AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Clinical Science and Epidemiology

# Atypical Divergence of SARS-CoV-2 Orf8 from Orf7a within the Coronavirus Lineage Suggests Potential Stealthy Viral Strategies in Immune Evasion

**Russell Y. Neches,[a] Nikos C. Kyrpides,[a] Christos A. Ouzounis[b]**

[a]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley California, USA
[b]Biological Computation and Process Laboratory, Chemical Process and Energy Resources Institute, Centre for Research and Technology Hellas, Thessalonica, Greece

**ABSTRACT** Orf8, one of the most puzzling genes in the SARS lineage of coronaviruses, marks a unique and striking difference in genome organization between SARS-CoV-2 and SARS-CoV-1. Here, using sequence comparisons, we unequivocally reveal the distant sequence similarities between SARS-CoV-2 Orf8 with its SARS-CoV-1 counterparts and the X4-like genes of coronaviruses, including its highly divergent "paralog" gene Orf7a, whose product is a potential immune antagonist of known structure. Supervised sequence space walks unravel identity levels that drop below 10% and yet exhibit subtle conservation patterns in this novel superfamily, characterized by an immunoglobulin-like beta sandwich topology. We document the high accuracy of the sequence space walk process in detail and characterize the subgroups of the superfamily in sequence space by systematic annotation of gene and taxon groups. While SARS-CoV-1 Orf7a and Orf8 genes are most similar to bat virus sequences, their SARS-CoV-2 counterparts are closer to pangolin virus homologs, reflecting the fine structure of conservation patterns within the SARS-CoV-2 genomes. The divergence between Orf7a and Orf8 is exceptionally idiosyncratic, since Orf7a is more constrained, whereas Orf8 is subject to rampant change, a peculiar feature that may be related to hitherto-unknown viral infection strategies. Despite their common origin, the Orf7a and Orf8 protein families exhibit different modes of evolutionary trajectories within the coronavirus lineage, which might be partly attributable to their complex interactions with the mammalian host cell, reflected by a multitude of functional associations of Orf8 in SARS-CoV-2 compared to a very small number of interactions discovered for Orf7a.

**IMPORTANCE** Orf8 is one of the most puzzling genes in the SARS lineage of coronaviruses, including SARS-CoV-2. Using sophisticated sequence comparisons, we confirm its origins from Orf7a, another gene in the lineage that appears as more conserved, compared to Orf8. Orf7a is a potential immune antagonist of known structure, while a deletion of Orf8 was shown to decrease the severity of the infection in a cohort study. The subtle sequence similarities imply that Orf8 has the same immunoglobulin-like fold as Orf7a, confirmed by structure determination. We characterize the subgroups of this superfamily and demonstrate the highly idiosyncratic divergence patterns during the evolution of the virus.

**KEYWORDS** SARS-CoV-2, coronavirus, Orf7a, X4-like, Orf8, protein superfamily, structure prediction, virus evolution

**Pandemic.** During 2019, a new coronavirus detected in China and found to be capable of human-to-human transmission (1), was shown to be most similar to strains

**Citation** Neches RY, Kyrpides NC, Ouzounis CA. 2021. Atypical divergence of SARS-CoV-2 Orf8 from Orf7a within the coronavirus lineage suggests potential stealthy viral strategies in immune evasion. mBio 12:e03014-20. https://doi.org/10.1128/mBio.03014-20.

**Editor** Diane E. Griffin, Johns Hopkins Bloomberg School of Public Health



Address correspondence to Christos A. Ouzounis, ouzounis@certh.gr.

This article is a direct contribution from Nikos C. Kyrpides, a Fellow of the American Academy of Microbiology, who arranged for and secured reviews by Paul Janssen, SCK-CEN, and Karen Beemon, Johns Hopkins University.

**Received** 14 November 2020
**Accepted** 11 December 2020
**Published** 19 January 2021

from the bat species *Rhinolophus affinis* (2, 3). This strain, subsequently named SARS-CoV-2 (4), is the causal agent for the COVID-19 pandemic and related to the coronavirus strains responsible for the SARS and MERS epidemics (5).

**Genome.** Although five genes of SARS-CoV-2 related to genome replication (Orf1, see reference 6) and virion structure (S, E, M, and N) are common to the two SARS-related viruses (1, 2), as well as omnipresent in other coronaviruses (5, 7), the so-called "accessory" genes play roles that are not well understood (8). In fact, Orf6, Orf7a, and Orf8 appear to be critical genes with hitherto-obscure roles in virus biology (2). These genes are present only in the "SARS" lineage that includes bat viruses (8) transcribed downstream from the genes S-Orf3a-E-M found in other coronavirus groups, followed by N (9).

**Genes.** Genes such as Orf3a and Orf8 have been found in surveillance studies to exhibit variation even within a single bat cave colony (10). Orf8 was seen as a highly variable gene in coronaviruses when present, even before the discovery of SARS-CoV-2 (5). Although Orf7a is reported to share 87.7% identity between SARS-CoV-1 and SARS-CoV-2 (11), Orf8 was presumed unique in SARS-CoV-2 (12), with a fragmented sequence in SARS-CoV-1 present as an Orf8a/b pair, due to a 29-nucleotide (nt) deletion that inactivates the Orf8ab tandem formation (13). This split structure was subsequently detected in bat colonies (10). The present evidence suggests that Orf8 is an evolutionary hot spot in the lineage (14, 15), confirmed by the comparison between SARS-CoV-2 and a pangolin strain: Orf7a exhibits 97.5% sequence identity, in contrast to Orf8 at 94.1% (and 40% with SARS-CoV-1) (16).

**Structure/variations.** The Orf7a protein is a probable immune antagonist of the host cell and forms a family with an immunoglobulin (Ig)-like fold (PF08779) (17), known for its structural and functional versatility (18). SARS-CoV-2 instances from Singapore (382 nt, SG/12-14) are marked by a deletion that maintains Orf7a and eliminates Orf8 entirely, an event also seen in late cases of SARS-CoV-1 (LC2) (14) or elsewhere (19, 20). This genotype was associated with lower incidence of hypoxia in a cohort study, indicating a milder manifestation of symptoms for patients infected with this variant (21). Orf7a has been found to lack its N terminus in one case (81 nt/27 amino acids [aa], AZ-ASU2923), removing its putative signal peptide and one beta-strand pair (22). Profile-profile, but not single-sequence driven, comparisons and protein modeling suggest a common origin of Orf8 (PF12093) with Orf7a (PF08779), with Orf8 being one of the most rapidly evolving segments of the SARS-CoV-2 genome (23). Additional mutants have been detected for Orf7a from Thailand (4-nt frameshift, BKK-0018, C terminus) (24) and Washington (392 nt, fusion with Orf8, WA-UW-4570) (25). The latter instance, coupled with the prediction of a similar Ig-like fold, points to a possible role of tandem Ig domains as important for virus growth.

**Function/interactions.** The Orf8b protein (84 aa) in SARS-CoV-1 induces the activation of ATF6 (10, 26) and triggers intracellular stress by activating the NLRP3 inflammasome (27), whereas protein Orf8a alone (39 aa) induces cellular apoptosis. The Orf8a/b protein pair suppresses the interferon signaling pathway, and Orf8b can also mediate this process via the degradation of IRF3 (28). Moreover, Orf8b downregulates the expression of the viral envelope (E) protein, but not in concert as Orf8ab does (29). Notably, Orf8a interacts with S and Orf8b with M/E/3a/7a—compared to Orf8ab, which interacts with S/3a/7a (29), and less so with M (29, 30). The Orf8ab configuration was found in earlier human samples and animal isolates (31), while the split structure (not seen in SARS-CoV-2 so far) suggests that it facilitates a more efficient replication against interferon, thereby increasing virulence (26, 27). Since Orf8b downregulates protein E, which in turn is known to have a positive effect on virus growth, it has been suggested that Orf8b might play a crucial role in modulating virulence (30). Recent studies confirm that Orf8 is indeed the least conserved protein of SARS-CoV-2 (32, 33), in contrast to the limited variability of other genes across major clades (34).

**Origins.** It has been previously suggested that gene Orf8 in SARS-CoV-1 was acquired from related bat viruses via recombination, since Orf8 proteins from *Rhinolophus ferrumequinum* bats exhibit just 23 to 37% identity to strains in other bat species and ~80%

identity to the human/civet strain (35). Further confirming the origin of Orf8 from bat species, it has also been argued that this region may play a role in tracing the origin of SARS strains in epidemic outbreaks, since the reported deletions do not affect the survival of the SARS-CoV-1 virus (20). Importantly, SARS-CoV-2 has an intact Orf8 (non-split, Orf8ab) gene, which is known to be absent from the more lethal MERS strain (20).

**Motivation.** The Orf7a protein has a known Ig-like structure and exhibits large genome-level variation and yet relative conservation as a single protein family, whereas Orf8, which is predicted to have a similar Ig-like structure, exhibits fewer genome-level variations apart from the split a/b pair or a full deletion and is quite variable at the protein sequence level. It should be noted that neither specific disordered regions have been detected for either of the two genes (36) nor any peculiar codon usage patterns (37). However, this peculiar trajectory of Orf8/Orf7a in SARS-CoV-2 indicates a stealth viral strategy that might be key to control virulence and pathogenicity, with roles in host cell-virus interactions. Here, we investigate the origins and evolution of Orf8 across the coronavirus lineage by sequence and recombination analysis, ultimately merging the two Ig-like families by sensitive searches and identifying the shared conserved residues that define the common Ig fold.

## RESULTS

**Patterns of the multiple sequence alignment.** The multiple alignment, as presented, reveals a turbulent evolutionary history across multiple coronavirus strains for this pair of SARS-CoV-2 genes and their homologs (Fig. 1). At the top of the alignment, there are nine members of the SARS Orf8a and Orf8b lineage, not always corresponding to a cognate pair, and the three truncated structures (block i, Fig. 1A). The next set (block ii) is composed of 93 Orf7a homologs, excluding those 3 of known structure (2 for SARS-CoV-1 and 1 for SARS-CoV-2, intertwined within the first block, due to their truncated N termini, block i thus containing 12 members). The Orf7a group is followed by a quite heterogeneous set of 24 X4-like/Orf9 sequences from bats, Orf8 from pangolin and Orf9 from SARS-CoV-1 (block iii): despite different names and significant variation, they represent similar sequences of common origin. Finally, the bottom group (block iv) comprises 67 Orf8 sequences from SARS-CoV-2 intermixed with those from bat or pangolin hosts. The mutations V71L (V62L) and L96S (L84S) reported elsewhere (38) are clearly seen within the Orf8 family, among others (Fig. 1A, see DS04 at https://doi.org/10.6084/m9.figshare.12678491.v1). Of 265 processed SARS-CoV-2 genomes, only two samples from Wuhan, China (WH19002/2019 and WH19004/2020), along with the USAWA-UW413/2020 case, exhibit no mutations compared to the original reference strain (Wuhan-Hu-1) (see DS.snp).

**Sequence conservation across virus strains.** The first 15 residues are considered to function as signal peptides for Orf7a and Orf8 proteins and were long thought to be the only element shared between them (13). We now extend the signal peptide similarity throughout the superfamily and suggest a common origin for those genes, augmenting independently obtained results by parallel efforts (23). The conservation pattern extends across the entire length of the superfamily, based on evidence from the database searches and the multiple alignment (Fig. 1A). It is notable that the N-terminal region is aligned separately from the downstream protein sequence, suggesting a potential cleavage site. Interestingly, most residues at position 2 are lysines, with two exceptions shown for SARS-CoV-2: one substitution by glutamate (QJF76096) and another by arginine (QIU91254), similar to bat X4-like/Orf9 entries (in the middle of the alignment, block iii). Then, from position 15 onward, the proteins of known structure suggest the presence of the Ig-like beta-sandwich, and an insertion of ~20 residues in the bottom part of the group (block iv, X4-like and Orf8), mostly covered by the regular expression {W.[I,L,V][K,R][I,V,Y]}. The Orf8a/b pair of SARS-CoV-1 underlines the structural flexibility of this segment and the possibility of having two distinct protein products that associate to perform the role of Orf8 in this viral strain. A shorter insert of about 10 residues is also found in Orf8b and some of its Orf9 bat homologs, mostly covered by the regular expression {L[I,V].RC}. Note that these two segments are

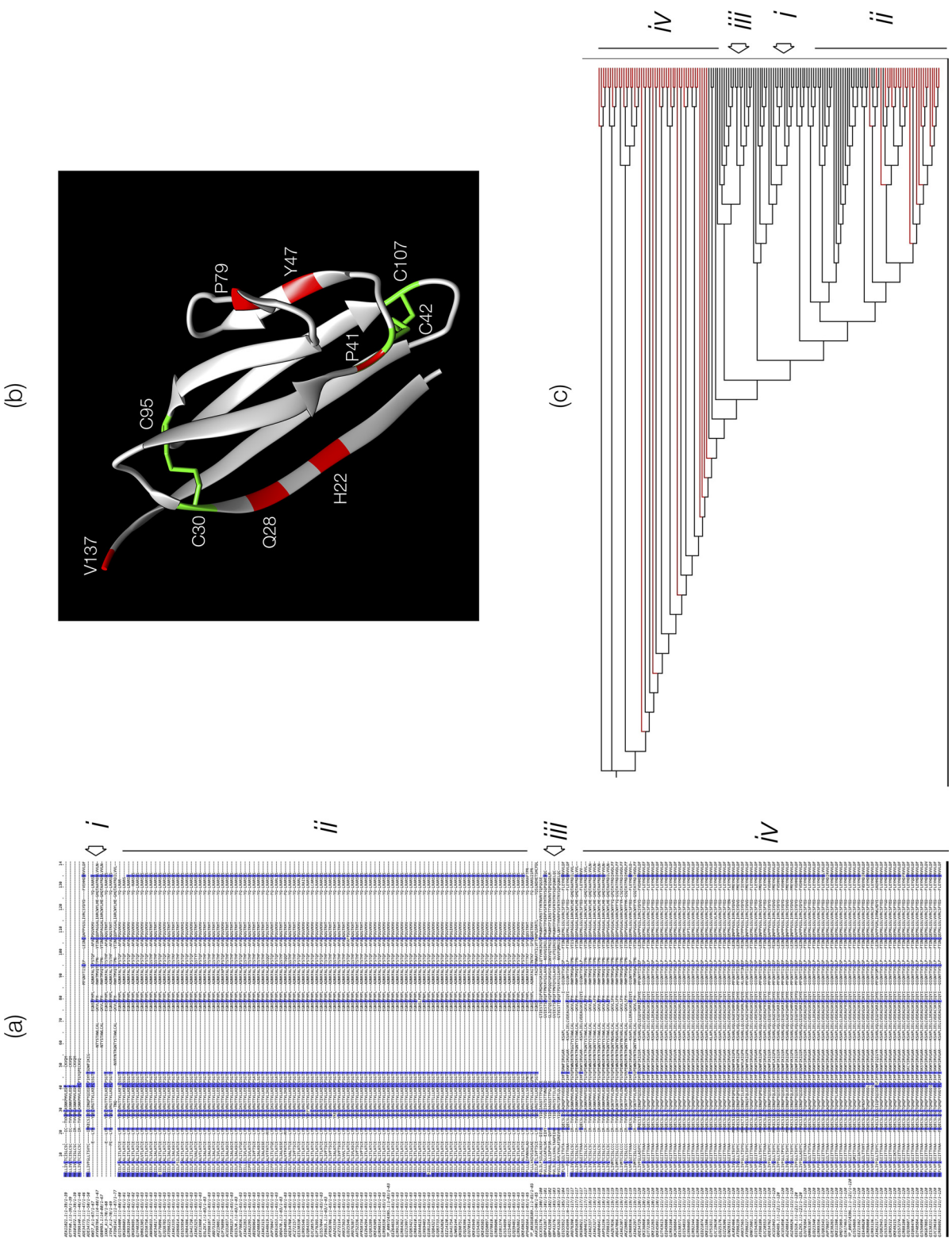


**FIG 1** Sequence-structure-evolution of the coronavirus Orf7a/Orf8 superfamily. (a) Final alignment (see DS04 at https://doi.org/10.6084/m9.figshare.12678491.v1), generated by JalView 2 (52). (b) Sequence-structure conservation for SARS-CoV-1 Orf7a (1xak), 10 conserved residues (apart from three residues found in the signal sequence but not available in the structure) are indicated by red; disulfide bridges are indicated by green, rendered by UCSF Chimera (59). (c) Cladogram based on the final alignment, with SARS-CoV-2 sequences marked by red color, generated by FastTree (55) (DS.tree), visualized by IcyTree (57). The block iv group containing Orf8 homologs exhibits the widest variation. Due to significant variation, a large number of gaps or a fragmented gene structure, a cladogram is selected instead of a more typical phylogenetic tree to depict the structure of the superfamily.

matched onto members of the Orf8ab pair, underlining the likely origin of the intact SARS-CoV-2 Orf8 from those, as mentioned above.

**Conserved positions and structural interpretation.** We define 13 highly conserved positions in the full (final) alignment: M1, K2, and L7 (all in the predicted signal peptide) and H22, Q28, C30, P41, C42, Y47, P79, C95, C107, and V135 (residue coordinates follow Fig. 1A, see DS04 at https://doi.org/10.6084/m9.figshare.12678491.v1). It is known, for Orf7a, that two disulfide bonds are present in positions C30-C90 and C42-C107 (17). Only a few sequences show mutations in positions C30F, C42S, and C107F. The 10 conserved positions are mapped onto the three-dimensional structure of SARS-CoV-1 Orf7a (1XAK) (17), thus defining the invariant amino acid residues for the Ig-like fold of the SARS lineage (Fig. 1B), presumed critical for structural and perhaps functional integrity for the entire protein superfamily. The dendrogram derived from the final alignment reveals the groupings for the superfamily and the distribution of SARS-CoV-2 sequences (DS.tree; Fig. 1C).

**Block annotation and consistency checking.** An annotated version of the alignment was generated, manually trimming low-occupancy positions, and marking the block structure (108 residues long, see DS05 at https://doi.org/10.6084/m9.figshare.12678491.v1 see Fig. S2A). From the annotated alignment (DS05), an automatically trimmed alignment (39) was also generated (54 residues long, see DS06 at https://doi.org/10.6084/m9.figshare.12678491.v1; see Fig. S2B), with two of the conserved positions not included (Y47, P79), in addition to removing most of the variation from rapidly changing residues across family members. This alignment reveals a consistent picture further suggesting that the common origin of these genes is indeed determined by the positions that define the Orf7a Ig-like fold, with Orf8 being subjected to rapid evolutionary change.

**Phylogenetic tree and sequence space.** An orthogonal approach to tree inference is a multidimensional alignment embedding in sequence space (40), which confirms the unique evolutionary history for the superfamily, not readily seen in a tree graph (Fig. 2). The distances in the three-dimensional (3D) embedding represent variation seen across multiple groups: in particular, within-family distances from Orf7a groups are clearly less variable compared to the Orf8 groups, supporting the notion of Orf8 as a highly variable protein in coronaviruses. However, the distances between SARS-CoV-1 and SARS-CoV-2 strain groups are comparable (Fig. 2). Predictably, when nonconserved residues are ignored, the within-family distances appear similar. SARS-CoV-1 Orf7a and Orf8 (also the concatenated ab) come closer than their SARS-CoV-2 counterparts, indicating that there is a faster evolution in the latter, also affected by more extensive sampling of the sequence space. This pattern is also reflected in a tree comparison tanglegram, with the major groups remaining stable, as the fast-evolving nodes alter the tree topology (see Fig. S3).

**Phylogenetic profiling across the coronavirus pangenome.** For both the Orf7a and Orf8 SARS-related families, the phylogenetic distribution of their members is distinctly restricted, supported by an ongoing coronavirus pangenome study (8), which lists these genes separately. We have confirmed a single key exception (X4-like, block iii) in the alpha group (QCX35187.1, https://www.ncbi.nlm.nih.gov/nuccore/MK720946.1/) and its X4-like homologs (e.g., QCX35176.1), as previously shown (23). When mapped onto the tree of 89 representative coronavirus strains derived from DNA-based whole-genome alignment (41), the restricted distribution pattern is clearly visible (see Fig. S4A). In addition, no clear evidence for recombination for this region of the SARS-CoV-2 genome is detectable across the 89 strains, markedly challenging previous claims (35, 42) and further supporting the scenario of gene duplication and extreme divergence. No homologs of this superfamily have ever been detected outside the coronavirus group. Furthermore, we were not able to confirm any other similarities at the structural level (23); these remain particularly valuable predictions for future research.

**From structure to function, as well as constraints from protein interactions.** The discovered interactions of these proteins for SARS-CoV-2 with the mammalian host

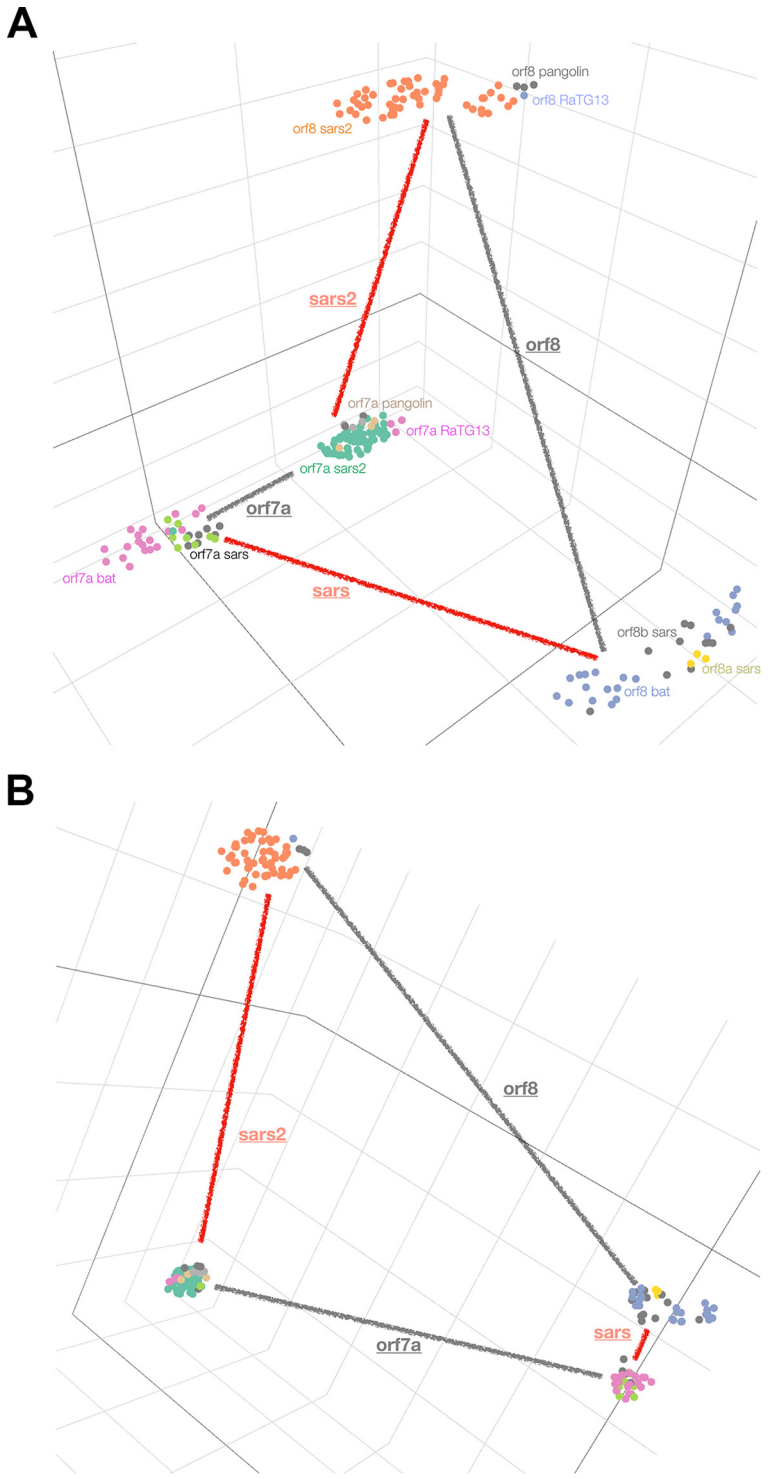


**FIG 2** Variation of Orf7a and Orf8 in the coronavirus lineage. Projections of alignment embedding in a 3D space, using the UMAP (53) function of AlignmentViewer (40), are shown. Groups are denoted by their annotations as follows: pink-light green Orf7a bat/SARS1 and purple-gold-green Orf7a bat/pangolin/SARS2, connected with a gray line and with the family name underlined; yellow-gray-blue Orf8a/Orf8b/bat and magenta-gray-orange bat/pangolin/SARS2, connected as for Orf7a. The groups are also connected by the corresponding names for SARS-CoV-1 (as "sars") and SARS-CoV-2 (as "sars2")—both underlined. (A) Annotated full alignment (see DS05 at https://doi.org/10.6084/m9.figshare.12678491.v1); (B) automatically trimmed alignment (see DS06 at https://doi.org/10.6084/m9.figshare.12678491.v1). This particular depiction is a consistently obtained representation of the embedding from multiple simulations.

proteins reveal a contrasting picture (43): while Orf7a is relatively more conserved, only two interactions have been detected for this protein (https://amp.pharm.mssm.edu/covid19/genesets/20). On the other hand, 47 interactions have been detected for Orf8, which has a much faster evolutionary rate (https://amp.pharm.mssm.edu/covid19/genesets/22) (43). This could be an artifact related to the coverage of protein interactions between the protein complement of the virus and the host cell, and yet it may also point to a more complex interaction landscape for Orf8, underlining its importance in viral strategies to invade cells and propagate.

## DISCUSSION

Here, we show for the first time that remote, nontrivial sequence similarities between the SARS-CoV-2 proteins Orf7a and Orf8 are detectable using supervised sequence space walks in database searches, aimed at precision and reproducibility (44). The detection of similarity between Orf8 and Orf7a, a protein of known structure with an Ig-like fold, implies the importance of this gene duplication for virus biology and confirms previous results, independently derived by different approaches, i.e., profile-profile searches and protein modeling (23). We assessed the extent at which sequence comparisons alone can establish unambiguously the homology between Orf8 and Orf7a family members within the coronavirus lineage and entire pangenome. The embedding of sequence conservation patterns in a multidimensional space reveals atypical divergence patterns, with "equidistant" groupings for Orf7a and Orf8 across SARS-CoV-1 and SARS-CoV-2 but significant divergence of Orf8 compared to within-family distances for Orf7a (Fig. 2). Moreover, the conservation of Orf7a compared to Orf8 is clearly demonstrated not only by its similarity to bat viruses for SARS-CoV-1 and pangolin-*Manis javanica* viruses for SARS-CoV-2, respectively, but with the stark difference between Orf8 in these strains (~35% identity for SARS-CoV-1 Chinese bat-*Rhinolophus sinicus* virus homologs compared to ~88% identity for SARS-CoV-2 pangolin virus homologs), a puzzling pattern (Fig. 3).

During the final stages of this study, the structure of Orf8 from SARS-CoV-2 has been announced in a preprint (PDB 7JTL), confirming the presence of an Ig-like fold (45) and further supporting the notion of a common origin between the Orf7a and Orf8 families. Of the 13 conserved alignment positions, 9 are available (4 are missing at the N terminus): Q28, C30 (strand 1), P41, C42 (loop between strands 2 and 3), Y47 (strand 3), P79 (loop before last three strands in both structures, at 16.141 Å distance of C-alpha atoms, due to the presence of an Orf8-specific region [45], thus excluded from superposition measurements), and C95, C107, and V135 (the latter also excluded due to uncertainty) are shifted and interpreted differently between the automatically derived profile-driven alignment here and the structure-based alignment (45). Remarkably, when the two cysteines are shifted, the root mean square deviation between 7 C-alpha atom pairs drops from 9.883 Å (from their sequence-based match) to 3.398 Å (to their structure-based match), thus confirming the rather different outcomes of the sequence- and structure-based alignments.

We provide strong evidence for the peculiar divergence of Orf8 from Orf7a, within an otherwise dense viral genome and delimit the phylogenetic distribution of these two genes across coronaviruses. Neither of these genes is found in the gamma or delta coronavirus groups, suggestive of a likely loss in those coronavirus sublineages (see Fig. S4BC). It is quite perplexing that no member of this family is present in the MERS clade (23). Given that Orf7a with an Ig-like structure is potentially an immune antagonist with a pivotal role in the viral infection strategy and the recent observation that Orf8 downregulates MHC-I (46), the Orf7a/Orf8 superfamily might be a key system for immune evasion, known for other analogous cases, including herpesviruses, poxviruses, and adenoviruses (47). The detected protein interactions of SARS-CoV-2 exhibit such a sharp contrast between Orf7a and Orf8 (43) that, barring a technical artifact with respect to coverage, the hypothesis arises for Orf7a being used as a conserved template, to generate variants, such as Orf8, wreaking havoc through immune evasion in the host cell.

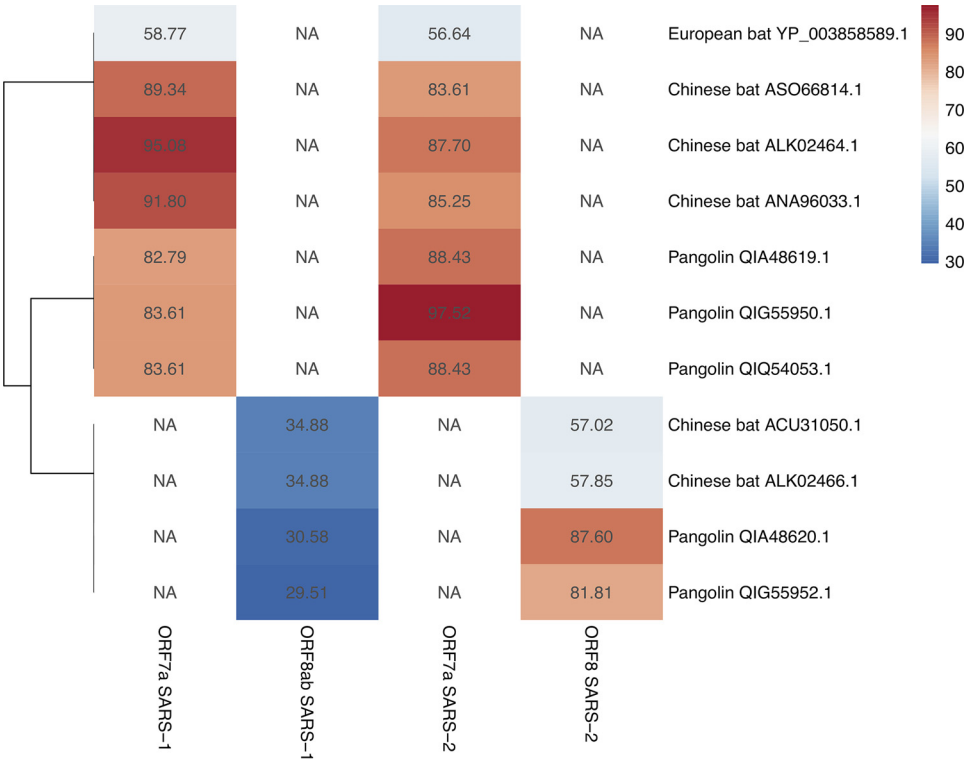


**FIG 3** Closest homologs of the SARS-CoV-1 and SARS-CoV-2 for Orf7a and Orf8. A restricted search against the virus subset of 'nr' that excludes SARS-1/-2 (taxid 694009 and 2697049, respectively) as targets reveals the similarity of superfamily members to coronavirus strains from other hosts. Orf7a of SARS-1 is most similar to Chinese bat (*Rhinolophus sinicus*) strains (~95% identity). The concatenated Orf8ab (from Orf8a/Orf8b) shows the closest—yet lower—similarity to Chinese bats (~35% identity), since the host species *Rhinolophus ferrumequinum* is excluded. In contrast, Orf7a of SARS-2 is most similar to pangolin (*Manis javanica*) strains (~97% identity), with Orf8 exhibiting high levels of similarity again to pangolin (~88% identity). Identifiers next to the host name are provided. The scale (sequence identity) is shown on the right. "NA" signifies that there are no similarities detected in this restricted sequence search. The diagram was generated by using ClustVis (66). SARS-CoV-1 is referred to here as "SARS-1," and SARS-CoV-2 is referred to here as "SARS-2." Orf8 for "SARS-1" corresponds to a concatenated Orf8ab.

In summary, the pair Orf7a/Orf8 may be an under appreciated element in SARS-CoV-2 biology, given its peculiar and quite unusual patterns of divergence and functional properties that might be related to virulence and the pathogenicity of this strain which has caused the COVID-19 pandemic.

## MATERIALS AND METHODS

**Database searches and documentation.** Using the Orf8 protein sequence of SARS-CoV-2 as a primary query, we extensively searched NCBI's 'nr' protein collection (48), following an approach described elsewhere (44), including masking by CAST (49) and permissive E value thresholds in supervised mode. This approach, named "sequence space walk," allows the use of permissive thresholds to detect weak sequence similarities, by inspecting all hits at each single iteration. We specifically searched the Virus subset of 'nr' using PSI-BLAST (50) and the following parameters: max target sequences, 500; Expect 10; word-size 2; BLOSUM62; gap costs, 11/1; compositional adjustment, none; Filter, none; and E value, 0.1. Against the full 'nr', the search exits early at ~200 hits with the same parameters (not shown). Fragments and sequences with unspecified residues were excluded (see below). Search statistics are provided (Table 1), along with the hit tables and PSSMs (the PSSM is unavailable for step 1 only) (DS.runs).

**Other software tools.** Sequence alignments were performed with MAFFT and default parameters (51) and visualized with JalView 2 (52). Family analysis was performed with AlignmentViewer (40), and dimensionality reduction for sequence space exploration (2D and 3D) was aided by UMAP (53), as implemented in AlignmentViewer. The color coding for sequence glyphs was selected according to mView (54) using a residue width of 2 pixels and a residue height of 1 pixel. Validation of sequence database searches, identifier tracking, and alignment quality checks were performed using various scripts and alignments were automatically trimmed with TrimAl (39), unless stated otherwise. Trees were inferred using FastTree (55) on the NGPhylogeny.fr servers with the LG/gamma options (56) and visualized with IcyTree (57). Tree annotation for

**TABLE 1** Summary of the iterative PSI-BLAST profile search[a]

| Step | T+ves | New | All | Lost | Sum | PDB ID | Min identity | Pfam | Archive |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 181 | 0 | 181 | 0 | 181 | | 27.05 | Orf8: 12093 | EFX15XAH01R |
| 2 | 181 | 26 | 207 | 1 | 206 | | 26.19 | | EFXHBT3G016 |
| 3 | 206 | 235 | 441 | 11 | 430 | 1xak, 1yo4, 6w37 | 20.00 | Orf7a: 08779 | EFXZ8AHS01R |
| 4 | 430 | 22 | 452 | 11 | 441 | | 19.05 | | EFZD8C4901R |
| 5 | 441 | 13 | 454 | 0 | 454 | | 19.05 | | EG07J0HG016 |
| 6 | 454 | 2 | 456 | 1 | 455 | | 14.61 | | EG0RNDT7016 |
| 7 | 455 | 3 | 458 | 0 | 458 | | 13.64 | | EG126CM8014 |
| 8 | 458 | 7 | 465 | 0 | 465 | | 8.06 | | EG1C8C66016 |
| 9 | 465 | 0 | 465 | 0 | 465 | | 8.06 | | EG1UF3ZC014 |

[a]Columns: Step, incremental count of the profile search; T+ves, number of entries detected; New, number of new entries detected (excluding step 1); All, entries detected at the corresponding step; Lost, entries lost and considered as false negatives (see text); Sum, total number of entries recovered; PDB ID, PDB identifiers for proteins of known structure detected; Min identity, minimum sequence identity; Pfam, Pfam identifiers for proteins detected; Archive, file name of results (hits/PSSM), available in DS.runs. The search converges at step 9, with no "new" hits.

genes and clades was supported by MicroReact (58). Protein structure analysis and annotation was performed with UCSF Chimera (59). Protein interaction data were obtained from the Covid-19 drug/gene set library (60). Variations were calculated using Virulign (61) over a representative sample of SARS-CoV-2 genomes. Recombination breakpoints were identified using BALi-Phy (62) to compute the pairwise homoplasy index across the genome alignment (63). Regions without breakpoints were identified as nonrecombining regions (15). Further details are given in Text S1 in the supplemental material.

**Supervised sequence space walk.** The initial query sequence Orf8 from SARS-CoV-2 detects its closest homologs within the Orf8 family (PF12093) (29), followed at step 3 by distant homologs of the Orf7a family (PF08779) (17, 64), variably called X4 or Orf10, from a range of viral hosts that include bats, pangolins and civets. The search converges at step 9 with high precision, i.e., no known false positives (Fig. 4a). The detected region of homology spans the length of these proteins beyond the signal sequence, as observed independently (23), and unifies the two Pfam entries into a single superfamily (65), with few invariant residues across its members (see below). All PSSMs are stored for future searches (see Text S1, DS.runs) and re-use by the community. Since the results are time sensitive at present (July 2020), the PSSM collection is provided to ensure reproducibility: very many and highly similar sequences will be deposited in the meanwhile, without substantially modifying the main conclusions or challenging the validity of the reported sequence search results.

**Identifier processing.** There are 24 sequence entries that are lost (as false negatives) in the iterative search, 20 of which reappear in subsequent iterations (Fig. 4b). Of those, 1 entry is lost at step 2, 11 are lost at step 3, another 11 are lost at step 4, and just 1 is lost at step 6. All lost entries were eventually recovered, except for 4 (all of representing sequences shorter than 25 residues). A link with the 24 entries is provided for further inspection (identifiers are available in DS.runs).

**Quality control and alignment.** The iterative search returns 465 unique entries, many of which are partial sequences (of short length) or have multiple undefined residues (see DS01 at https://doi.org/10.6084/m9.figshare.12678491.v1); just 1 entry (QJI07349.1) returns two hits in separate regions, due to 70 undefined middle positions (one duplicate ID). When partial/undefined sequences were removed, 288 entries remained, which were then aligned by MAFFT, revealing the conservation patterns of the superfamily (Fig. 4c, initial alignment, 288 sequences, 146 residues long, see DS02 at https://doi.org/10.6084/m9.figshare.12678491.v1). With filtering single undefined residues, 92 sequences were removed (edited alignment, 196 sequences, 146 residues long, see DS03 at https://doi.org/10.6084/m9.figshare.12678491.v1): 21 are marked as partial, and 71 contain single undefined positions. The only short sequences retained reflect the Orf8a/b (or Orf10) gene configuration (31).

**Multiple alignment editing and statistics.** To define a more accurate, reference alignment, all retained ($n$ = 196) sequences were set to be ≥80 residues long (≤100 for Orf7a and >100 for Orf8, plus other members of the superfamily to demonstrate variation), manually removing the C-terminal end for five low-occupancy positions. Exceptions were the Orf8a/b pair and three (truncated) structure entries from the PDB (1XAK 68/83 [17], 1YO4 69/87 [64], and 6W37 67/67). The reference alignment is provided (Fig. 4d, final alignment, 196 sequences, 141 residues long, see DS04 at https://doi.org/10.6084/m9.figshare.12678491.v1): at 80% identity threshold, we defined 13 conserved residues (see below), 9% (13/141) of the length of the aligned sequences.

**Cross-family sequence similarities.** The sequence space walk with Orf8 revealed its closest homologs at the first steps and at step 3 connected them with Orf7a and its homologs (Fig. 4a), with minimum sequence identity dropping from 20 to 8% at convergence (Table 1). This outcome is also confirmed by the resulting alignments, with an average sequence identity dropping from 40 to 10%. The alignment depiction as glyphs for both the initial (Fig. 4c) and the final (Fig. 4d) alignments is also reflected in the matrix of pairwise sequence similarities for the former (see Fig. S1a) and the latter (see Fig. S1b), respectively; these cross-similarity patterns were generated by AlignmentViewer (40). (Note that in this study, for data annotation purposes only, SARS-CoV-1 and SARS-CoV-2 are marked as "SARS-1" and "SARS-2," respectively.)

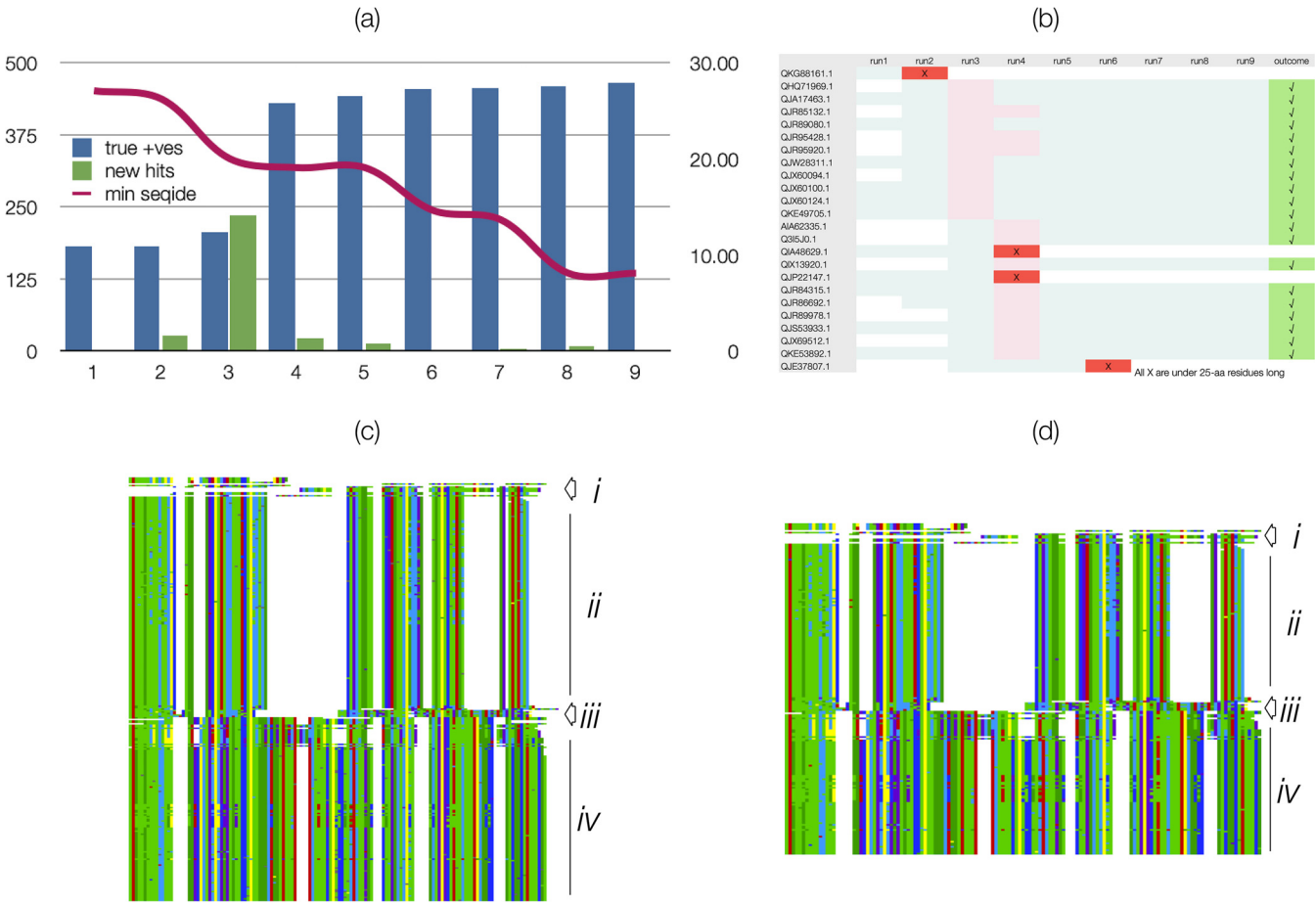


**FIG 4** Sequence searches and resulting multiple alignments. (a) Depiction of the sequence space walk, as executed by multiple database search iterations (*x* axis, 9 in total). The left *y* axis shows the number of entries (true positives [true +ves], blue bars) and the number of new hits (new hits, green bars); the right *y* axis shows the level of minimum sequence identity (min seqide, red line), dropping from approximately 30 to 8% at the final iteration. (b) Tabular representation for the 24 entries that are lost and recovered during the search (listed in the left column). White indicates absence (including those during the first and second iterations; green boxes indicates detection, pink boxes indicate temporary loss, and red boxes for four cases (see text) indicate permanent loss. The right column shows the 20 sequences that were recovered. (c) Alignment glyph of the initial alignment (see DS02 at https://doi.org/10.6084/m9.figshare.12678491.v1, blocks i to iv are shown [see text]). (d) Alignment glyph of the final alignment (see DS04 at https://doi.org/10.6084/m9.figshare.12678491.v1, blocks i to iv as described above), generated by AlignmentViewer (40), with a color scheme according to mview (54).

**Additional supplement data.** Additional supplemental data can be found at https://doi.org/10.6084/m9.figshare.12678491.v1.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**FIG S1**, TIF file, 4.2 MB.

**FIG S2**, TIF file, 6.3 MB.

**FIG S3**, TIF file, 2.6 MB.

**FIG S4A**, TIF file, 0.2 MB.

**FIG S4BC**, TIF file, 0.2 MB.

**TEXT S1**, DOC file, 0.02 MB.

## ACKNOWLEDGMENTS

We thank Konstantinos Kyritsis (School of Pharmacy, Aristotle University of Thessalonica) for help with the "dendextend" tree comparison.

C.A.O. acknowledges support by the project Elixir-GR, implemented under the Action "Reinforcement of the Research and Innovation Infrastructure," funded by the Operational Program Competitiveness, Entrepreneurship, and Innovation (NSRF 2014-2020) and cofinanced by Greece and the European Union (European Regional

Development Fund). This study was supported in part by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. The information presented here does not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred.

C.A.O. is an Affiliate Scholar of Lawrence Berkeley National Laboratory.

## REFERENCES

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W, China Novel Coronavirus Investigating and Research Team. 2020. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733. https://doi.org/10.1056/NEJMoa2001017.
2. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang Y-Z. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265–269. https://doi.org/10.1038/s41586-020-2008-3.
3. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7.
4. *Coronaviridae* Study Group. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 5:536–544. https://doi.org/10.1038/s41564-020-0695-z.
5. Cui J, Li F, Shi ZL. 2019. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17:181–192. https://doi.org/10.1038/s41579-018-0118-9.
6. Neuman BW. 2016. Bioinformatics and functional analyses of coronavirus nonstructural proteins involved in the formation of replicative organelles. Antiviral Res 135:97–107. https://doi.org/10.1016/j.antiviral.2016.10.005.
7. Chen B, Tian E-K, He B, Tian L, Han R, Wang S, Xiang Q, Zhang S, El Arnaout T, Cheng W. 2020. Overview of lethal human coronaviruses. Signal Transduct Target Ther 5:89. https://doi.org/10.1038/s41392-020-0190-2.
8. Alam I, Kamau A, Kulmanov M, Arold ST, Pain AT, Gojobori T, Duarte CM. 2020. Functional pangenome analysis suggests inhibition of the protein E as a readily available therapy for COVID-2019. bioRxiv https://www.biorxiv.org/content/10.1101/2020.02.17.952895v2.
9. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. Cell 181:914–921. https://doi.org/10.1016/j.cell.2020.04.011.
10. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, Xie J-Z, Shen X-R, Zhang Y-Z, Wang N, Luo D-S, Zheng X-S, Wang M-N, Daszak P, Wang L-F, Cui J, Shi Z-L. 2017. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog 13:e1006698. https://doi.org/10.1371/journal.ppat.1006698.
11. Xu J, Zhao S, Teng T, Abdalla AE, Zhu W, Xie L, Wang Y, Guo X. 2020. Systematic comparison of two animal-to-human transmitted human coronaviruses: SARS-CoV-2 and SARS-CoV. Viruses 12:244. https://doi.org/10.3390/v12020244.
12. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 7:1012–1023. https://doi.org/10.1093/nsr/nwaa036.
13. Oostra M, de Haan CA, Rottier PJ. 2007. 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 81:13876–13888. https://doi.org/10.1128/JVI.01631-07.
14. Su YCF, Anderson DE, Young BE, Zhu F, Linster M, et al. 2020. Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2. bioRxiv https://www.biorxiv.org/content/10.1101/2020.03.11.987222v1.
15. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, Rambaut A, Robertson DL. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol 5:1408–1417. https://doi.org/10.1038/s41564-020-0771-4.
16. Liu P, Jiang J-Z, Wan X-F, Hua Y, Li L, Zhou J, Wang X, Hou F, Chen J, Zou J, Chen J. 2020. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog 16:e1008421. https://doi.org/10.1371/journal.ppat.1008421.
17. Nelson CA, Pekosz A, Lee CA, Diamond MS, Fremont DH. 2005. Structure and intracellular targeting of the SARS-coronavirus Orf7a accessory protein. Structure 13:75–85. https://doi.org/10.1016/j.str.2004.10.010.
18. Bodelón G, Palomino C, Fernández LÁ. 2013. Immunoglobulin domains in *Escherichia coli* and other enterobacteria: from pathogenesis to applications in antibody technologies. FEMS Microbiol Rev 37:204–250. https://doi.org/10.1111/j.1574-6976.2012.00347.x.
19. Song H-D, Tu C-C, Zhang G-W, Wang S-Y, Zheng K, Lei L-C, Chen Q-X, Gao Y-W, Zhou H-Q, Xiang H, Zheng H-J, Chern S-WW, Cheng F, Pan C-M, Xuan H, Chen S-J, Luo H-M, Zhou D-H, Liu Y-F, He J-F, Qin P-Z, Li L-H, Ren Y-Q, Liang W-J, Yu Y-D, Anderson L, Wang M, Xu R-H, Wu X-W, Zheng H-Y, Chen J-D, Liang G, Gao Y, Liao M, Fang L, Jiang L-Y, Li H, Chen F, Di B, He L-J, Lin J-Y, Tong S, Kong X, Du L, Hao P, Tang H, Bernini A, Yu X-J, Spiga O, Guo Z-M, et al. 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci U S A 102:2430–2435. https://doi.org/10.1073/pnas.0409608102.
20. Wu Z, Yang L, Ren X, Zhang J, Yang F, Zhang S, Jin Q. 2016. ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. J Infect Dis 213:579–583. https://doi.org/10.1093/infdis/jiv476.
21. Young BE, Fong SW, Chan YH, Mak TM, Ang LW, et alet al. 2020. 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 396:603–611. https://doi.org/10.1016/S0140-6736(20)31757-8.
22. Holland LA, Kaelin EA, Maqsood R, Estifanos B, Wu LI, Varsani A, Halden RU, Hogue BG, Scotch M, Lim ES. 2020. An 81- nucleotide deletion in SARS-CoV-2 ORF7a identified from sentinel surveillance in Arizona (Jan-Mar 2020). J Virol 94:e00711-20. https://doi.org/10.1128/JVI.00711-20.
23. Tan Y, Schneider T, Leong M, Aravind L, Zhang D. 2020. Novel immunoglobulin domain proteins provide insights into evolution and pathogenesis of SARS-CoV-2-related viruses. mBio 11:e00760-20. https://doi.org/10.1128/mBio.00760-20.
24. Batty EM, Kochakarn T, Panthan B, Kumpornsin K, Jiaranai P, et al. 2020. Genomic surveillance of SARS-CoV-2 in Thailand reveals mixed imported populations, a local lineage expansion and a virus with truncated ORF7a. medRxiv https://www.medrxiv.org/content/10.1101/2020.05.22.20108498v1.
25. Addetia A, Xie H, Roychoudhury P, Shrestha L, Loprieno M, et alet al. 2020. Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates. J Clin Virol https://doi.org/10.1016/j.jcv.2020.104523.
26. Sung SC, Chao CY, Jeng KS, Yang JY, Lai MM. 2009. The 8ab protein of SARS-CoV is a luminal ER membrane-associated protein and induces the activation of ATF6. Virology 387:402–413. https://doi.org/10.1016/j.virol.2009.02.021.
27. Shi CS, Nabar NR, Huang NN, Kehrl JH. 2019. SARS-coronavirus open reading frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes. Cell Death Discov 5:101. https://doi.org/10.1038/s41420-019-0181-7.
28. Wong HH, Fung TS, Fang S, Huang M, Le MT, Liu DX. 2018. Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3. Virology 515:165–175. https://doi.org/10.1016/j.virol.2017.12.028.
29. Keng C-T, Choi Y-W, Welkers MRA, Chan DZL, Shen S, Gee Lim S, Hong W, Tan Y-J. 2006. 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 354:132–142. https://doi.org/10.1016/j.virol.2006.06.026.
30. Keng C-T, Tan Y-J. 2010. Molecular and biochemical characterization of the SARS-CoV accessory proteins ORF8a, ORF8b, and ORF8ab, p 177–191.

*In* Lal YJ (ed), Molecular biology of the SARS-coronavirus. Springer-Verlag, Berlin, Germany.
31. McBride R, Fielding BC. 2012. The role of severe acute respiratory syndrome (SARS)-coronavirus accessory proteins in virus pathogenesis. Viruses 4:2902–2923. https://doi.org/10.3390/v4112902.
32. Cui H, Gao Z, Liu M, Lu S, Mkandawire W, et al. 2020. Structural genomics and interactomics of 2019 Wuhan novel coronavirus, SARS-CoV-2, indicate evolutionary conserved functional regions of viral proteins. Preprints preprints202002.200372.v202001.
33. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, Chen Y, Gnanakaran S, Korber B, Gao F. 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci Adv 6:eabb9153. https://doi.org/10.1126/sciadv.abb9153.
34. Chiara M, Horner DS, Pesole G. 2020. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-Cov-2. bioRxiv https://doi.org/10.1101/2020.03.30.016790.
35. Lau SKP, Feng Y, Chen H, Luk HKH, Yang W-H, Li KSM, Zhang Y-Z, Huang Y, Song Z-Z, Chow W-N, Fan RYY, Ahmed SS, Yeung HC, Lam CSF, Cai J-P, Wong SSY, Chan JFW, Yuen K-Y, Zhang H-L, Woo PCY. 2015. Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. J Virol 89:10532–10547. https://doi.org/10.1128/JVI.01048-15.
36. Giri R, Bhardwaj T, Shegane M, Gehi BR, Kumar P, Gadhavi K. 2020. Dark proteome of newly emerged SARS-CoV-2 in comparison with human and bat coronaviruses. bioRxiv https://doi.org/10.1101/2020.03.13.990598.
37. Kames J, Holcomb DD, Kimchi O, DiCuccio M, Hamasaki-Katagiri N, Wang T, Komar AA, Alexaki A, Kimchi-Sarfaty C. 2020. Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. Sci Rep 10:15643. https://doi.org/10.1038/s41598-020-72533-2.
38. Nguyen TM, Zhang Y, Pandolfi PP. 2020. Virus against virus: a potential treatment for 2019-nCov (SARS-CoV-2) and other RNA viruses. Cell Res 30:189–190. https://doi.org/10.1038/s41422-020-0290-0.
39. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
40. Reguant R, Antipin Y, Sheridan R, Dallago C, Diamantoukos D, Luna A, Sander C, Gauthier NP. 2020. AlignmentViewer: sequence analysis of large protein families (version 1). F1000Res 9:213. https://doi.org/10.12688/f1000research.22242.1.
41. Ou Z, Ouzounis C, Wang D, Sun W, Li J, Chen W, Marliere P, Danchin A. 2020. A path towards SARS-CoV-2 attenuation: metabolic pressure on CTP synthesis rules the virus evolution. Genome Biol Evol 12:2467–2485. https://doi.org/10.1093/gbe/evaa229.
42. Mohammad S, Bouchama A, Mohammad Alharbi B, Rashid M, Saleem Khatlani T, Gaber NS, Malik SS. 2020. SARS-CoV-2 ORF8 and SARS-CoV ORF8ab: genomic divergence and functional convergence. Pathogens 9:677. https://doi.org/10.3390/pathogens9090677.
43. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL, Tummino TA, Hüttenhain R, Kaake RM, Richards AL, Tutuncuoglu B, Foussard H, Batra J, Haas K, Modak M, Kim M, Haas P, Polacco BJ, Braberg H, Fabius JM, Eckhardt M, Soucheray M, Bennett MJ, Cakir M, McGregor MJ, Li Q, Meyer B, Roesch F, Vallet T, Mac Kain A, Miorin L, Moreno E, Naing ZZC, Zhou Y, Peng S, Shi Y, Zhang Z, Shen W, Kirby IT, Melnyk JE, Chorba JS, Lou K, Dai SA, Barrio-Hernandez I, Memon D, Hernandez-Armenta C, et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583:459–468. https://doi.org/10.1038/s41586-020-2286-9.
44. Promponas VJ, Katsani KR, Blencowe BJ, Ouzounis CA. 2016. Sequence evidence for common ancestry of eukaryotic endomembrane coatomers. Sci Rep 6:22311. https://doi.org/10.1038/srep22311.
45. Flower TG, Buffalo CZ, Hooy RM, Allaire M, Ren X, Hurley JH. 2020. Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. bioRxiv https://doi.org/10.1101/2020.08.27.270637.
46. Zhang Y, Zhang J, Chen Y, Luo B, Yuan Y, et alet al. 2020. The ORF8 protein of SARS-CoV-2 mediates immune evasion through potently downregulating MHC-I. bioRxiv https://doi.org/10.1101/2020.05.24.111823.
47. Farre D, Martinez-Vicente P, Engel P, Angulo A. 2017. Immunoglobulin superfamily members encoded by viruses and their multiple roles in immune evasion. Eur J Immunol 47:780–796. https://doi.org/10.1002/eji.201746984.
48. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A, Kitts PA, Kuznetsov A, Lathrop S, Lu Z, McGarvey K, Madden TL, Murphy TD, O'Leary N, Phan L, Schneider VA, Thibaud-Nissen F, Trawick BW, Pruitt KD, Ostell J. 2020. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 48:D9–D16. https://doi.org/10.1093/nar/gkz899.
49. Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA. 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Bioinformatics 16:915–922. https://doi.org/10.1093/bioinformatics/16.10.915.
50. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
51. Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20:1160–1166. https://doi.org/10.1093/bib/bbx108.
52. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191. https://doi.org/10.1093/bioinformatics/btp033.
53. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 37:38–44. https://doi.org/10.1038/nbt.4314.
54. Brown NP, Leroy C, Sander C. 1998. MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 14:380–381. https://doi.org/10.1093/bioinformatics/14.4.380.
55. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
56. Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, Gascuel O. 2019. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. Nucleic Acids Res 47:W260–W265. https://doi.org/10.1093/nar/gkz303.
57. Vaughan TG. 2017. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. Bioinformatics 33:2392–2394. https://doi.org/10.1093/bioinformatics/btx155.
58. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2:e000093. https://doi.org/10.1099/mgen.0.000093.
59. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612. https://doi.org/10.1002/jcc.20084.
60. Kuleshov M, Clarke DJB, Kropiwnicki E, Jagodnik K, Bartal A, et al. 2020. The COVID-19 gene and drug set library. SSRN https://ssrn.com/abstract=3606799.
61. Libin PJK, Deforche K, Abecasis AB, Theys K. 2019. VIRULIGN: fast codon-correct alignment and annotation of viral genomes. Bioinformatics 35:1763–1765. https://doi.org/10.1093/bioinformatics/bty851.
62. Suchard MA, Redelings BD. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22:2047–2048. https://doi.org/10.1093/bioinformatics/btl175.
63. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet 10:e1004342. https://doi.org/10.1371/journal.pgen.1004342.
64. Hanel K, Stangler T, Stoldt M, Willbold D. 2006. Solution structure of the X4 protein coded by the SARS-related coronavirus reveals an immunoglobulin like fold and suggests a binding activity to integrin I domains. J Biomed Sci 13:281–293. https://doi.org/10.1007/s11373-005-9043-9.
65. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995.
66. Metsalu T, Vilo J. 2015. ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. Nucleic Acids Res 43:W566–W570. https://doi.org/10.1093/nar/gkv468.
67. Galili T. 2015. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics 31:3718–3720. https://doi.org/10.1093/bioinformatics/btv428.