolutionary rate per site analysis

Mean (relative) evolutionary rates for each site in the alignment were estimated under the General Time Reversible model, including all 3 codon positions. These rates were scaled considering the average evolutionary rate across all sites is 1. This means that sites showing a rate <1 are evolving slower than average, and those with a rate >1 are evolving faster than average. This analysis was conducted using MEGA 7 (Kumar et al., 2016).

Inference of selective pressures

Since natural selection can be manifested as different modes (diversifying, directional or purifying), we used a combination of different evolutionary analyses to enhance the detection of relevant sites in the genome of SARS-CoV-2 experiencing diversifying (positive) and purifying (negative) selection: Single Likelihood Ancestor Counting (SLAC) (Kosakovsky Pond and Frost,

2005), Fixed Effects Likelihood (FEL) (Kosakovsky Pond and Frost, 2005), Mixed Effects Model of Evolution (MEME) (Murrell et al., 2012), and Fast Unbiased Bayesian Approximation (FUBAR) (Murrell et al., 2013). These methods use a maximum likelihood or Bayesian approach (FUBAR) to infer nonsynonymous (dN) and synonymous (dS) substitution rates on a per site basis for given coding alignment and corresponding phylogeny (Weaver et al., 2018). SLAC, FEL and FUBAR were methods used to identify sites experiencing pervasive diversifying or purifying selection, while MEME was used to detect sites experiencing episodic diversifying selection.

The presence of recombination in the sequence dataset potentially affecting the detection of positive selection was assessed using the algorithm GARD (Kosakovsky Pond et al., 2006). All methods were performed on the adaptive evolution server Datamonkey 2.0 (Weaver et al., 2018).

Evidence of directional selection was assessed on amino acid sequences using the Directional Evolution of Protein Sequences (DEPS) method, implemented on the Datamonkey webserver (classic) (Delport et al., 2010). This method is a model-based phylogenetic maximum likelihood test that looks for evidence of preferential substitution toward a given residue at individual positions of a protein alignment (Kosakovsky Pond et al., 2008).

Coevolution analysis

Evidence of coevolution among different sites in the SARS-CoV-2 genome was evaluated using the method Bayesian Graphical Models for co-evolving sites (BGM) (Poon et al., 2007). This method detects coevolutionary interactions between amino acids in a protein, where amino acid substitutions are mapped to branches in the phylogenetic tree.

Blosum 62 substitution matrix (BSM62)

BSM62 was used to infer the biological significance of amino acid replacements found during the evolutionary analysis of SARS-CoV-2, where positive values reflect that the substitution is most likely a product of random substitution, while negative values are indicative of selection (Henikoff and Henikoff, 1992).

Results

Phylogenetic dynamics of SARS-CoV-2

To evaluate potential divergence events of SARS-CoV-2, indicating the rise of new variants early during the pandemic, we reconstructed the evolution of SARS-CoV-2 using full-length genome sequences of viruses collected between late December of 2019 and early March of 2020 from patients infected in different countries around the world. The results of the phylogenetic analysis demonstrate the rapid divergence of SARS-CoV-2 into three distinct phylogenetic groups, differentiated only by a few changes at the nucleotide and amino acid levels (Figure 2A and 2B).

Group A includes the deduced ancestral sequence (NC_045512.2) obtained from the index case in Wuhan, China, and reported to the WHO on 31/December/ 2019, as well as multiple viral isolates from different Chinese provinces. The position of these sequences among multiple branches within the Group A cluster suggests the emergence of multiple viral variants in China, especially from Wuhan before the start of the global pandemic. Furthermore, the basal branch position of some of these variants indicates that they were the ancestors of viral isolates obtained

from patients in the USA, Japan, Finland, Taiwan, Nepal, and India between January and February 2020.

Similarly, in the Group B cluster, we found viral isolates from multiple Chinese provinces between December of 2019 and January of 2020. These isolates are likely ancestors of viral isolates recovered from patients in the USA, India, and Taiwan between January and March 2020. Interestingly, one isolate from Wuhan (LR757995.1) is part of the Group B cluster, supporting the hypothesis that multiple viral variants emerged in China before the start of the pandemic.

The Group C cluster was the only cluster that did not contain sequences from China. This cluster includes viral isolates collected from the USA, Italy, Australia, Sweden, Brazil, and South Korea between January and February of 2020. The absence of viral isolates from China and the increased genetic distance from Group A suggests that the emergence of these variants might have come from a second wave of transmission outside of China after the start of the pandemic.

Evolutionary divergence in the genome of SARS-CoV-2

Once we reconstructed the phylodynamic of SARS-CoV-2 isolates obtained early during the pandemic event, we attempted to determine which nucleotide positions in the SARS-CoV-2 genome were related to the early divergence of this virus. Overall, the evolutionary rate of SARS-CoV-2 is 1.15×10^{-3} substitutions/site/year (95% HPD 7.41×10^{-4} -1.57×10^{-3}), while pairwise analysis at nucleotide and amino acid levels revealed an average identity of 99.93-99.98% and 99.86-99.97%, respectively. Given the short divergence time, a high level of identity is to be expected; however, a few synonymous and non-synonymous substitutions were observed in the ORF1ab, S, OFR3a M, ORF8a, N and OFR10 genes (Figure 3A and 3B). When pairwise distances were calculated based on gene length, the highest levels of divergence were observed within genes ORF10 and ORF8a when considering synonymous and non-synonymous substitutions, respectively (Figure 3C and 3D).

Also, the estimated per site evolutionary rate in the coding regions revealed that 98.85% of the sites in the genome are evolving at expected rates of evolution, while 1.15% of the sites are evolving faster than expected (Figure 3E). In this context, and consistent with the length of the OFR1ab gene, most of these synonymous and non-synonymous substitutions (82 sites) were distributed among different protein-encoding segments of this gene; the segment encoding nsp3 had the highest number of polymorphic sites (Figure 3E).

Detection of purifying and diversifying selection

Once we identified fast-evolving positions within different genes of SARS-CoV-2, we used a combination of different algorithms centered on a codon-based phylogenetic framework to detect specific codons evolving under natural selection. Overall, no recombination events potentially affecting the results of these analyses were detected using the GARD algorithm.

Using SLAC we obtained a broad picture of the extent of natural selection acting upon the SARS-CoV-2 genome. We found an overall dN/dS ratio of 0.937 along the genome. In particular, 75 codons located within 5 genes (ORF1ab > S > N > ORF8a >ORF3a) showed evidence of increased fixation of non-synonymous mutations (dN/dS >1). Conversely, a small number of codons (35 codons) located within 5 genes (ORF1ab > N > S >M = ORF10) were accumulating a higher number of synonymous mutations (dN/dS <1). Interestingly, evaluation of

dN/dS at the level of individual genes showed higher ratios for the ORF3a and ORF8 genes (Figure 4A).

Significant purifying (negative) selection was observed in 12 out of the 35 codons evolving at dN/dS <1 using the FEL (12 sites) and FUBAR (1 site) methods; the codons were located in the ORF1ab, S, and N genes (Figure 4B). At these codons, increased fixation of synonymous substitutions seems to be favoring the phenotypic preservation of SARS-CoV-2 at specific residues of the proteins encoded by these genes. By tracking these mutations within different isolates, we observed that these changes could explain the divergence of different viruses within different phylogenetic groups. In some cases, mutations were associated with multiple isolates, supporting the relevance of these findings.

On the other hand, evidence of diversifying positive selection on non-synonymous sites was detected in just 4 of the 75 codons evolving at dN/dS >1 in genes ORF1ab and ORF3a, with the FUBAR and FEL methods providing the highest power of detection (Figure 4C). Based on this analysis, these four sites appear to be evolving under pervasive diversifying selection.

Interestingly, the detection of diversifying selection at codon 3606 (nsp6) was significantly supported by three different tests. Also, the selection of this site was observed in isolates from all three phylogenetic groups, thus supporting the reliability of this finding. However, the conservative nature of the amino acid substitution at this site (L-F; BSM62= 0) suggests this may not affect the phenotype of SARS-CoV-2. The same pattern was observed in codon 75 (D-E; BSM62= 2).

Conversely, based on the nature of the substitution (G-V; BSM62= -3), diversifying selection at codon 251 of the ORF3a gene may produce a biologically relevant effect on the phenotype of SARS-CoV-2. The same situation was observed at codon 2244 (I-T; BSM62= -1). Interestingly, change at codon position 251 is highly conserved within isolates of group C, suggesting that this change might have promoted the divergence of this group.

Detection of directional selection

To maximize the inference of potential sites experiencing positive selection, amino acid alignments of SARS-CoV-2 were analyzed using the DEPS algorithm. Overall, DEPS identified a total of 4 amino acid residues that are experiencing directional selection. Of these four residues, isoleucine (I) has the strongest bias, affecting 16 out of 19 sites evolving via directional selection (figure 5A).

The majority of selected sites were located in nonstructural proteins (nsp) encoded by the ORF1ab gene, with nsp3 accounting for the highest proportion (Figure 5B). Interestingly, a low proportion of the total number of predicted sites had a conservative amino acid substitution (residues at positions 902, 1769, 2235, and 2908), suggesting that the majority of substitutions may have an adaptive effect. In this context, it is remarkable that codon 84 of protein ORF8a is synapomorphic in all Group B sequences. Also, similar to previous algorithms, DEPS identified positive selection of residue 251 of ORF3a, supporting the potential significance of this site in the early evolution of SARS-CoV-2.

Evidence of coevolution among sites

Finally, we attempted to find coevolutionary correlations between different codons within the genome that result in the positive selection of sites. Analysis by BMG produced evidence of 14 coevolving codon pairs; these interactions took place mostly within codons located within the ORFB1ab gene (Figure 6). Although most of the interactions were detected between nonsynonymous codons, coevolution between codons 4090-4269 and 818-4320 was detected by a synonymous substitution at one of the codons. Also, based on the nature of the amino acid replacement, just 6 of the 14 interactions produced a potential biological significance in the replacement. Interestingly, 8 of the 14 interactions appeared associated with sites evolving under some type of positive selection, suggesting that the selection of these sites might be the result of epistatic events (Figure 7).

Discussion

Herein, we evaluated the phylogenetic and evolutionary dynamics of SARS-CoV-2 during the first months of the pandemic event in 2020. Our phylogenetic analysis revealed the complex dynamic of the spread of infection throughout the world, suggesting that multiple viral variants might have emerged in China before the start of the pandemic event. The evolutionary rate calculated for SARS-CoV-2 in this study was consistent not only with previous reports for SARS-CoV (Salemi et al., 2004, Zhao et al., 2004), but also with the rate for other RNA viruses (Sanjuan et al., 2010), explaining the high levels of identity at nucleotide and amino acid levels calculated for SARS-CoV-2 in our study. In this context, the high conservation observed in the genome of SARS-CoV-2 early during the pandemic might also be attributed to the unique RNA correction machinery of coronaviruses (Ferron et al., 2018).

However, and despite the relative genome stability observed in SARS-CoV-2 at this stage of the pandemic, we were able to describe the existence of 3 phylogenetic groups. Interestingly, our evolutionary analysis supported the hypothesis regarding early divergence events produced during the pandemic. By using a combination of different evolutionary algorithms, we detected multiple codon sites that may be promoting the divergence of SARS-CoV-2. However, two primary considerations must be addressed regarding the biological relevance of multiple sites detected in this study. First, a considerable number of polymorphisms were detected in just one viral isolate, which might be a consequence of the small number of viral isolates available at the time we started this study. Another possibility is that some of the polymorphisms might be due to sequencing errors. Additionally, some positively selected sites resulted in conservative replacements in terms of the similar nature of these amino acids, suggesting that they may not have a potential biological effect on the phenotype of SARS-CoV-2. However, the last assumption must be taken with caution, since experimental work on Chikungunya virus has shown that a conservative replacement of aspartic acid (D) with glutamic acid (E) at residue 350 of the E1 protein is able to alter the affinity of a monoclonal antibody for this site (Tuekprakhon et al., 2018).

We are reporting three potentially relevant residues that may be driving the evolution of SARS-CoV-2 in human populations. Residue 251 of the ORF3a protein appears to be experiencing diversifying selection and might be related to the emergence of viruses in phylogenetic Group C. The early selection of this site might have a biological relevance since the ORF3a protein has been associated with virulence of human coronaviruses by controlling not only the expression of cytokines and chemokines but also inducing necrotic cell death (Shi et al., 2019).

Also, a residue located at position 84 of the ORF8a protein was found to be evolving under directional selection and might be related to the emergence of Group B. This protein has been implicated in viral pathogenesis by regulating the initial innate response (McBride and Fielding, 2012, Shi et al., 2019). Interestingly, ORF3a and ORF8a accessory proteins had the highest dN/dS, indicating that a positive selective pressure is being exerted on these proteins during the pandemic.

Based on the known roles in tropism and virulence associated with the helicase of MERS-CoV, we consider it important to mention codon 5865 (ORF1ab) that was found to be evolving under directional selection and might be related to the divergence of 5 isolates from Washington, USA, forming a sub-cluster in Group B.

Finally, our analysis of coevolution revealed some potential epistatic interactions that might be driving the evolution of SARS-CoV-2. This mechanism has been proposed to explain the emergence of an Ebola virus variant in 2014 (Ibeh et al., 2016), and its relevance in the evolution of coronaviruses should be explored in future studies. Also, it is interesting to mention that most of the co-evolving sites were located in nsp3; given the role of this protein in the virulence of coronaviruses (Fehr et al., 2015), this observation may be key in understanding the evolution of SARS-CoV-2. Furthermore, since two of the interactions detected by BGM were associated with synonymous mutations, the relevance of this type of substitution to viral fitness should not be underestimated, since selection of synonymous substitutions has been reported in other RNA viruses like VSV (Novella et al., 2004, Velazquez-Salinas et al., 2018).

Collectively, our results describe the early evolutionary events of SARS-CoV-2 during the current pandemic and the findings may support the hypothesis that different variants of SARS-CoV-2 with disparate levels of virulence might be circulating in the world. This possibility might have an important impact on public health and measures to control the pandemic. Subsequent studies using reverse genetics will be needed to confirm the relevance of our findings connecting specific residue substitutions with different virus phenotypes.

Author Contributions

LV-S, and SZ conceived and designed the experiments. LV-S, SZ and SE performed the experiments. LV-S, SZ, SE, DG, IN, and MB analyzed the data. LV-S, SZ, SE, DG, IN, and MB wrote the manuscript.

Funding

This work was performed under USDA Research Service CRIS Project No. 8064-32000-060-00D.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

We would like to particularly like to thank Melanie Prarat for editing the manuscript. Also, we thank Dr. Peter Simmonds for the use of the SSE program used for pairwise distance analysis.

Figure Legends

Figure 1. Sample summary . Description of the 86 SARS-Cov-2 full-length genome sequences included in this study. All sequences were obtained form I from the GenBank database, accession number, genome length, isolate name, source, host and country of origin are provided. N/A indicates information not available.

Figure 2. Phylogeny of SARS-Cov-2. A) A MCC tree reconstructed using 86 SARS-Cov-2 full-length genomes collected from patients naturally infected at different countries, showing the existence of 3 phylogenetic groups (A, B, C). Branches in the tree are colored based on the geographic location of each isolate. Presumed index case is pointed out with a red arrow. B) Average p-distance, and average nucleotide and amino acid differences between group A and groups B and C sequences, respectively. Analysis was conducted using the software MEGA 7.

Figure 3. Pairwise distance analysis. Pairwise distance analysis at: A) synonymous and B) non-synonymous nucleotide sites was conducted using the program Sequence Distances (software SSE). Red bars represent pairwise distance comparisons using an sliding window of 50 nucleotides. Average nucleotide pairwise distance for different genes is shown at C) synonymous and D) non-synonymous sites. E) Fast-evolving synonymous and non-synonymous sites at each coding region are shown. For these sites evolutionary rates oscillated between 4.97 and 4.95. Red numbers represent nucleotides at : 1)Leader protein, 2) nsp2, 3) nsp3, 4) nsp4, 5) nsp6, 6) nsp7, 7)nsp8, 8)nsp10, 9) RNA independent polymerase, 10) helicase, 11) 3'to 5' exonuclease, 12) endoRNAse, 13) 2'-O-ribose methyltransferase.

Figure 4. Diversifying and purifying selection on SARS-CoV-2. A) General overview obtained by SLAC analysis, showing the evolutionary rate (dN-dS or dN/dS) along the genome and at individual genes of SARS-CoV-2. Statistically significant codons were inferred by multiple evolutionary tests used in this study. Red asterisks represent codons with significant evidence for selection, Codons evolving at B) Purifying (negative) or C) Diversifying (positive) selection are shown numbers in red represent evolutionary tests with significant values according to the analysis :SLAC, FEL, MEME (p-value=0.1) and FUBAR (posterior probability =0.9). The criteria for considering a site positively or negatively selected was based on their identification by at least one of the tests. The Phylogenetic group column (assigned according with figure 1A) shows also the isolates carrying the substitutions . Abbreviations: LP (Leader protein), 3LP (3C-like proteinase), n9 (nsp9), 3'-5' exo (3'-to 5' exonuclease), EN (endoRNAse), and 2'M (2'-o-ribose methyltransferase).

Figure 5. Directional selection analysis on SARS-CoV-2. A)An amino acid alignment was evaluated by DEPS and 4 different residues producing 19 directionally evolving sites in the proteome of SARS-CoV-2 are reported. P-values show the statistical significance of each residue considering a model test of selection versus not selection. Bias term: Alignment-wide relative rate of substitution towards target residue. , Proportion of affected sites: Percentage of sites evolving under a directional model, versus a standard model with no directionality. Directionally evolving sites: Number of sites that show evidence of directional selection for focal

residue. B) Description of 19 directionally evolving sites. Sites were detected by Empirical Bayesian Factor (EBF) considering a cut-off of 100 or more. Numbers in red represent replacements between amino acids with different properties. The Phylogenetic group column (assigned according with figure 1A) shows also the isolates carrying the substitutions.

Figure 6. Coevolution between codon pairs in the genome of SARS-CoV-2. BMG analysis was conducted to detect coevolving codon pairs. Evidence of 14 coevolving codon pairs were detected and the specific locations of those in the genome of SARS-CoV-2 are presented. Posterior probability of pair associations was supported by Markov Chain Monte Carlo Analysis at cut-off of 50 or more. Numbers in red represent replacements between amino acids with different properties. The Phylogenetic group column (assigned according with figure 1A) shows also the isolates carrying the substitutions. *[1] represents viral isolated where the changes were not detected.

Figure 7. Potential epistatic network in the genome of SARS-CoV-2. Coevolving pair codons involving sites evolving under positive selection at different regions in the genome of SARS-CoV-2 are shown with stars, suggesting that the selection of these sites may be result of epistasis.

References

ANDERSEN, K. G., RAMBAUT, A., LIPKIN, W. I., HOLMES, E. C. & GARRY, R. F. 2020. The proximal origin of SARS-CoV-2. *Nature Medicine*.

CORONAVIRIDAE STUDY GROUP OF THE INTERNATIONAL COMMITTEE ON TAXONOMY OF, V. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol*.

DELPORT, W., POON, A. F., FROST, S. D. & KOSAKOVSKY POND, S. L. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics,* 26**,** 2455-7.

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. & RAMBAUT, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol,* 29**,** 1969-73.

FEHR, A. R., ATHMER, J., CHANNAPPANAVAR, R., PHILLIPS, J. M., MEYERHOLZ, D. K. & PERLMAN, S. 2015. The nsp3 macrodomain promotes virulence in mice with coronavirus-induced encephalitis. *J Virol,* 89**,** 1523-36.

FERRON, F., SUBISSI, L., SILVEIRA DE MORAIS, A. T., LE, N. T. T., SEVAJOL, M., GLUAIS, L., DECROLY, E., VONRHEIN, C., BRICOGNE, G., CANARD, B. & IMBERT, I. 2018. Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. *Proc Natl Acad Sci U S A,* 115**,** E162-E171.

FROST, S. D. W., MAGALIS, B. R. & KOSAKOVSKY POND, S. L. 2018. Neutral Theory and Rapidly Evolving Viral Pathogens. *Mol Biol Evol,* 35**,** 1348-1354.

HENIKOFF, S. & HENIKOFF, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A,* 89**,** 10915-9.

IBEH, N., NSHOGOZABAHIZI, J. C. & ARIS-BROSOU, S. 2016. Both Epistasis and Diversifying Selection Drive the Structural Evolution of the Ebola Virus Glycoprotein Mucin-Like Domain. *J Virol,* 90**,** 5475-5484.

KOSAKOVSKY POND, S. L. & FROST, S. D. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol,* 22**,** 1208-22.

KOSAKOVSKY POND, S. L., POON, A. F., LEIGH BROWN, A. J. & FROST, S. D. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol,* 25**,** 1809-24.

KOSAKOVSKY POND, S. L., POSADA, D., GRAVENOR, M. B., WOELK, C. H. & FROST, S. D. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics,* 22**,** 3096-8.

KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33**,** 1870-4.

MCBRIDE, R. & FIELDING, B. C. 2012. The role of severe acute respiratory syndrome (SARS)-coronavirus accessory proteins in virus pathogenesis. *Viruses,* 4**,** 2902-23.

MURRELL, B., MOOLA, S., MABONA, A., WEIGHILL, T., SHEWARD, D., KOSAKOVSKY POND, S. L. & SCHEFFLER, K. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol,* 30**,** 1196-205.

MURRELL, B., WERTHEIM, J. O., MOOLA, S., WEIGHILL, T., SCHEFFLER, K. & KOSAKOVSKY POND, S. L. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet,* 8**,** e1002764.

NOVELLA, I. S., ZARATE, S., METZGAR, D. & EBENDICK-CORPUS, B. E. 2004. Positive selection of synonymous mutations in vesicular stomatitis virus. *J Mol Biol,* 342**,** 1415-21.

PHAN, M. V. T., NGO TRI, T., HONG ANH, P., BAKER, S., KELLAM, P. & COTTEN, M. 2018. Identification and characterization of Coronaviridae genomes from Vietnamese bats and rats based on conserved protein domains. *Virus Evol,* 4**,** vey035.

POON, A. F., LEWIS, F. I., POND, S. L. & FROST, S. D. 2007. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput Biol,* 3**,** e231.

RAMBAUT, A., DRUMMOND, A. J., XIE, D., BAELE, G. & SUCHARD, M. A. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol,* 67**,** 901-904.

SALEMI, M., FITCH, W. M., CICCOZZI, M., RUIZ-ALVAREZ, M. J., REZZA, G. & LEWIS, M. J. 2004. Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. *J Virol,* 78**,** 1602-3.

SANJUAN, R., NEBOT, M. R., CHIRICO, N., MANSKY, L. M. & BELSHAW, R. 2010. Viral mutation rates. *J Virol,* 84**,** 9733-48.

SHI, C. S., NABAR, N. R., HUANG, N. N. & KEHRL, J. H. 2019. SARS-Coronavirus Open Reading Frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes. *Cell Death Discov,* 5**,** 101.

SIMMONDS, P. 2012. SSE: a nucleotide and amino acid sequence analysis platform. *BMC Res Notes,* 5**,** 50.

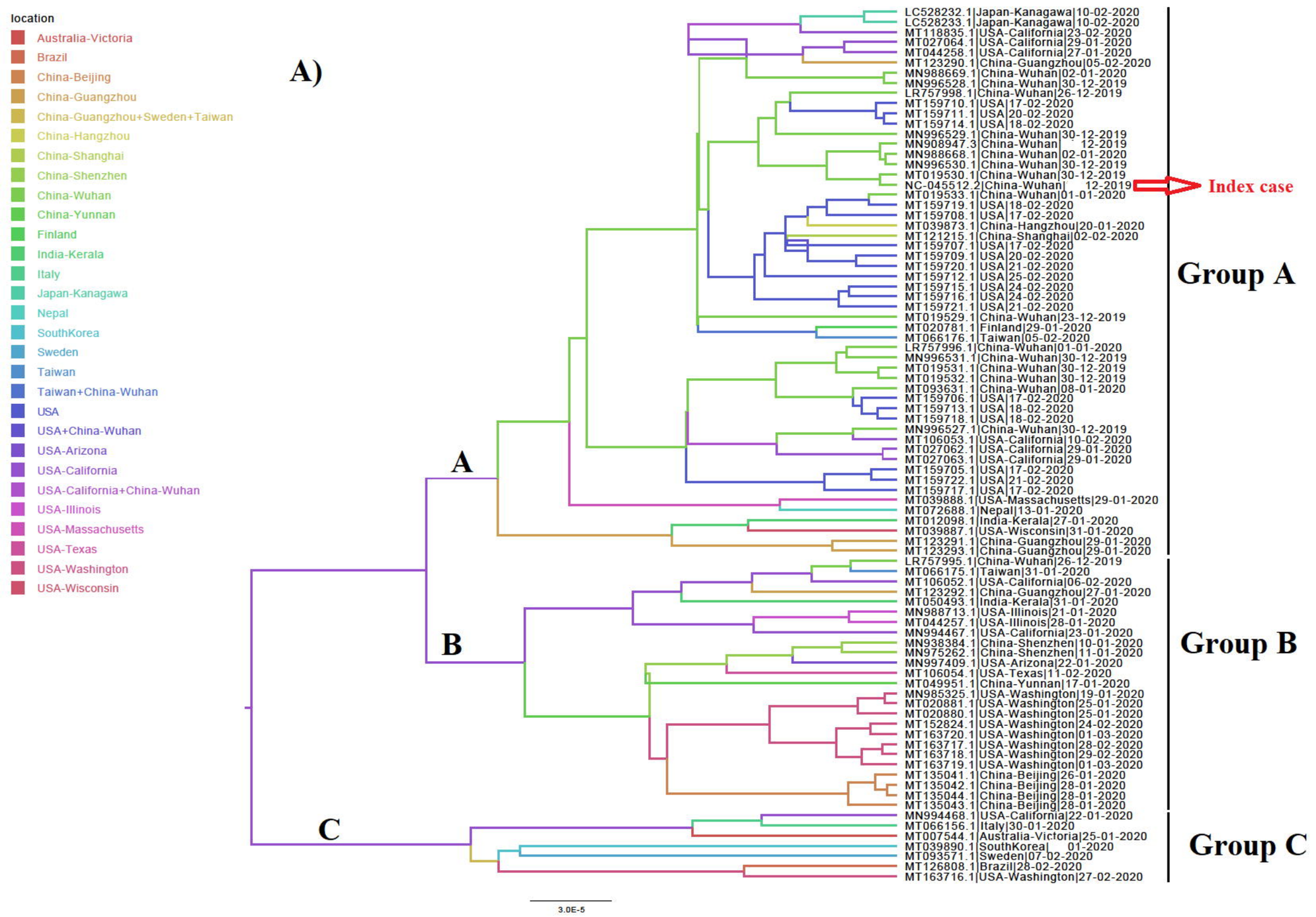
SUN, P., QIE, S., LIU, Z., REN, J., LI, K. & XI, J. 2020. Clinical characteristics of hospitalized patients with SARS-CoV-2 infection: A single arm meta-analysis. *J Med Virol.*

TUEKPRAKHON, A., NAKAYAMA, E. E., BARTHOLOMEEUSEN, K., PUIPROM, O., SASAKI, T., HUITS, R., LUPLERTLOP, N., KOSOLTANAPIWAT, N., MANEEKAN, P., ARIEN, K. K., SHIODA, T. & LEAUNGWUTIWONG, P. 2018. Variation at

455 position 350 in the Chikungunya virus 6K-E1 protein determines the sensitivity of
456 detection in a rapid E1-antigen test. *Sci Rep,* 8**,** 1094.
457 VELAZQUEZ-SALINAS, L., PAUSZEK, S. J., STENFELDT, C., O'HEARN, E. S.,
458 PACHECO, J. M., BORCA, M. V., VERDUGO-RODRIGUEZ, A., ARZT, J. &
459 RODRIGUEZ, L. L. 2018. Increased Virulence of an Epidemic Strain of Vesicular
460 Stomatitis Virus Is Associated With Interference of the Innate Response in Pigs. *Front*
461 *Microbiol,* 9**,** 1891.
462 WAN, Y., SHANG, J., GRAHAM, R., BARIC, R. S. & LI, F. 2020. Receptor Recognition by
463 the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural
464 Studies of SARS Coronavirus. *J Virol,* 94.
465 WEAVER, S., SHANK, S. D., SPIELMAN, S. J., LI, M., MUSE, S. V. & KOSAKOVSKY
466 POND, S. L. 2018. Datamonkey 2.0: A Modern Web Application for Characterizing
467 Selective and Other Evolutionary Processes. *Mol Biol Evol,* 35**,** 773-777.
468 WEISS, S. R. & NAVAS-MARTIN, S. 2005. Coronavirus pathogenesis and the emerging
469 pathogen severe acute respiratory syndrome coronavirus. *Microbiol Mol Biol Rev,* 69**,**
470 635-64.
471 WU, F., ZHAO, S., YU, B., CHEN, Y. M., WANG, W., SONG, Z. G., HU, Y., TAO, Z. W.,
472 TIAN, J. H., PEI, Y. Y., YUAN, M. L., ZHANG, Y. L., DAI, F. H., LIU, Y., WANG, Q.
473 M., ZHENG, J. J., XU, L., HOLMES, E. C. & ZHANG, Y. Z. 2020. A new coronavirus
474 associated with human respiratory disease in China. *Nature,* 579**,** 265-269.
475 ZHAO, Z., LI, H., WU, X., ZHONG, Y., ZHANG, K., ZHANG, Y. P., BOERWINKLE, E. &
476 FU, Y. X. 2004. Moderate mutation rate in the SARS coronavirus genome and its
477 implications. *BMC Evol Biol,* 4**,** 21.

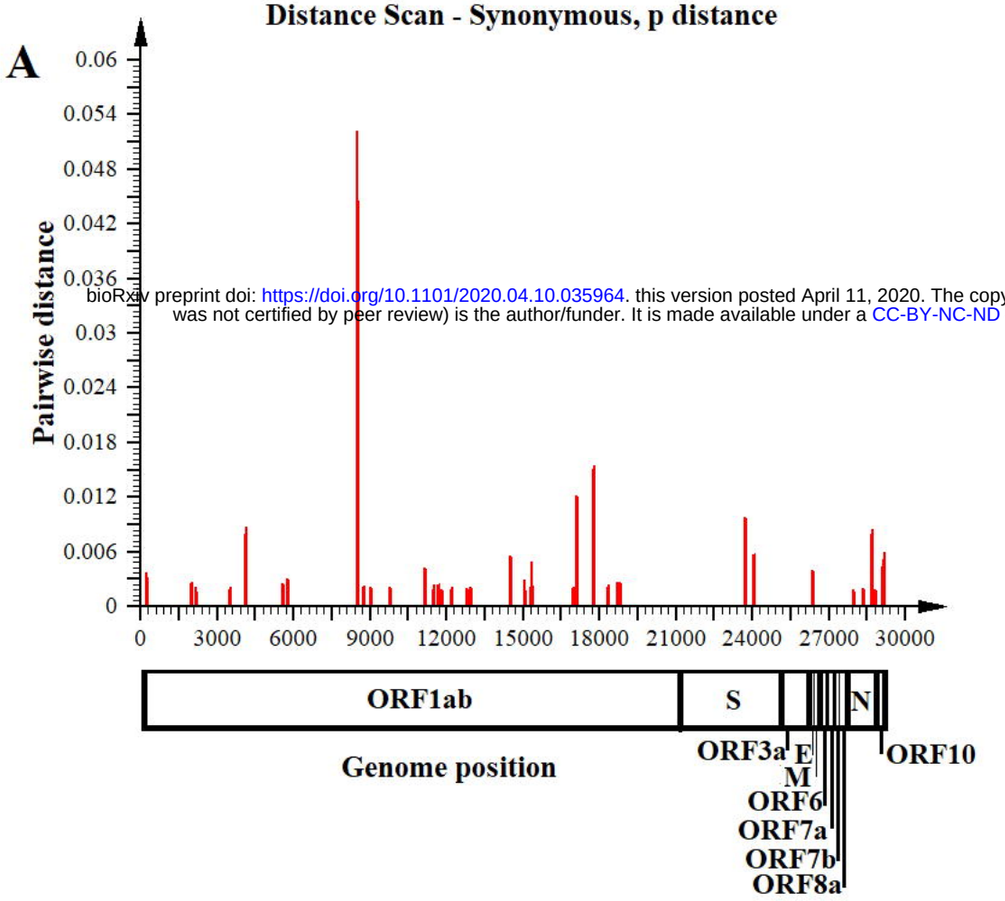
478

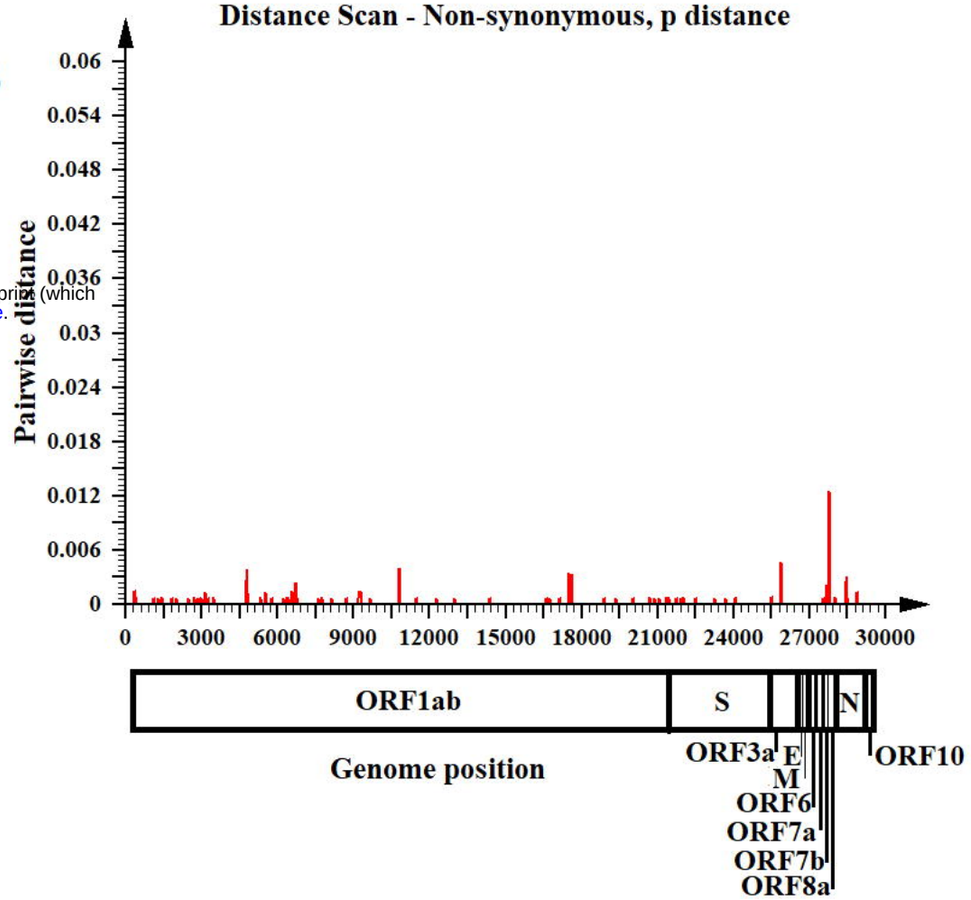
Genbank accession number	Genome bases	isolation source	Host	Country	State/City/Province	Collection Date
LC528232.1	29902	throat swab	Homo sapiens	Japan	Kanagawa	10-Feb-20
LC528233.1	29902	throat swab	Homo sapiens	Japan	Kanagawa	10-Feb-20
LR757995.1	29872	N/A	Homo sapiens	China	Wuhan	26-Dec-19
LR757996.1	29868	N/A	Homo sapiens	China	Wuhan	1-Jan-20
LR757998.1	29866	N/A	Homo sapiens	China	Wuhan	26-Dec-19
MN908947.3	29903	N/A	Homo sapiens	China	Wuhan	Dec-19
MN938384.1	29838	nasopharyngeal swab	Homo sapiens	China	Shenzhen	10-Jan-20
MN975262.1	29891	sputum	Homo sapiens	China	Shenzhen	11-Jan-20
MN985325.1	29882	oropharyngeal swab	Homo sapiens	USA	Washington	19-Jan-20
MN988668.1	29881	bronchoalveolar lavage fluid	Homo sapiens	China	Wuhan	2-Jan-20
MN988669.1	29881	bronchoalveolar lavage fluid	Homo sapiens	China	Wuhan	2-Jan-20
MN988713.1	29882	sputum	Homo sapiens	USA	Illinois	21-Jan-20
MN994467.1	29882	nasopharyngeal swab	Homo sapiens	USA	California	23-Jan-20
MN994468.1	29883	nasopharyngeal swab	Homo sapiens	USA	California	22-Jan-20
MN996527.1	29825	bronchoalveolar lavage fluid	Homo sapiens	China	Wuhan	30-Dec-19
MN996528.1	29891	bronchoalveolar lavage fluid	Homo sapiens	China	Wuhan	30-Dec-19
MN996529.1	29852	bronchoalveolar lavage fluid	Homo sapiens	China	Wuhan	30-Dec-19
MN996530.1	29854	bronchoalveolar lavage fluid	Homo sapiens	China	Wuhan	30-Dec-19
MN996531.1	29857	bronchoalveolar lavage fluid	Homo sapiens	China	Wuhan	30-Dec-19
MN997409.1	29882	buccal swab	Homo sapiens	USA	Arizona	22-Jan-20
MT007544.1	29893	N/A	Homo sapiens	Australia	Victoria	25-Jan-20
MT012098.1	29854	throat swab	Homo sapiens	India	Kerala	27-Jan-20
MT019529.1	29899	bronchoalveolar lavage fluid	Homo sapiens; male; age 65	China	Wuhan	23-Dec-19
MT019530.1	29889	bronchoalveolar lavage fluid	Homo sapiens; female; age 49	China	Wuhan	30-Dec-19
MT019531.1	29899	bronchoalveolar lavage fluid	Homo sapiens; male; age 41	China	Wuhan	30-Dec-19
MT019532.1	29890	bronchoalveolar lavage fluid	Homo sapiens; female; age 52	China	Wuhan	30-Dec-19
MT019533.1	29883	bronchoalveolar lavage fluid	Homo sapiens; male; age 61	China	Wuhan	1-Jan-20
MT020781.2	29806	N/A	Homo sapiens	Finland	N/A	29-Jan-20
MT020880.1	29882	nasopharyngeal swab	Homo sapiens	USA	Washington	25-Jan-20
MT020881.1	29882	oropharyngeal swab	Homo sapiens	USA	Washington	25-Jan-20
MT027062.1	29882	nasopharyngeal swab	Homo sapiens	USA	California	29-Jan-20
MT027063.1	29882	oropharyngeal swab	Homo sapiens	USA	California	29-Jan-20
MT027064.1	29882	oropharyngeal swab	Homo sapiens	USA	California	29-Jan-20
MT039873.1	29833	sputum	Homo sapiens; male	China	Hangzhou	20-Jan-20
MT039887.1	29879	nasopharyngeal swab	Homo sapiens	USA	Wisconsin	31-Jan-20
MT039888.1	29882	oropharyngeal swab	Homo sapiens	USA	Massachusetts	29-Jan-20
MT039890.1	29903	N/A	Homo sapiens	South Korea	N/A	Jan-20
MT044257.1	29882	sputum	Homo sapiens	USA	Illinois	28-Jan-20
MT044258.1	29858	respiratory swab	Homo sapiens	USA	California	27-Jan-20
MT049951.1	29903	sputum	Homo sapiens	China	Yunnan	17-Jan-20
MT050493.1	29851	throat swab	Homo sapiens	India	Kerala	31-Jan-20
MT066156.1	29867	sputum	Homo sapiens	Italy	N/A	30-Jan-20
MT066175.1	29870	N/A	Homo sapiens	Taiwan	N/A	31-Jan-20
MT066176.1	29870	N/A	Homo sapiens	Taiwan	N/A	5-Feb-20
MT072688.1	29811	oropharyngeal swab	Homo sapiens	Nepal	N/A	13-Jan-20
MT093571.1	29886	N/A	Homo sapiens	Sweden	N/A	7-Feb-20
MT093631.2	29860	throat swab	Homo sapiens	China	Wuhan	8-Jan-20
MT106052.1	29882	nasopharyngeal swab	Homo sapiens	USA	California	6-Feb-20
MT106053.1	29882	nasopharyngeal swab	Homo sapiens	USA	California	10-Feb-20
MT106054.1	29882	sputum	Homo sapiens	USA	Texas	11-Feb-20
MT118835.1	29882	bronchoalveolar lavage	Homo sapiens	USA	California	23-Feb-20
MT121215.1	29945	throat swab	Homo sapiens	China	Shanghai	2-Feb-20
MT123290.1	29891	oropharyngeal swab	Homo sapiens	China	Guangzhou	5-Feb-20
MT123291.2	29982	bronchoalveolar lavage fluid	Homo sapiens	China	Guangzhou	29-Jan-20
MT123292.2	29923	sputum	Homo sapiens	China	Guangzhou	27-Jan-20
MT123293.2	29871	stool	Homo sapiens	China	Guangzhou	29-Jan-20
MT126808.1	29876	nasopharyngeal swab	Homo sapiens	Brazil	N/A	28-Feb-20
MT135041.1	29903	N/A	Homo sapiens	China	Beijing	26-Jan-20
MT135042.1	29903	N/A	Homo sapiens	China	Beijing	28-Jan-20
MT135043.1	29903	N/A	Homo sapiens	China	Beijing	28-Jan-20
MT135044.1	29903	N/A	Homo sapiens	China	Beijing	28-Jan-20
MT152824.1	29878	mid-nasal swab	Homo sapiens	USA	Washington	24-Feb-20
MT159705.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	17-Feb-20
MT159706.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	17-Feb-20
MT159707.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	17-Feb-20
MT159708.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	17-Feb-20
MT159709.1	29882	oropharyngeal swab	Homo sapiens	USA	N/A	20-Feb-20
MT159710.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	17-Feb-20
MT159711.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	20-Feb-20
MT159712.1	29882	oropharyngeal swab	Homo sapiens	USA	N/A	25-Feb-20
MT159713.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	18-Feb-20
MT159714.1	29882	nasopharyngeal swab	Homo sapiens	USA	USA	18-Feb-20
MT159715.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	24-Feb-20
MT159716.1	29687	nasopharyngeal swab	Homo sapiens	USA	N/A	24-Feb-20
MT159717.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	17-Feb-20
MT159718.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	18-Feb-20
MT159719.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	18-Feb-20
MT159720.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	21-Feb-20
MT159721.1	29882	oropharyngeal swab	Homo sapiens	USA	N/A	21-Feb-20
MT159722.1	29882	nasopharyngeal swab	Homo sapiens	USA	N/A	21-Feb-20
MT163716.1	29903	N/A	Homo sapiens	USA	Washington	27-Feb-20
MT163717.1	29897	N/A	Homo sapiens	USA	Washington	28-Feb-20
MT163718.1	29903	N/A	Homo sapiens	USA	Washington	29-Feb-20
MT163719.1	29903	N/A	Homo sapiens	USA	Washington	1-Mar-20
MT163720.1	29732	N/A	Homo sapiens	USA	Washington	1-Mar-20
NC_045512.2	29903	N/A	Homo sapiens	China	Wuhan	Dec-19

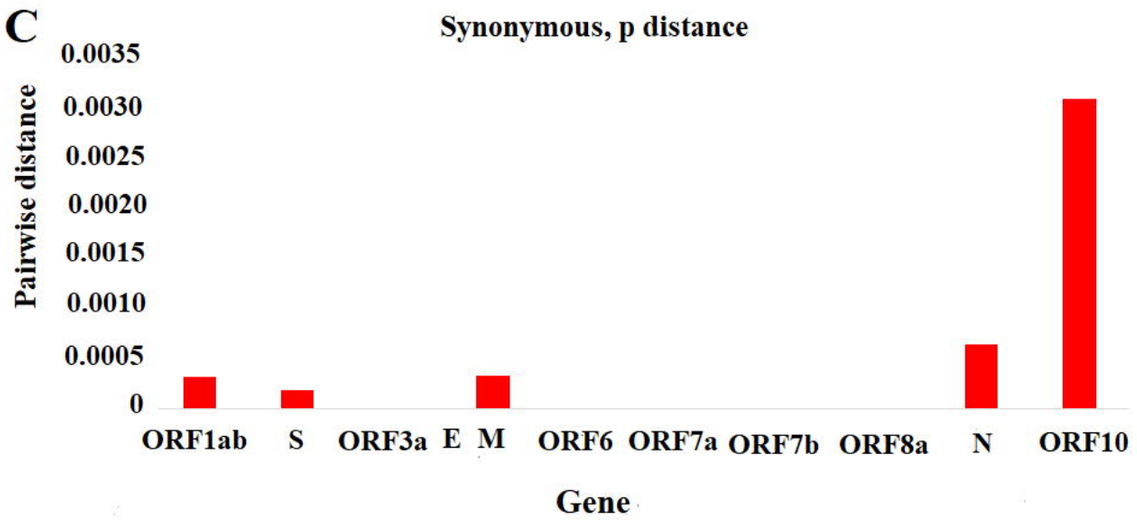


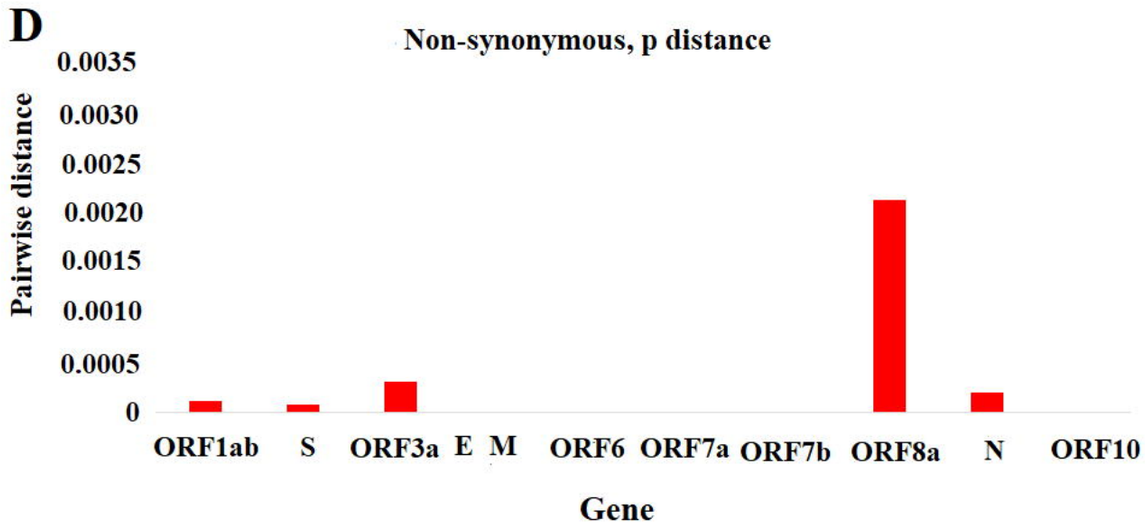
B)

	Average p-distance	Average nucleotide differences	Average amino acid differences
Group B	0.000128	3.75	2
Group C	0.000191	5.78	3.67



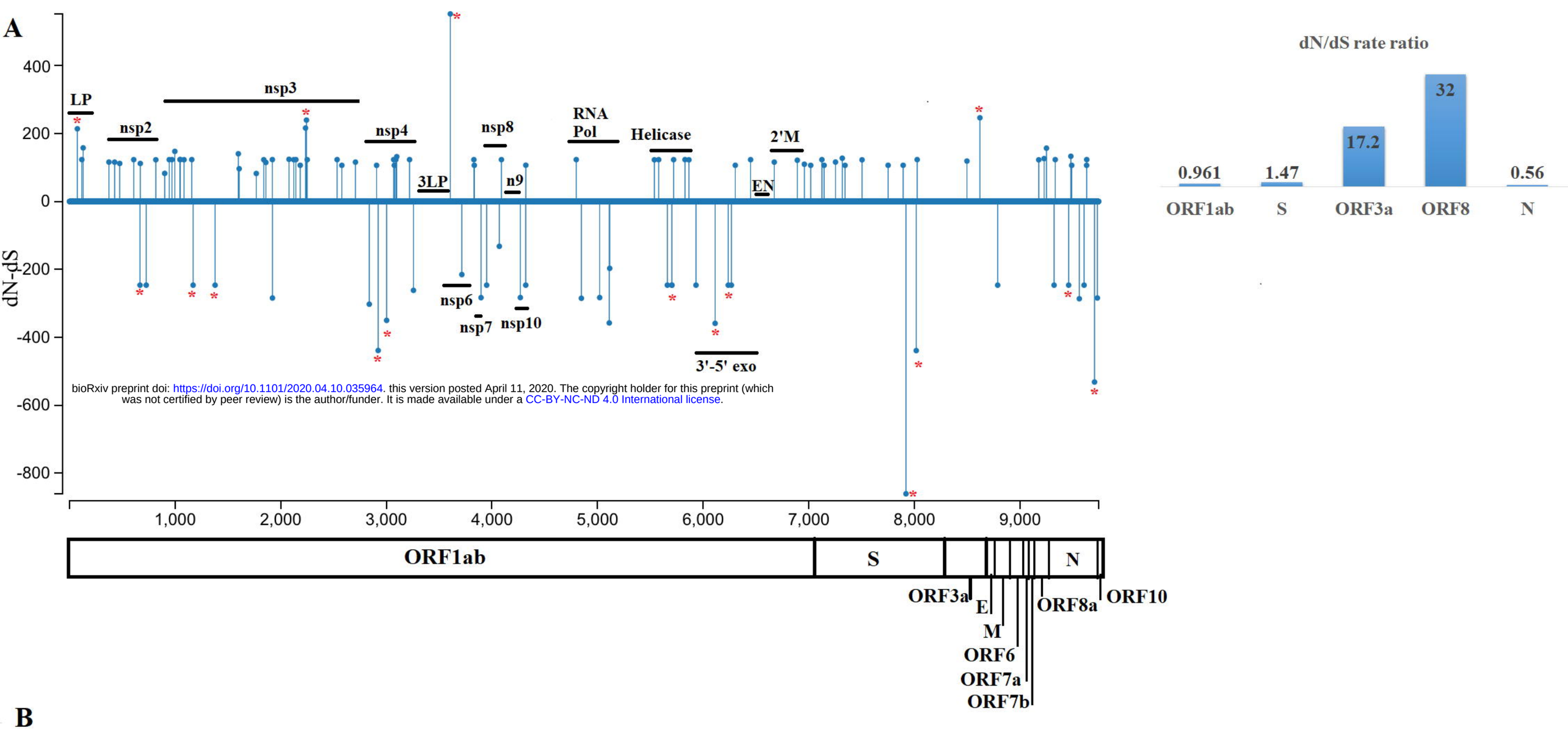






E

Gene	Total of evolving sites >1	# of synonymous sites	Gene nucleotide position	# of non-synonymous sites	Gene nucleotide position
ORF1ab	82	28	(349)[1], (2004, 2181)[2], (3513, 4137, 5766)[3], (8517, 8769, 9009, 9771)[4], (11145)[5],(11691)[6], (11850,12208)[7],(12807, 12960)[8], (14541, 15060, 15333, 15343)[9], (16983, 17109, 17112)[10], (17796, 18339,18711,18801)[11]	54	(225,389)[1], (1120, 1283, 1426, 1826, 2012, 2452)[2], (2706,2834, 2912, 2994, 3141, 3146, 3253, 3473, 4797, 4819, 5307, 5519, 5580, 5761, 6236, 6371, 6430, 6554, 6703, 6731, 6751, 7601, 7736, 8123)[3], (8722, 9226, 9269, 9659)[4], (10818, 11485, 11499)[5], (12269)[7], (12961)[8], (14393)[9], (16613, 16736, 17159, 17483, 17594)[10], (18911, 19346)[11], (20017)[12], (20672, 20878, 21052)[13]
S	11	3	471, 2472, 2763	8	82, 145, 662, 741, 1223, 1963, 2390, 2789
ORF3a	2	0		2	383, 752
M	1	1	207	0	
ORF8a	3	0		3	32, 184, 251
N	10	5	105, 519, 822, 957	5	136, 581, 605, 1028, 1030
ORF10	1	1	78	0	



Gene	Gene codon position	SLAC dN-dS	SLAC P [dN/dS < 1]	FEL dN/dS	FEL p-value	FUBAR dN-dS	FUBAR Prob[dS>dN]	Inferred substitution	Phylogenetic group
ORF1ab/ nsp2	668	-246.55	0.33	0	0.05	-18.84	0.88	$_{(V)}GCA_{85\ (V)}GCU_{1}$	Group C (MT066156.1)
ORF1ab/ nsp3	1171	-246.55	0.33	0	0.06	-18.81	0.88	$_{(T)}ACA_{85\ (T)}AGG_{1}$	Group A (MT019529.1)
ORF1ab/ nsp3	1379	-246.55	0.33	0	0.10	-18.15	0.88	$_{(L)}CUU_{82\ (L)}CUC_{4}$	Group B (MT135041.1,MT135042.1, MT13043.1, MT135044.1)
ORF1ab/nsp4	2923	-439.02	0.21	0	0.05	-19.03	0.88	$_{(K)}AAA_{85\ (K)}AAG_{1}$	Group A (MT066176.1)
ORF1ab/nsp4	3003	-350.33	0.26	0	0.05	-18.99	0.88	$_{(R)}AGA_{85\ (R)}AGG_{1}$	Group C (MT093571.1)
ORF1ab/helicase	5704	-246.55	0.33	0	0.06	-18.80	0.88	$_{(A)}GCC_{82\ (A)}GCU_{4}$	Group A (MT012098.1, MT039887.1, MT123291.1) Group B (MT123292.1)
ORF1ab/3'-to-5' exonuclease	6113	-359.07	0.23	0	0.09	-18.48	0.88	$_{(H)}CAU_{85\ (H)}CAC_{1}$	Group B(MT106054.1)
ORF1ab/3'-to-5' exonuclease	6267	-246.55	0.33	0	0.09	-18.18	0.88	$_{(P)}CCU_{85\ (P)}CCC_{1}$	Group C (MT007544.1)
S	824	-861.17	0.02	0	0.07	-39.51	0.98	$_{(N)}AAC_{81\ (N)}AAU_{5}$	Group A (MT039888.1, MT072688.1) Group B (MN988713.1, MN994467.1, MT044257.1)
S	921	-439.36	0.21	0	0.05	-19.03	0.88	$_{(K)}AAA_{84\ (K)}AAG_{2}$	Group A (MT106053.1, MN996527.1)
N	173	-246.55	0.33	0	0.05	-18.85	0.88	$_{(A)}GCA_{85\ (A)}GCU_{1}$	Group B (MN994467.1)
N	418	-531.95	0.22	0	0.10	-18.69	0.88	$_{(Q)}CAG_{84\ (Q)}CAA_{2}$	Group A (MT123291.1) Group B (MT123292.1)

C

Gene	Gene codon position	SLAC dN-dS	SLAC P [dN/dS < 1]	FEL dN/dS	FEL p-value	MEME dN/neutral evolution component	MEME p-value	FUBAR dN-dS	FUBAR Prob[dS<dN]	Inferred substitution	Phylogenetic group
ORF1ab/leader protein	75	214.12	0.60	Infinity	0.25	293.63	0.26	31.52	0.94	$_{(D)}GAU_{84\ (E)}GAA_{2}$	Group B (MN988713.1, MT04427.1)
ORF1ab/nsp3	2244	239.71	0.47	Infinity	0.10	480.91	0.12	31.52	0.95	$_{(I)}AUC_{84\ (T)}ACC_{2}$	Group A (MT019531.1, MT123293.1)
ORF1ab/nsp6	3606	552.28	0.14	Infinity	0.08	3129.69	0.10	41.36	0.97	$_{(L)}UUG_{80\ (F)}UUU_{5\ (F)}UUC_{1}$*	Group A (LC528232.1, LC5282233.1) Group B (MT049951.1*) Group C (MT163716.1, MT049951.1)
ORF3a	251	246.55	0.44	Infinity	0.19	517.70	0.21	31.38	0.95	$_{(G)}GGU_{79\ (V)}GUU_{7}$	Group C (all 7 taxon)

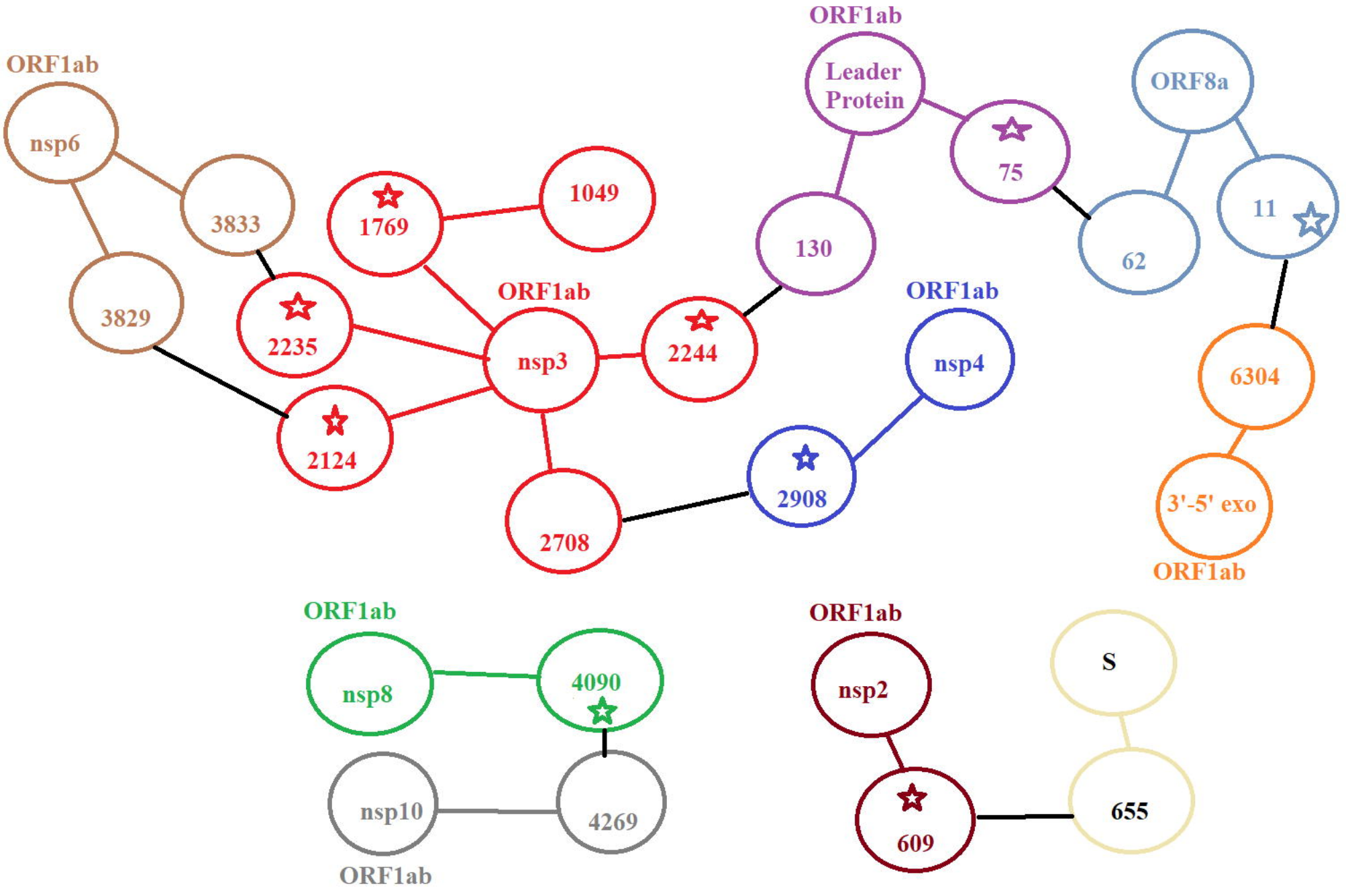
A

Residue	p-value	Bias term	Proportion of affected sites	Directionally evolving sites
C	0.0028	10.89	50.00%	1
I	0	716.43	0.69%	16
S	0	47.28	1.71%	1
V	0	24.36	3.80%	1

B

Protein	Site	Composition in the alignment	Reconstructed most common ancestor at site	Inferred Substitutions	DEPS EBF	Blosum 62 Matrix Score	Phylogenetic group
ORF1ab/nsp2	**609**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 375.3	-1	Group A (MT027064.1)
ORF1ab/nsp3	**902**	$M_{85}I_1$	M	$I_1 \leftrightarrow_0 M$	I: 224.4	1	Group C (MT039890.1)
ORF1ab/nsp3	**945**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 374.2	-1	Group A (MT159717.1)
ORF1ab/nsp3	**1769**	$M_{85}I_1$	M	$I_1 \leftrightarrow_0 M$	I: 223.3	1	Group C (MT163716.1)
ORF1ab/nsp3	**1840**	$T_{84}I_2$	T	$I_2 \leftrightarrow_0 T$	$I:>10^5$	-1	Group B (MT152824.1, MT163720.1)
ORF1ab/nsp3	**2124**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 375.3	-1	Group A (MT159712.1)
ORF1ab/nsp3	**2185**	$S_{84}I_2$	S	$I_2 \leftrightarrow_0 S$	$I:>10^5$	-2	Group A (MT123291.1, MT123293.1)
ORF1ab/nsp3	**2235**	$L_{85}I_1$	L	$I_1 \leftrightarrow_0 L$	I: 367.5	2	Group A (MT019529.1)
ORF1ab/nsp4	**2908**	$F_{85}I_1$	F	$F_0 \leftrightarrow_1 I$	I: 558.4	1	Group A (MN996531.1)
ORF1ab/nsp4	**3090**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 375.3	-1	Group A (LR757998.1)
ORF1ab/nsp8	**4090**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 375.3	-1	Group A (MT123292.1)
ORF1ab/helicase	**5538**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 378.5	-1	Group B (MT050493.1)
ORF1ab/helicase	**5579**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 375.3	-1	Group C (MN994468.1)
ORF1ab/helicase	**5865**	$Y_{81}C_5$	Y	$C_5 \leftrightarrow_0 Y$	C: 695.1	-2	Group B (MT163717.1, MT163718.1, MT163719.1, MT163720.1, MT152824.1)
ORF1ab/ 3'-5' exonuclease	**6449**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 375.5	-1	Group A (MT123291.1)
S	**408**	$R_{85}I_1$	R	$I_1 \leftrightarrow_0 R$	I: 652.7	-3	Group A MT012098.1
ORF3a	**251**	$G_{79}V_7$	G	$G_0 \leftrightarrow_7 V$	$V:>10^5$	-3	Group C (all 7 taxon)
ORF8a	**11**	$T_{85}I_1$	T	$I_1 \leftrightarrow_0 T$	I: 376.3	-1	Group B (MT106054.1)
ORF8a	**84**	$L_{61}S_{25}$	L	$L_0 \leftrightarrow_{25} S$	$S:>10^5$	-2	Group B (all 25 taxon)

Gene site 1	Codon position	Gene site 2	Codon position	Probability that sites 1 and 2 are not conditionally independent	Inferred substitution site 1	Blosum 62 matrix score	Inferred substitution site 2	Blosum 62 matrix score	Phylogenetic group
ORF1ab/leader protein	75	ORF8a	62	0.5	$_{(D)}GAU_{84\ (E)}GAA_{2}$	2	$_{(V)}GUG_{83\ (L)}CUG_{3}$	1	Group B (MT044257.1, MN988713.1) MN994467.1*[1]
ORF1ab/leader protein	117	ORF1ab/nsp3	1607	0.76	$_{(A)}GCU_{84\ (T)}ACU_{2}$	0	$_{(I)}AUA_{84\ (V)}GUA_{2}$	3	Group A (MT027062.1, MT027063.1)
ORF1ab/leader protein	130	ORF1ab/nsp3	2244	0.56	$_{(G)}GGA_{85\ (E)}GAA_{1}$	-2	$_{(I)}AUC_{83\ (T)}ACC_{2}$	-1	Group A (MT123293.1) MT019353.1*[1]
ORF1ab/nsp2	609	S	49	0.74	$_{(T)}ACU_{85\ (I)}AUU_{1}$	-1	$_{(H)}CUA_{85\ (Y)}UAU_{1}$	2	Group A (MT027064.1)
ORF1ab/nsp2	818	ORF1ab/nsp10	4320	0.76	$_{(G)}GGU_{85\ (S)}AGU_{1}$	0	$_{(S)}UCC_{85\ (S)}UCG_{1}$	N/A	Group C (MT093571.1)
ORF1ab/nsp3	1047	S	655	0.73	$_{(E)}GAA_{85\ (D)}GAC_{1}$	2	$_{(H)}CAU_{85\ (Y)}UAU_{1}$	2	Group B (MT163720.1)
ORF1ab/nsp3	1049	ORF1ab/nsp3	1769	0.76	$_{(A)}GCU_{85\ (V)}GUU_{1}$	0	$_{(M)}AUG_{85\ (I)}AUU_{1}$	1	Group C (MT163716.1)
ORF1ab/nsp3	2124	ORF1ab/nsp6	3829	0.75	$_{(T)}ACU_{85\ (I)}AUU_{1}$	-1	$_{(L)}CUC_{85\ (I)}UUC_{1}$	2	Group A (MT159712.1)
ORF1ab/nsp3	2235	ORF1ab/nsp6	3833	0.74	$_{(L)}CUA_{85\ (I)}AUA_{1}$	2	$_{(N)}AAU_{85\ (K)}AAA_{1}$	0	Group B (LR757998.1)
ORF1ab/nsp3	2251	ORF1ab/2'-O-ribose methyltransferase	6958	0.74	$_{(G)}GGU_{85\ (S)}AGU_{1}$	0	$_{(K)}AAG_{85\ (R)}AGG_{1}$	2	Group A (MN99629.1)
ORF1ab/nsp3	2579	ORF1ab/nsp4	3090	0.72	$_{(D)}GAU_{85\ (A)}GCU_{1}$	-2	$_{(T)}ACU_{85\ (I)}AUU_{1}$	-1	Group A (MN996531.1)
ORF1ab/nsp3	2708	ORF1ab/nsp4	2908	0.73	$_{(N)}AAC_{85\ (S)}AGC_{1}$	1	$_{(F)}UUU_{85\ (I)}AUU_{1}$	0	Group A (MT019529.1)
ORF1ab/nsp8	4090	ORF1ab/nsp10	4269	0.73	$_{(T)}ACU_{85\ (I)}AUU_{1}$	-1	$_{(F)}UUC_{85\ (F)}UUU_{1}$	N/A	Group B (MT123292.1)
ORF1ab/3'-to-5' exonuclease	6304	ORF8a	11	0.73	$_{(D)}GAU_{85\ (A)}GCU_{1}$	-2	$_{(T)}ACA_{85\ (I)}AUA_{1}$	-1	Group B (MT106054.1)


ORF1ab
nsp6
3833
3829
1769
1049
2235
ORF1ab
nsp3
2244
2124
2708
130
ORF1ab
Leader Protein
75
ORF8a
62
11
6304
3'-5' exo
ORF1ab
ORF1ab
nsp4
2908
ORF1ab
nsp8
4090
nsp10
4269
ORF1ab
ORF1ab
nsp2
609
S
655