RESEARCH

Open Access

The 29-nucleotide deletion in SARS-CoV: truncated versions of ORF8 are under purifying selection

Anastassia Bykova[1], Andreu Saura[1], Galina V. Glazko[2], Abiel Roche-Lima[3], Vyacheslav Yurchenko[1*] and Igor B. Rogozin[4*]

Abstract

Background Accessory proteins have diverse roles in coronavirus pathobiology. One of them in SARS-CoV (the causative agent of the severe acute respiratory syndrome outbreak in 2002–2003) is encoded by the open reading frame 8 (*ORF8*). Among the most dramatic genomic changes observed in SARS-CoV isolated from patients during the peak of the pandemic in 2003 was the acquisition of a characteristic 29-nucleotide deletion in *ORF8*. This deletion cause splitting of *ORF8* into two smaller *ORF*s, namely *ORF8a* and *ORF8b*. Functional consequences of this event are not entirely clear.

Results Here, we performed evolutionary analyses of *ORF8a* and *ORF8b* genes and documented that in both cases the frequency of synonymous mutations was greater than that of nonsynonymous ones. These results suggest that *ORF8a* and *ORF8b* are under purifying selection, thus proteins translated from these *ORF*s are likely to be functionally important. Comparisons with several other SARS-CoV genes revealed that another accessory gene, *ORF7a*, has a similar ratio of nonsynonymous to synonymous mutations suggesting that ORF8a, ORF8b, and ORF7a are under similar selection pressure.

Conclusions Our results for SARS-CoV echo the known excess of deletions in the *ORF7a-ORF7b-ORF8* complex of accessory genes in SARS-CoV-2. A high frequency of deletions in this gene complex might reflect recurrent searches in "functional space" of various accessory protein combinations that may eventually produce more advantageous configurations of accessory proteins similar to the fixed deletion in the SARS-CoV *ORF8* gene.

Keywords Rearrangements, Evolution, Natural selection, Protein truncation, Gene regulation, Recombination

*Correspondence:
Vyacheslav Yurchenko
vyacheslav.yurchenko@osu.cz
Igor B. Rogozin
rogozin@ncbi.nlm.nih.gov
[1]Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava 710 00, Czech Republic
[2]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA
[3]Center for Collaborative Research in Health Disparities–RCMI Program, Medical Sciences Campus, University of Puerto Rico, San Juan, PR 00936, USA
[4]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Background

The severe acute respiratory syndrome coronavirus (SARS-CoV) genome is a ~ 30 kb long, single-stranded, positive RNA molecule with the gene organization typical of coronaviruses including that of infamous SARS-CoV-2. There are 12 open reading frames (*ORF*s; hereafter gene names are Italicized, protein names are capitalized) that encode 26 proteins: 16 non-structural proteins (NSP1 to NSP16), four structural proteins (M, N, S, and E), and six accessory proteins (3a, 6, 7a, 7b, 8, 10) [1]. As a rule, accessory proteins have diverse functions in coronavirus pathobiology [2, 3]. They are usually dispensable for replication in cell culture, but appear to have regulatory roles during the viral cycle and, thus, likely contribute to the virus fitness by increasing its ability to evade the human innate immune response [4–7]. Different groups of coronaviruses usually differ in those accessory proteins and more infective species have specific pathogenic features [8, 9]. One of such proteins is encoded by the unique to the SARS-CoV lineage *ORF8* [2]. The intact *ORF8* (present in viruses of animals and some early-in-the-pandemic human isolates) encodes a 123 amino-acid polypeptide, consisting of an N-terminal signal sequence followed by the predicted Ig-like and transmembrane domains. Notably, *ORF7a* and *ORF8* genes have similar length and domain architecture suggesting similarities of their functions [10, 11]. It has been posited that the cleavable signal sequence directs the ORF8 precursor to the endoplasmic reticulum (ER) and mediates its translocation into the lumen [12, 13]. The cleaved SARS-CoV ORF8 protein became N-glycosylated, assembled into disulfide-linked homomultimeric complexes, and remained stable in the ER [12]. Another study reported that this protein induces ATF6-dependent transcription triggering the expression of chaperones and leading to attenuation of the protein translation level, thus, modifying the unfolded protein response [13].

The sequence identity of SARS-CoV and SARS-CoV2 *ORF8*s is about 40% implying that ORF8 is a relatively fast evolving in comparison to other viral proteins. The second fastest evolving gene is *ORF6* with 70% identity. Identity for other genes varies between 72% and 95%: for example, *ORF3a*, *N*, and spike (*S*) genes have 72, 91, and 76% identity, respectively [14]. Although *ORF8* is the fastest evolving gene, there is no doubt that it evolves under strong purifying selection (natural selection acting against deleterious mutations) [2, 14, 15].

The SARS-CoV and SARS-CoV-2 share almost identical gene architecture except for the *OFR8* gene [1]. One of the most striking and dramatic genomic changes observed in the SARS-CoV isolates from humans during the peak of the pandemic in 2003, most likely soon after its zoonotic transmission from palm civets, was acquisition of the 29-nucleotide deletion of *ORF8*, which splits *ORF8* into *ORF8a* and *ORF8b* (with a frameshift of 35 bp) encoding 39- and 84-residue polypeptides, respectively (Fig. 1A). The functional role of SARS-CoV ORF8 and consequences of the 29-nucleotide deletion are not entirely clear [15, 16]. For example, it was suggested that replication of the SARS-CoV is affected in cells that overexpress the protein encoded by *ORF8a* [16, 17]. This protein is likely to remain in the cytoplasm, as it is too small for its signal sequence to function, and will, therefore, be directly released from the ribosome [12]. A soluble, unmodified and monomeric ORF8b protein is also present in the cytoplasm. Yet, it is highly unstable and get rapidly degraded [16]. This protein, when overexpressed, induces apoptosis and gets involved in cellular degradation of the viral envelope protein [17–19].

It has been widely speculated that the truncated products of ORF8a/b led to a modulation of pathogenicity and/or replication that favored adaptation of SARS-CoV to humans [3, 20, 21]. Based on this hypothesis, it was suggested that further comprehensive genomics and structural-functional studies of *ORF8* and *ORF8a/b* are needed to reach a definitive conclusion about their function(s) that can eventually define their relevance to future therapeutics development [2].

Here, we performed evolutionary analyses of *ORF8a* and *ORF8b* genes and found that in both cases the number of synonymous mutations was greater than that for nonsynonymous ones. These results suggested that both *ORF8a* and *ORF8b* are under purifying selection implying that proteins translated from these *ORF*s are functional.

Methods

We used the term "*ORF8*" for sequences without the 29-nucleotide deletion, as well as "*ORF8a*" and "*ORF8b*" (outcomes of the 29-nucleotide deletion) according to the commonly used nomenclature [12]. We performed BLASTN searches with default parameters except for "Organism:" = HCoV-SARS (taxid:694,009) and "Maximum #seq" = 5000 using the "SARS coronavirus HSZ-Cc" sequence (GenBank Accession Number AY394995) as a query. For the SARS-CoV *ORF8* sequences, the last sequence was "SARS coronavirus isolate HC/SZ/266/03" (AY545916), whereas for the *ORF8a/b* sequences, the last sequence was "SARS coronavirus Urbani isolate icSARS-C7" (MK062183). The "last sequence" means a human-associated sequence with the highest *e*-value compared to other sequences, in which human is listed as the host, in the BLASTN output.

To estimate Ka and Ks, we extracted mutations from multiple alignments of *ORF8* and *ORF8a/b* sequences (Figs. S1 and S2, Additional file 1) using an "ad hoc" Python script (Additional file 2). We operated with raw numbers of nonsynonymous and synonymous mutations/

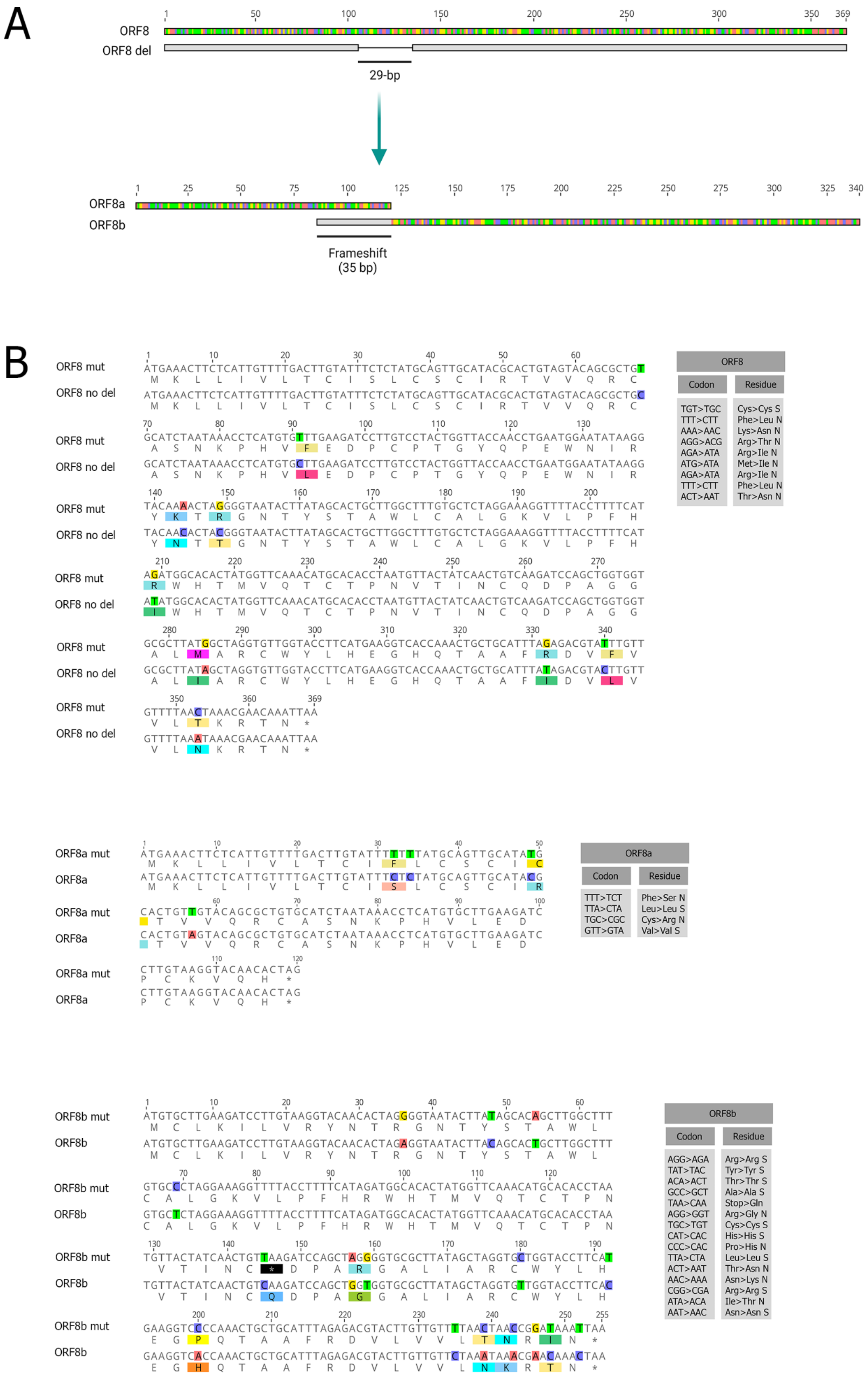

A
ORF8
ORF8 del
1 50 100 150 200 250 300 350 369
29-bp
ORF8a
ORF8b
1 25 50 75 100 125 150 175 200 225 250 275 300 325 340
Frameshift
(35 bp)
B
ORF8 mut
ATGAAACTTCTCATTGTTTTGACTTGTATTTCTCTATGCAGTTGCATACGCACTGTAGTACAGCGCTGT
M K L L I V L T C I S L C S C I R T V V Q R C
ORF8 no del
ATGAAACTTCTCATTGTTTTGACTTGTATTTCTCTATGCAGTTGCATACGCACTGTAGTACAGCGCTGC
M K L L I V L T C I S L C S C I R T V V Q R C
ORF8 mut
GCATCTAATAAACCTCATGTGTTTGAAGATCCTTGTCCTACTGGTTACCAACCTGAATGGAATATAAGG
A S N K P H V F E D P C P T G Y Q P E W N I R
ORF8 no del
GCATCTAATAAACCTCATGTGCTTGAAGATCCTTGTCCTACTGGTTACCAACCTGAATGGAATATAAGG
A S N K P H V L E D P C P T G Y Q P E W N I R
ORF8 mut
TACAAAACTAGGGGTAATACTTTATAGCACTGCTTGGCTTTGTGCTCTAGGAAAGGTTTTACCTTTTCAT
Y K T R G N T Y S T A W L C A L G K V L P F H
ORF8 no del
TACAACACTACGGGTAATACTTTATAGCACTGCTTGGCTTTGTGCTCTAGGAAAGGTTTTACCTTTTCAT
Y N T T G N T Y S T A W L C A L G K V L P F H
ORF8 mut
AGATGGCACACTATGGTTCAAACATGCACACCTAATGTTACTATCAACTGTCAAGATCCAGCTGGTGGT
R W H T M V Q T C T P N V T I N C Q D P A G G
ORF8 no del
ATATGGCACACTATGGTTCAAACATGCACACCTAATGTTACTATCAACTGTCAAGATCCAGCTGGTGGT
I W H T M V Q T C T P N V T I N C Q D P A G G
ORF8 mut
GCGCTTATGGCTAGGTGTTGGTACCTTCATGAAGGTCACCAAACTGCTGCATTTAGAGACGTATTTGTT
A L M A R C W Y L H E G H Q T A A F R D V F V
ORF8 no del
GCGCTTATAGCTAGGTGTTGGTACCTTCATGAAGGTCACCAAACTGCTGCATTTATAGACGTACTTGTT
A L I A R C W Y L H E G H Q T A A F I D V L V
ORF8 mut
GTTTTAACTAAACGAACAAATTAA
V L T K R T N *
ORF8 no del
GTTTTAAATAAACGAACAAATTAA
V L N K R T N *
ORF8
Codon
TGT>TGC
TTT>CTT
AAA>AAC
AGG>ACG
AGA>ATA
ATG>ATA
AGA>ATA
TTT>CTT
ACT>AAT
Residue
Cys>Cys S
Phe>Leu N
Lys>Asn N
Arg>Thr N
Arg>Ile N
Met>Ile N
Arg>Ile N
Phe>Leu N
Thr>Asn N
ORF8a mut
ATGAAACTTCTCATTGTTTTGACTTGTATTTTTTTATGCAGTTGCATATG
M K L L I V L T C I F L C S C I C
ORF8a
ATGAAACTTCTCATTGTTTTGACTTGTATTTCTCTATGCAGTTGCATACG
M K L L I V L T C I S L C S C I R
ORF8a mut
CACTGTTGTACAGCGCTGTGCATCTAATAAACCTCATGTGCTTGAAGATC
T V V Q R C A S N K P H V L E D
ORF8a
CACTGTAGTACAGCGCTGTGCATCTAATAAACCTCATGTGCTTGAAGATC
T V V Q R C A S N K P H V L E D
ORF8a mut
CTTGTAAGGTACAACACTAG
P C K V Q H *
ORF8a
CTTGTAAGGTACAACACTAG
P C K V Q H *
ORF8a
Codon
TTT>TCT
TTA>CTA
TGC>CGC
GTT>GTA
Residue
Phe>Ser N
Leu>Leu S
Cys>Arg N
Val>Val S
ORF8b mut
ATGTGCTTGAAGATCCTTGTAAGGTACAACACTAGGGGTAATACTTATAGCACAGCTTGGCTTT
M C L K I L V R Y N T R G N T Y S T A W L
ORF8b
ATGTGCTTGAAGATCCTTGTAAGGTACAACACTAGAGGTAATACTTACAGCACTGCTTGGCTTT
M C L K I L V R Y N T R G N T Y S T A W L
ORF8b mut
GTGCCCTAGGAAAGGTTTTACCTTTTCATAGATGGCACACTATGGTTCAAACATGCACACCTAA
C A L G K V L P F H R W H T M V Q T C T P N
ORF8b
GTGCTCTAGGAAAGGTTTTACCTTTTCATAGATGGCACACTATGGTTCAAACATGCACACCTAA
C A L G K V L P F H R W H T M V Q T C T P N
ORF8b mut
TGTTACTATCAACTGTTAAGATCCAGCTAGGGGTGCGCTTATAGCTAGGTGCTGGTACCTTCAT
V T I N C * D P A R G A L I A R C W Y L H
ORF8b
TGTTACTATCAACTGTCAAGATCCAGCTGGTGGTGCGCTTATAGCTAGGTGTTGGTACCTTCAC
V T I N C Q D P A G G A L I A R C W Y L H
ORF8b mut
GAAGGTCCCCAAACTGCTGCATTTAGAGACGTACTTGTTGTTTTAACTAACCGGATAAATTAA
E G P Q T A A F R D V L V V L T N R I N *
ORF8b
GAAGGTCACCAAACTGCTGCATTTAGAGACGTACTTGTTGTTCTAAATAAACGAACAAACTAA
E G H Q T A A F R D V L V V L N K R T N *
ORF8b
Codon
AGG>AGA
TAT>TAC
ACA>ACT
GCC>GCT
TAA>CAA
AGG>GGT
TGC>TGT
CAT>CAC
CCC>CAC
TTA>CTA
ACT>AAT
AAC>AAA
CGG>CGA
ATA>ACA
AAT>AAC
Residue
Arg>Arg S
Tyr>Tyr S
Thr>Thr S
Ala>Ala S
Stop>Gln
Arg>Gly N
Cys>Cys S
His>His S
Pro>His N
Leu>Leu S
Thr>Asn N
Asn>Lys N
Arg>Arg S
Ile>Thr N
Asn>Asn S

Fig. 1 (**A**) Schematic representation of *ORF8*, highlighting the – 29 bp deletion and subsequent split into *ORF8a* and *ORF8b* with an overlapping region corresponding to the extension of the 5′ end due to the frameshift. (**B**) Predicted mutations (the simplest parsimony analysis) mapped onto *ORF8* and *ORF8a/b* sequences sites to perform statistical analyses. The number of predicted nonsynonymous and synonymous sites was estimated using the parsimony method as implemented in the PBL (Pamilo-Bianchi-Li) approach [22, 23]. We used the Fisher's Exact Test (https://www.langsrud.com/fisher.htm) to analyze a significance of heterogeneity of 2×2 tables assuming independence of 4 variables: the raw number of nonsynonymous and synonymous mutations vs. raw numbers of nonsynonymous and synonymous sites. The right-tailed Fisher Exact Test was used because the alternative to independence in the case of *ORF8* is that there is positive association between the variables (we do expect that numbers of nonsynonymous and synonymous sites/mutations are positively associated, e.g., larger numbers of nonsynonymous mutations are indeed expected when larger numbers of synonymous mutations are observed). We used the PBL method as implemented in DnaSP v. 5.10.01 [24, 25] and MEGA7 software [24, 25] to estimate Ka/Ks values illustrating putative modes of selection. In addition to *ORF8* and *ORF8a/b* genes, we analyzed raw numbers of nonsynonymous and synonymous mutations in *ORF3a*, *ORF6*, *ORF7a*, *ORF7b*, *S*, and *N* genes (Additional file 1). We also analyzed the distribution of mutations across *N*, *S*, and *ORF3a* genes using a sliding window approach. Specifically, the length of each non-overlapping window was equal to the length of the *ORF8* gene (369 bp); incomplete windows at 3′ ends were not used for analyses of nonsynonymous and synonymous mutations. The *ORF3a* gene was split into 3 windows. The *N* gene was split into 4 windows. The *S* gene was split into 11 windows.

To map the mutations onto phylogenetic trees, we first inferred phylogenies using the maximum parsimony approach, which is suitable for closely related sequences [26, 27]. We have also used the unweighted pair group method with arithmetic mean (UPGMA) approach with the number of differences used as model of substitution events as an alternative method to reconstruct phylogenetic trees because it can infer a root of a tree [28, 29]. Reconstruction of ancestral sequences and mutation events across phylogenetic trees were performed in the MEGA7 software with default parameters. The input data are presented in Figs. S3 and S4, the removal of identical positions (to make visual representations of results clearer) does not change estimates of MP and UPGMA topologies that were used in our study. To find candidate parallel mutations, each position of an alignment was analyzed using the MEGA7 software (Tree Explorer). We did not specify outgroup sequences for *ORF8*. For the *ORF8a/b* alignment, the number of mutations is larger compared to that of *ORF8*, thus, more uncertainty of root location is expected. To resolve this issue, we arbitrarily chose AY394995 as a putative outgroup (the sequence that was used as a query in BLASTN searches).

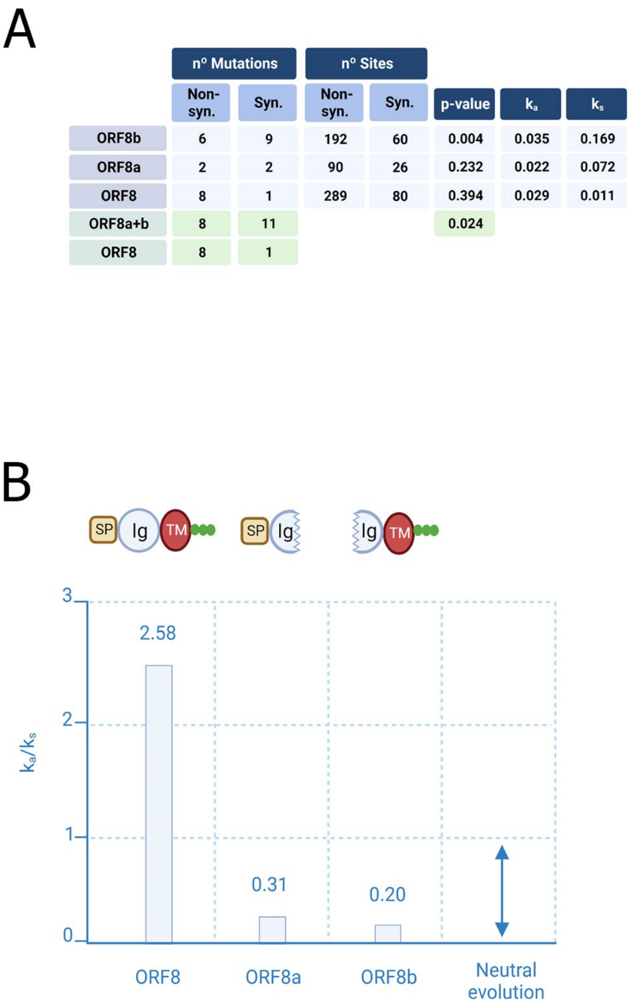
Results

Analysis of nonsynonymous and synonymous substitutions is a powerful tool to analyze modes of natural selection and tendencies in evolution of protein-coding genes [30–32]. Purifying selection acts against deleterious mutations eliminating them from the population. This is by far the predominant form of selection operating in evolution preserving the *status quo* in terms of fitness. The Ka/Ks ratio (the ratio of the rate of non-synonymous nucleotide substitutions, which lead to a change in the encoded amino acid, to the rate of synonymous ones) is commonly used to distinguish between purifying and positive selection. It is widely accepted that Ka/Ks < 1 reflects purifying selection, whereas Ka/Ks > 1 may indicate positive (Darwinian) selection [31, 32].

To estimate Ka and Ks, we extracted mutations from multiple alignments of *ORF8* and *ORF8a*/b sequences. First, we estimated numbers of nonsynonymous and synonymous sites. Lists of mutations are relatively short due to the limited number of SARS-CoV sequences (Fig. 1B, S1 and S2), however these numbers still allowed statistical analyses. The number of nonsynonymous and synonymous sites are 90 and 26 in *ORF8a*, 192 and 60 in *ORF8b*, and 289 and 80 in *ORF8* sequences, respectively (Fig. 2A).

To infer mutations in *ORF8* and *ORF8a/b* genes, we used a simple parsimony approach assuming that (i) consensus sequences are ancestral and (ii) all parallel mutations (if any) are the result of recombination events. In other words, we assumed no parallel (independent recurrent) mutations. We found that the number of synonymous mutations in *ORF8b* is greater than that for nonsynonymous ones (9 vs. 6). Considering the larger number of predicted nonsynonymous sites compared to the synonymous ones, this excess is statistically significant ($p = 0.004$, Fig. 2A). Thus, *ORF8b* is under purifying selection (Ka/Ks is 0.20) (Fig. 2B). We obtained a similar result for the *ORF8a* (Ka/Ks is approximately 0.31, Fig. 2B), although much smaller numbers (2 each of synonymous and nonsynonymous mutations, Fig. 2A) expectedly produced insignificant result ($p = 0.232$). After merging *ORF8a* and *ORF8b* mutations into one set (11 synonymous vs. 8 nonsynonymous mutations), the probability value was 0.002. These results strongly suggested that at least *ORF8b* (or both *ORF8a* and *ORF8b* genes) is(are) under purifying selection and, subsequently, proteins translated from these *ORF*s are functional.

For the intact *ORF8* gene, the number of nonsynonymous mutations is greater than that of the synonymous ones (8 vs. 1) (Fig. 2A). The number of synonymous and nonsynonymous mutations in *ORF8* and *ORF8a/b* is significantly different ($p = 0.024$, Fig. 2A) suggesting different modes of evolution for these *ORF*s. The Ka/Ks value for *ORF8* was greater than 1 (2.58, Fig. 2B) indicating a



	nº Mutations		nº Sites				
	Non-syn.	Syn.	Non-syn.	Syn.	p-value	k_a	k_s
ORF8b	6	9	192	60	0.004	0.035	0.169
ORF8a	2	2	90	26	0.232	0.022	0.072
ORF8	8	1	289	80	0.394	0.029	0.011
ORF8a+b	8	11			0.024		
ORF8	8	1					

Fig. 2 Analyses of nonsynonymous and synonymous mutations. (**A**) Numbers of nonsynonymous/ synonymous mutations and sites, associated *p*-values, and Ka - Ks values. The Ka and Ks values are shown for illustrative purposes only. (**B**) Representative domain architectures of ORF8, ORF8a and ORF8b and corresponding Ka/Ks values. Ka/Ks below 1 reflects purifying selection. "SP" is the signal peptide (light-brown), "Ig" is the immunoglobulin-like domain (light-blue), "TM" is the transmembrane domain (red). Glycosylation sites are shown as green dots. Numbers of nonsynonymous/synonymous, mutations/sites and associated Ka and Ks values

putative positive selection pressure. However, we cannot consider excess of Ka over Ks in *ORF8* as a true signature of positive selection because the excess is not statistically significant (p=0.394, Fig. 2A). It appears that *ORF8* in human SARS-CoV evolve under relaxed purifying selection or even (nearly-)neutrally suggesting weaker functional constrains on the protein encoded by *ORF8*. It is parsimonious to suggest that ORF8 and ORF8a/b experienced different evolutionary forces. For example, emergence of ORF8a/b could be a result of "unsuccessful" searches for "optimal" variants of ORF8. This could cause a "successful" fixation of the ORF8a/b variant in the SARS-CoV population.

Mapping mutations onto phylogenetic trees is problematic for viral sequences. However, we attempted to predict mutations using a conventional phylogenetic approach (maximum parsimony) under the assumption that there are no recombination events. We were not able to define reliable outgroup sequence(s) because more distantly related sequences are unlikely to reflect a complex history of zoonotic transfer events. Therefore, it was not possible to reliably root phylogenetic trees. In other words, different locations of the root across phylogenetic trees depending on various methods of phylogenetic reconstructions should be considered. Alignments of *ORF8* and *ORF8a/b* sequences (Figs. S1 and S2) were reduced to small matrices after removal of noninformative positions (positions without mutations, Figs. S3 and S4). We used two phylogenetic methods that are likely to be consistent with the extremely small volumes of data: UPGMA and maximum parsimony MP). The MP and UPGMA trees have different topologies (Fig. 3).

For *ORF8*, we found one position that contains parallel A>C mutations (the UPGMA method, position #9) (Fig. 3A). No parallel mutations were detected for the MP phylogenetic tree, however, a reverse mutation event C>A was detected in the position #9 at the branch leading to the AY304486 sequence (Fig. 3B). Thus, scenarios appear dramatically different for the position #9 for the MP and UPGMA reconstructions: parallel mutations or one reversal. The uncertainty of these predictions reflects the complexity of the problem. The prediction of parallel mutations increases the number of nonsynonymous mutations (adding one nonsynonymous mutation). However, this increase does not change results of the simple parsimony approach (Fig. 2A): the small probability value for comparison of *ORF8* and *ORF8a/b* would become even smaller (the last p-value in the Fig. 2A). In addition, the p-value for the selection mode detection (the third p-value in the Fig. 2A) remains insignificant. For the *ORF8a/b* alignment, no parallel mutations have been detected for both MP and UPGMA. These results suggested that the simple consensus approach is a good proxy for mutation predictions in the *ORF8a/b* genes. In general, the directionality of mutations does not influence the simplest estimates of numbers of nonsynonymous and synonymous mutations (Fig. 1), although some non-trivial hidden biases cannot be excluded. In addition, potential variations in directionality are not expected to cause major problems for estimates of Ka and Ks in pairwise comparisons [22, 23] that have been used for illustrative purposes only.

We analyzed several other SARS-CoV genes (Fig. 4A). For the *ORF8a*, *ORF8b*, *ORF7a*, and *N* genes, the ratio (R) of the number of nonsynonymous mutations and the number of synonymous mutations is close to or equal 1. For *ORF3a*, *ORF6*, and *S* genes, the R varies between

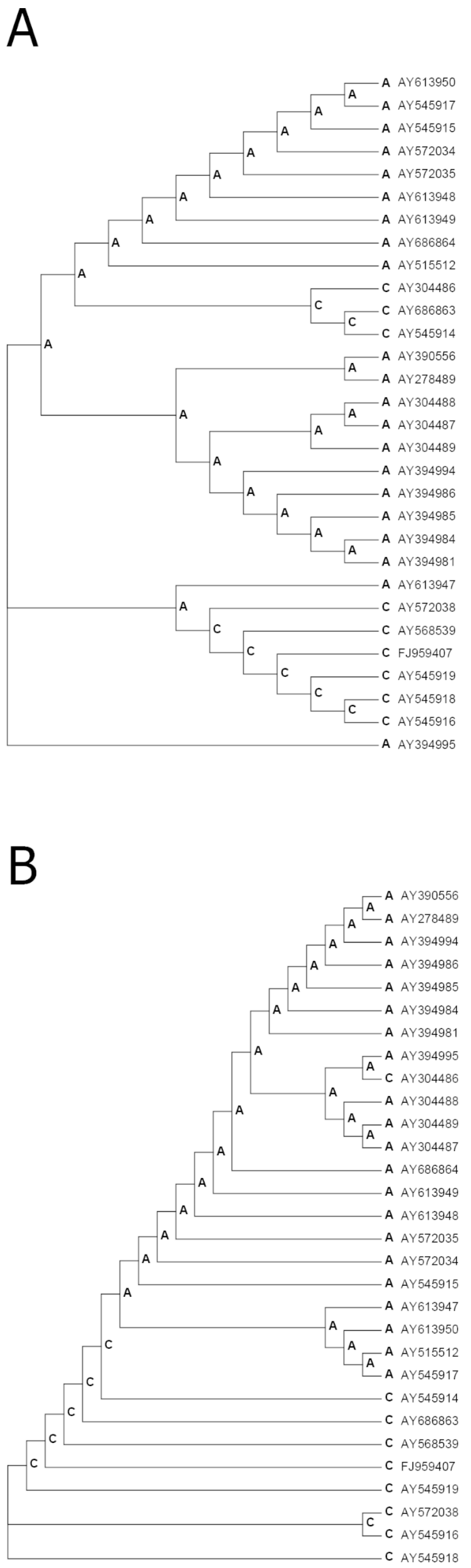


Fig. 3 Mapping mutations onto phylogenetic trees of *ORF8* genes. Ancestral nucleotides are shown at each node of a tree. Changes between nodes indicate mutation events. (**A**) the position #9 that contains parallel A > C mutations (the UPGMA method); (**B**) a reverse mutation event C > A in the position #9 at the branch leading to the AY304486 sequence. Ancestral sequences and mutation events across phylogenetic trees were reconstructed using the MP method

1.6 (*S*) and 3.1 (*ORF3a*) (Fig. 4A). These genes appear to evolve under "relaxed" (less constrained) purifying selection or contain substantial fractions of positively selected sites. Recombination events cannot be excluded. In addition, we analyzed non-overlapping sliding windows with the length 349 nucleotides (the length of *ORF8*) (Fig. 4B). For *ORF3a*, we documented a substantial variation of R: 6.7 (window #1), 4.0 (window #2), and 0.8 (window #3) (Fig. 4B), indicating a sharp decline of R toward the 3′ end of the gene. For the *N* gene, no noticeable trends were detected: R varies from 1.6 (window #4) to 1.7 (window #3). For the *S*, we revealed a mix of windows with relatively small (0.8–2) and windows with relatively large R values (3.3–7.0) (Fig. 4B).

In more distantly related viral genomes (Additional file 3), *ORF8* has smaller Ka compared to Ks. For example, for the SARS coronavirus HSZ-Cc vs. bat SARS-like coronavirus YNLF_34C, the Ka=0.099 and Ks=0.476, giving Ka/Ks of 0.21 (Fig. 5). Analyses of several other bat/civet sequences suggested that for almost all pairwise distances comparisons Ks is greater than Ka. There is only one case of closely related sequences (AY394995 and AY572035), where Ka is similar to Ks, however, a small number of variable sites (Fig. S5) is likely to bias estimates in this case. In general, the Ka/Ks values for distantly related sequences are similar to *ORF8a* and *ORF8b* (Figs. 2 and 5). The mean value of Ka is 0.089, whereas the mean value of Ks is 0.0.781. These results indicate strong purifying selection acting on ORF8 except the two closely related sequences discussed above.

Discussion

The structural properties of proteins encoded by *ORF8a/b* look somewhat puzzling. The Ig-like domain is separated into two parts of approximately equal lengths [10] (Fig. 2B). Thus, functionality of this domain was unclear. In addition to the incomplete Ig-like domain, ORF8b contains the putative transmembrane helix and a glycosylation site. Considering incompleteness of the Ig-like domain and absence of a signal peptide, it is mystifying why ORF8b remains under the strong purifying selection. Therefore, the function(s) of the ORF8b may be separated from that(those) of the ORF8a and these proteins may play independent roles. This puzzling observation echoes with an observed evolutionary trend in SARS-CoV-2 accessory proteins: it is well-known that there are many deletions in the *ORF7a-ORFb-ORF8* complex in SARS-CoV-2 [33–38]. Our recent analysis of deletion in the SAR-CoV-2 genome suggested that the density of deletions (the number of deletions divided by the gene length) for *ORF7a* and *ORF8* is always much higher than that for other SARS-CoV-2 genes [39]. Interestingly, a prominent feature of all the studied deletions is that they predominantly located around the middle of

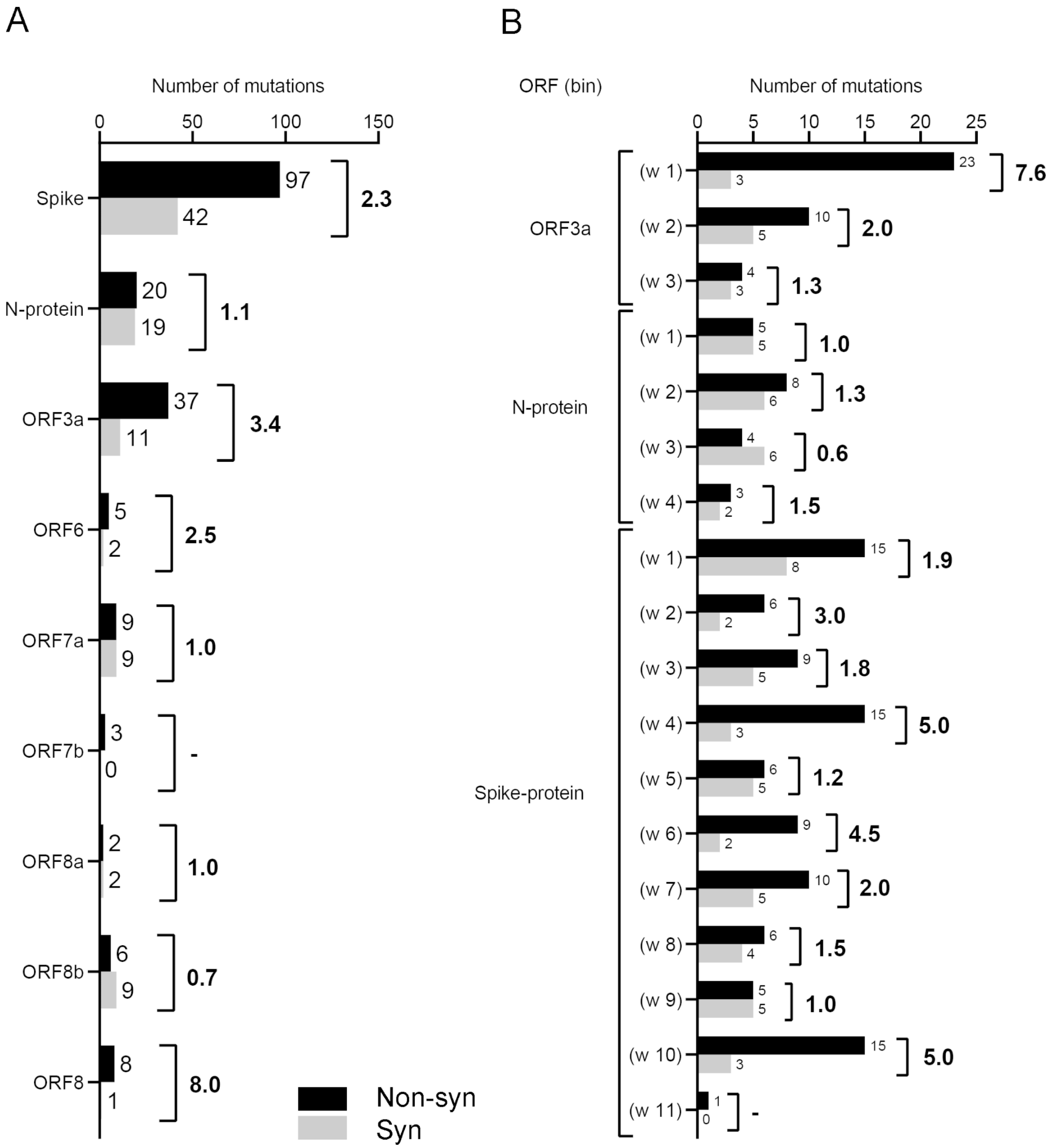


Fig. 4 Analysis of nonsynonymous and synonymous mutations in several SARS-CoV genes. (**A**) Number of nonsynonymous/ synonymous mutations and the ratio of these two numbers in SARS-CoV genes. (**B**) Numbers of nonsynonymous/ synonymous mutations and the ratio of these two numbers in nonoverlapping sliding windows of length 369 (the length of the *ORF8* gene)

the SARS-CoV-2 genes resembling the situation with a 29-nt deletion in SAR-CoV *ORF8*. It should be noted that a hypervariability in SARS-CoV-2 *ORF7A* and *ORF8* is not associated with mechanisms of deletions [39]. Thus, these deletions are likely to reflect recurrent searches of functional “space” of accessory protein combinations to achieve their more advantageous configurations. In the case of SARS-CoV, such a search of optimal configurations/ combinations of accessory proteins led to the ORF8a-ORF8b variant of the ORF8 (Fig. 2B). The observed differences in the rates of nonsynonymous and synonymous mutations in *ORF8* and *ORF8a/b* (Figs. 1

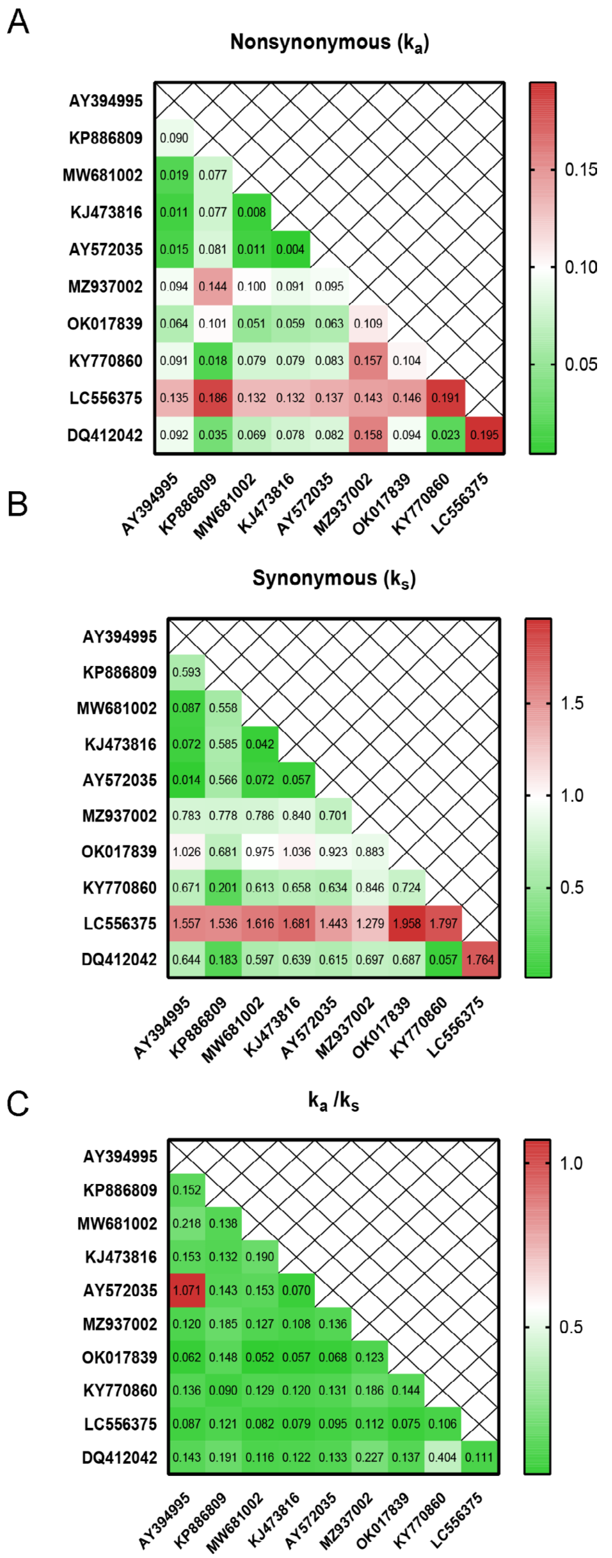


Fig. 5 Estimates of evolutionary divergence between viral sequences. The number of synonymous (**A**) and nonsynonymous (**B**) substitutions per site. (**C**) shows Ka/Ks estimates. Analyses were conducted using the Pamilo-Bianchi-Li model. Numbers of mismatches are shown in Fig. S5

and 2) could be the result of "unsuccessful" searches for "optimal" variants of ORF8. This could cause a "successful" fixation of the ORF8a/b variant in the SARS-CoV population.

The fairly small available dataset of SARS-CoV sequences means that we dealt with relatively small numbers of mutations. Nevertheless, the results of statistical analyses are significant with reasonably small probability values that reject the null hypothesis of homogeneity. Another technical note is that we have counted one nonsense mutation as a nonsynonymous one. This is a conservative approach: after removal of this mutation, all probability values associated with *ORF8b* become even smaller. The reason why we did not remove the missense mutation is that although readthrough events are rare, they are still possible [40]. We also detected one tandem double substitution AGG -> GGT (Fig. 1B), but it appear to evolve neutrally [41]. The most parsimonious scenario of this mutation involves one nonsynonymous mutation and one synonymous mutation; however, we conservatively counted this event as a single nonsynonymous mutation.

A high density of mutations (3 synonymous and 3 nonsynonymous) has been detected at the 3′ end of the *ORF8b* gene (Fig. 1). This clustering is statistically significant: (p=0.015 according to the Fisher exact test, the ancestral sequence was split in two fragments: positions 1-220 and 221–235) (Fig. 1B) suggesting that there a possibility of an episodic adaptation of the C-end of the ORF8b protein to the dramatic 29-nucleotide deletion events. There are several short out-of-frame deletions in the studied *ORF8* and *ORF8a/b* sequences detected in 2 or more sequences (two in *ORF8* and two in *ORF8a/b*) and an additional 1-nucleotide insertion in the *ORF8b* gene (Figs. S1 and S2). Functional implications of these deletions are not clear although some of them may reflect additional functional variability of *ORF8* and *ORF8a/b* genes (similar to the mutations at the 3′ end of *ORF8b* as discussed above).

Prediction of mutations in multiple alignments of viral sequences is a complicated problem because of frequent recombination events and absence of outgroup sequences. Both factors are likely to severely affect prediction of ancestral sequences and mapping of putative mutation events onto phylogenetic trees considering their non-negligible frequency [42–45]. Reconstructions of mutation events using phylogenetic trees may be even misleading because phylogenetic methods assume a single history underlying the data for each position of an alignment [42]. To avoid those major problems, we used a simple parsimony approach assuming that consensus sequences are ancestral, and all parallel mutations have resulted from recombination events. If the first assumption is violated, the changes in directionality of inferred mutations would be expected. However, this will not affect results of mutation analyses. Violations of the second assumption is likely to influence results of mutation analyses; however, we consider a parallel mutation is unlikely event considering an extremely low frequency of mutations in *ORF8* and *ORF8a/b* genes (Fig. 1B). Nevertheless, we also attempted to predict mutations (including parallel and reversal ones) using phylogenetic trees to check this assumption. No parallel mutations were detected for *ORF8a/b* and only one potential event was found in the *ORF8* gene not affecting the conclusion of this paper. Thus, consistent signs of purifying selection were detected for both variants of mutation inferences.

Synonymous substitutions can affect the efficiency of translation and the stability of mRNAs and proteins [46–49]. It is generally accepted that translation efficiency is affected by codon usage bias *via* tuning the rate of elongation [50, 51]. This effect manifests both at the genome-wide scale [46, 49] or in short genomic regions [52–54]. Furthermore, synonymous mutations are likely to experience positive or purifying selection [55–58]. Thus, there is a possibility that a high frequency of synonymous mutations in *ORF8a/b* (Fig. 1) could be due to the positive selection. However, the number of synonymous mutations in *ORF7a* (where positive selection is not expected) and *ORF8b* is the same (Fig. 4). Thus, this hypothesis is not supported by our comparative analysis. As for the RNA stability, our analysis of synonymous mutations W>S and S>W (W=A or T; S=G or C) did not reveal any obvious trends that are expected to reflect substantial changes in the overall stability of *ORF8b* RNA sequence compared to that of its remote homolog, *ORF7a*: the number of S>W and W>S is 4/4 for *ORF7a* and 4/3 for *ORF8b* (Fig. 1B and Additional File 1). It should be noted that this result does not refute the potential functional importance of the viral RNA stability or other factors that may affect the frequency of synonymous mutations.

Conclusions

Deletions causing changes in architecture of accessory proteins are likely to be important in evolution of SARS-CoV lineages. Evolutionary analyses of *ORF8a* and *ORF8b* suggested that the frequency of synonymous mutations was greater than that of nonsynonymous ones. Comparisons with several other SARS-CoV genes suggested that another accessory gene, *ORF7a* has similar ratio of nonsynonymous to synonymous mutations indicating that *ORF8a/b* and *ORF7a* have experienced similar levels of selection pressure. These results imply that *ORF8a* and *ORF8b* are under purifying selection, thus, proteins translated from these *ORF*s are likely to be functionally important. It is well-known that there are many deletions in the *ORF7a-ORFb-ORF8* complex in SARS-CoV-2, which may reflect recurrent searches of functional “space” of accessory protein combinations to achieve their more advantageous configuration. In the case of SARS-CoV, such searches of optimal configurations/combinations of accessory proteins led to the ORF8a-ORF8b variant of ORF8. Such evolutionary trends are largely unexplored field of virology, and they are likely to be important for our understanding of biology and evolution of viruses.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09482-3.

Supplementary Material 1: Fig. S1. Intact SARS-CoV *ORF8*, alignment and delineated mutations. A consensus sequence (reconstructed using the “majority” rule) and “mutated” sequence (all mutations merged together) are shown below the alignment.

Supplementary Material 2: Fig. S2. *ORF8ab* with the deletion, alignment and delineated mutations. A consensus sequence (reconstructed using the “majority” rule) and “mutated” sequence (all mutations merged together) are shown below the alignment.

Supplementary Material 3: Fig. S3. Reduced matrices containing positions with mutations for the alignments of *ORF8* sequences.

Supplementary Material 4: Fig. S4. Reduced matrices containing positions with mutations for the alignments of *ORF8a/b* sequences.

Supplementary Material 5: Fig. S5. Numbers of differences among distantly related viral sequences.

Additional file 1: Alignments of all studied SARS-CoV genes (ZIP file).

Additional file 2: Python script for extraction of mutations from BLASTN multiple alignments (ZIP file).

Additional file 3: Alignments of distantly related viral sequences (ZIP file).

Acknowledgements

We are grateful to the members of our laboratories for stimulating discussions.

Authors & contributions

VY and IBR conceived the study; AB, IBR, GVG, AR-L and AS performed the analyses. I.B.R. and A.S. prepared Figs. 1, 2, 3, 4 and 5. All authors reviewed the manuscript.

Funding

This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health (IBR), the Ministry of Education, Youth and Sports of Czech Republic and ERD Funds (project OPVVV 16_019/0000759) (VY and AB), RCMI grant U54 MD007600 (National Institute on Minority Health and Health Disparities) from the National Institutes of Health (to AR-L).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

VY is an Associate Editor and IBR is an Editorial Advisor of BMC Genomics. Other authors declare that they have no competing interests.

Received: 4 January 2023 / Accepted: 23 June 2023
Published online: 10 July 2023

References

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.
2. Mohammad S, Bouchama A, Mohammad Alharbi B, Rashid M, Saleem Khatlani T, Gaber NS, et al. SARS-CoV-2 ORF8 and SARS-CoV ORF8ab: genomic divergence and functional convergence. Pathogens. 2020;9(9):677.
3. Liu DX, Fung TS, Chong KK, Shukla A, Hilgenfeld R. Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res. 2014;109:97–109.
4. Narayanan K, Huang C, Makino S. SARS coronavirus accessory proteins. Virus Res. 2008;133(1):113–21.
5. Li JY, Liao CH, Wang Q, Tan YJ, Luo R, Qiu Y, et al. The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. Virus Res. 2020;286:198074.
6. López-Muñoz AD, Kosik I, Holly J, Yewdell JW. Cell surface SARS-CoV-2 nucleocapsid protein modulates innate and adaptive immunity. Res Sq. 2021;pre–print:rs3rs–1162804.
7. Su CM, Wang L, Yoo D. Activation of NF-κB and induction of proinflammatory cytokine expressions mediated by ORF7a protein of SARS-CoV-2. Sci Rep. 2021;11(1):13464.
8. Stadler K, Masignani V, Eickmann M, Becker S, Abrignani S, Klenk HD, et al. SARS–beginning to understand a new virus. Nat Rev Microbiol. 2003;1(3):209–18.
9. Zhou Z, Huang C, Zhou Z, Huang Z, Su L, Kang S, et al. Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14^{+} monocytes. iScience. 2021;24(3):102187.
10. Tan Y, Schneider T, Leong M, Aravind L, Zhang D. Novel immunoglobulin domain proteins provide insights into evolution and pathogenesis of SARS-CoV-2-related viruses. mBio. 2020;11(3):e00760–00720.
11. Neches RY, Kyrpides NC, Ouzounis CA. Atypical divergence of SARS-CoV-2 Orf8 from Orf7a within the coronavirus lineage suggests potential stealthy viral strategies in immune evasion. mBio. 2021;12(1):e03014–03020.
12. Oostra M, de Haan CA, Rottier PJ. The 29-nucleotide deletion present in human but not in animal severe acute respiratory syndrome coronaviruses disrupts the functional expression of open reading frame 8. J Virol. 2007;81(24):13876–88.
13. Sung SC, Chao CY, Jeng KS, Yang JY, Lai MM. The 8ab protein of SARS-CoV is a luminal ER membrane-associated protein and induces the activation of ATF6. Virology. 2009;387(2):402–13.
14. Mohammed MEA. The percentages of SARS-CoV-2 protein similarity and identity with SARS-CoV and BatCoV RaTG13 proteins can be used as indicators of virus origin. J Proteins Proteom. 2021;12(2):81–91.
15. Lau SK, Feng Y, Chen H, Luk HK, Yang WH, Li KS, et al. Severe Acute Respiratory Syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. J Virol. 2015;89(20):10532–47.
16. Muth D, Corman VM, Roth H, Binger T, Dijkman R, Gottula LT, et al. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Sci Rep. 2018;8(1):15177.
17. Chen CY, Ping YH, Lee HC, Chen KH, Lee YM, Chan YJ, et al. Open reading frame 8a of the human severe acute respiratory syndrome coronavirus not only promotes viral replication but also induces apoptosis. J Infect Dis. 2007;196(3):405–15.
18. Keng CT, Choi YW, Welkers MR, Chan DZ, Shen S, Gee Lim S, et al. The human severe acute respiratory syndrome coronavirus (SARS-CoV) 8b protein is distinct from its counterpart in animal SARS-CoV and down-regulates the expression of the envelope protein in infected cells. Virology. 2006;354(1):132–42.
19. Law PY, Liu YM, Geng H, Kwan KH, Waye MM, Ho YY. Expression and functional characterization of the putative protein 8b of the severe acute respiratory syndrome-associated coronavirus. FEBS Lett. 2006;580(15):3643–8.
20. Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science. 2003;302(5643):276–8.
21. Chinese SARS, Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science. 2004;303(5664):1666–9.
22. Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 1993;36(1):96–9.
23. Pamilo P, Bianchi NO. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol. 1993;10(2):271–81.
24. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2.
25. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870–4.
26. Herbst L, Fischer M. Ancestral sequence reconstruction with Maximum Parsimony. Bull Math Biol. 2017;79(12):2865–86.
27. Nei M, Kumar S, Takahashi K. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc Natl Acad Sci U S A. 1998;95(21):12390–7.
28. Sneath PH, Sokal RR. Numerical taxonomy. Nature. 1962;193:855–60.
29. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
30. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267(5608):275–6.
31. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000;15(12):496–503.
32. Koonin EV, Rogozin IB. Getting positive about selection. Genome Biol. 2003;4(8):331.
33. Holland LA, Kaelin EA, Maqsood R, Estifanos B, Wu LI, Varsani A, et al. An 81-nucleotide deletion in SARS-CoV-2 ORF7a identified from sentinel surveillance in Arizona (January to March 2020). J Virol. 2020;94(14):e00711–00720.
34. Michel CJ, Mayer C, Poch O, Thompson JD. Characterization of accessory genes in coronavirus genomes. Virol J. 2020;17(1):131.
35. Panzera Y, Calleros L, Goni N, Marandino A, Techera C, Grecco S, et al. Consecutive deletions in a unique uruguayan SARS-CoV-2 lineage evidence the genetic variability potential of accessory genes. PLoS ONE. 2022;17(2):e0263563.
36. Wu A, Wang L, Zhou HY, Ji CY, Xia SZ, Cao Y, et al. One year of SARS-CoV-2 evolution. Cell Host Microbe. 2021;29(4):503–7.
37. Su YCF, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J, et al. Discovery and genomic characterization of a 382-nucleotide deletion in *ORF7b* and *ORF8* during the early evolution of SARS-CoV-2. mBio. 2020;11(4):e01610–01620.
38. Young BE, Fong SW, Chan YH, Mak TM, Ang LW, Anderson DE, et al. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. Lancet. 2020;396(10251):603–11.
39. Rogozin IB, Saura A, Bykova A, Brover V, Yurchenko V. Deletions across the SARS-CoV-2 genome: molecular mechanisms and putative functional consequences of deletions in accessory genes. Microorganisms. 2023;11(1):229.
40. Belinky F, Ganguly I, Poliakov E, Yurchenko V, Rogozin IB. Analysis of stop codons within prokaryotic protein-coding genes suggests frequent readthrough events. Int J Mol Sci. 2021;22(4):1876.
41. Belinky F, Sela I, Rogozin IB, Koonin EV. Crossing fitness valleys via double substitutions within codons. BMC Biol. 2019;17(1):105.
42. Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol. 2002;54(3):396–402.
43. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol. 2010;84(7):3134–46.
44. Turakhia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, Ye C, et al. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. Nature. 2022;609(7929):994–7.
45. Ignatieva A, Hein J, Jenkins PA. Ongoing recombination in SARS-CoV-2 revealed through genealogical reconstruction. Mol Biol Evol. 2022;39(2):msac028.
46. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008;134(2):341–52.

47. Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 2013;41(4):2073–94.
48. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon optimality is a major determinant of mRNA stability. Cell. 2015;160(6):1111–24.
49. Rogozin IB, Gertz EM, Baranov PV, Poliakov E, Schaffer AA. Genome-wide changes in protein translation efficiency are associated with autism. Genome Biol Evol. 2018;10(8):1902–19.
50. Tsai CJ, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. J Mol Biol. 2008;383(2):281–91.
51. Quax TE, Claassens NJ, Soll D, van der Oost J. Codon bias as a means to fine-tune gene expression. Mol Cell. 2015;59(2):149–61.
52. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, et al. A "silent" polymorphism in the *MDR1* gene changes substrate specificity. Science. 2007;315(5811):525–8.
53. Tuller T, Zur H. Multiple roles of the coding sequence 5' end in gene expression regulation. Nucleic Acids Res. 2015;43(1):13–28.
54. Postnikova OA, Uppal S, Huang W, Kane MA, Villasmil R, Rogozin IB, et al. The functional consequences of the novel ribosomal pausing site in SARS-CoV-2 spike glycoprotein RNA. Int J Mol Sci. 2021;22(12):6490.
55. Chamary JV, Hurst LD. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol Biol Evol. 2004;21(6):1014–23.
56. Resch AM, Carmel L, Mariño-Ramírez L, Ogurtsov AY, Shabalina SA, Rogozin IB, et al. Widespread positive selection in synonymous sites of mammalian genes. Mol Biol Evol. 2007;24(8):1821–31.
57. Akashi H, Eyre-Walker A. Translational selection and molecular evolution. Curr Opin Genet Dev. 1998;8(6):688–93.
58. Shen X, Song S, Li C, Zhang J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature. 2022;606(7915):725–31.

Publisher's Note