Sarbecovirus comparative genomics elucidates gene content of SARS-CoV-2 and functional impact of COVID-19 pandemic mutations

Irwin Jungreis[1,2], Rachel Sealfon[3], Manolis Kellis[1,2†]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA;
[2]Broad Institute of MIT and Harvard, Cambridge, MA;
[3]Center for Computational Biology, Flatiron Institute, New York, NY

†Corresponding author: manoli@mit.edu

Abstract

Despite its overwhelming clinical importance for understanding and mitigating the COVID-19 pandemic, the protein-coding gene content of the SARS-CoV-2 genome remains unresolved, with the function and even protein-coding status of many hypothetical proteins unknown and often conflicting among different annotations, thus hindering efforts for systematic dissection of its biology and the impact of recent mutations. Comparative genomics is a powerful approach for distinguishing protein-coding versus non-coding functional elements, based on their characteristic patterns of change, which we previously used to annotate protein-coding genes in human, fly, and other species. Here, we use comparative genomics to provide a high-confidence set of SARS-CoV-2 protein-coding genes, to characterize their protein-level and nucleotide-level evolutionary constraint, and to interpret the functional implications for SARS-CoV-2 mutations acquired during the current pandemic. We select 44 complete Sarbecovirus genomes at evolutionary distances well-suited for protein-coding and non-coding element identification, create whole-genome alignments spanning all named and putative genes, and quantify their protein-coding evolutionary signatures using PhyloCSF and their overlapping constraint using FRESCo. We find strong protein-coding signatures for all named genes and for hypothetical ORFs 3a, 6, 7a, 7b, and 8, indicating protein-coding roles, and provide strong evidence of protein-coding status for a recently-proposed alternate-frame novel ORF within 3a. By contrast, ORF10 shows no protein-coding signatures but shows unusually-high nucleotide-level constraint, indicating it has important but non-coding functions, and ORF14 and SARS-CoV-1 ORF3b, which overlap other genes, lack evolutionary signatures expected for dual-coding regions, indicating they do not produce functional proteins. ORF9b has ambiguous protein-coding signatures, preventing us from resolving its protein-coding status. ORF8 shows extremely fast nucleotide-level evolution, lacks a known function, and was deactivated in SARS-CoV-1, but shows clear signatures indicating protein-coding function worthy of further investigation given its rapid evolution and potential role in replication. SARS-CoV-2 mutations are preferentially excluded from evolutionarily-constrained amino acid residues and synonymously-constrained nucleotides, indicating purifying constraint acting at both coding and non-coding levels. In contrast, we find a conserved region in the nucleocapsid that is enriched for recent mutations, which could indicate a selective signal, and find that several spike-protein mutations previously identified as candidates for increased transmission and several mutations in isolates found to generate higher viral load in-vitro disrupt otherwise-perfectly-conserved amino-acids, consistent with adaptations for human-to-human transmission.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for the COVID-19 pandemic (Wu et al. 2020), is a betacoronavirus in the subgenus Sarbecovirus, which also includes SARS-CoV-1 (also known as SARS-CoV), the strain responsible for the 2003 SARS outbreak. The large and complex positive-strand RNA genome of SARS-CoV-2 consists of approximately 30,000 nucleotides, and encodes approximately 30 known or hypothetical mature proteins (**Fig. 1A, Fig. 3**). Despite its extreme medical importance, its gene content remains surprisingly unresolved, with several hypothetical open reading frames (ORFs) whose function or even protein-coding status is unknown.

More than two-thirds of the SARS-CoV-2 genome is spanned by a large open reading frame (ORF1ab), which includes an internal programmed translational frameshift triggered by a translation-slippery sequence UUUAAAC and a downstream RNA pseudoknot structure that is generally conserved among coronaviruses

(Baranov et al. 2005). Translation of ORF1ab yields a large polyprotein that is cleaved into non-structural proteins nsp1-nsp10 and nsp12-nsp16. When the frameshift does not occur (ORF1a) translation terminates at a stop codon four codons past the frameshift site and the product is cleaved into nsp1-nsp11. Mature proteins encoded by ORF1 include an RNA-dependent RNA polymerase (Pol); a helicase (Hel); and proteins involved in viral transcription, proofreading (ExoN), translation, cleavage (3CL-PRO), assembly, and suppression of host cell and immune system function (Supplemental Table S2).

Since the virus uses the human translation machinery, which translates the first ORF of a given transcript, in order to translate genes in the remaining third of the genome it generates varying-length subgenomic RNAs (Miller and Koev 2000), believed to result from RNA-dependent transcription of the positive strand genomic RNA to a negative strand RNA molecule beginning at the 3' end and continuing until a transcription-regulatory sequence (TRS), followed by transcription from a common leader on the 5' end, and a second round of RNA-dependent negative-to-positive RNA-dependent transcription (Kim et al. 2020).

The named genes in the last third of the genome encode the spike surface glycoprotein S (ORF2), which is cleaved into S1 and S2 and is responsible for viral attachment and entry by binding the human ACE2 receptor; the envelope protein E (ORF4) and membrane glycoprotein M (ORF5), responsible for virus morphogenesis and assembly; and the nucleocapsid phosphoprotein N (ORF9), responsible for packaging the RNA genome.

The remaining ORFs are unnamed, were annotated primarily by homology and prediction algorithms rather than functional evidence, and are subject to disagreement on which encode functional accessory proteins. NCBI annotates the SARS-CoV-2 reference genome NC_045512.2 as containing ORFs 3a, 6, 7a, 7b, 8, and 10. UniProt annotates two additional ORFs, 9b and 14, both of which overlap the nucleocapsid phosphoprotein N in a different reading frame. The Nature paper that introduced SARS-CoV-2 shows all of these, and also 3b, which overlaps 3a in a different reading frame in SARS-CoV-1 but is not an open reading frame in SARS-CoV-2 due to several in-frame stop codons, and refers to UniProt ORFs 9b and 14 as 9a and 9b, respectively (Wu et al. 2020), but the most recent GenBank record of the paper, MN908947.3, does not include 3b, 9b, or 14 and also lacks 7b. A recent Lancet paper refers to UniProt 3a as 3, 6 as 7, 7a as 8, 7b as 9, 8 as 10b, 9b as 13, and is missing 10, but includes 14 (Lu et al. 2020). For consistency, we use the UniProt numbering here. Orthologs of ORFs 3a, 6, 7a, 7b, and 9b, are also annotated in the NCBI reference genome NC_004718.3 for SARS-CoV-1, but ORF8 is split into 8a and 8b, 3b is included, and neither 14 nor 10 are included. Supplemental Table S2 includes a summary of what is known about each proposed SARS-CoV-2 ORF and mature protein product, extracted from the UniProt annotations.

Several high-throughput experimental techniques have been used to try to determine the protein-coding content of the SARS-CoV-2 genome. Proteomics experiments identified peptides for a subset of ORFs: 1ab, S, 3a, M, 6, 7a, 8, N, and 9b, but not E, 7b, 14, or 10 (Davidson et al. 2020; Bojkova et al. 2020). Direct-RNA sequencing found subgenomic RNAs indicating translation potential for a different subset: S, 3a, E, M, 6, 7a, 7b, 8, and N, but with limited or no support for 9b, 14, and 10 (Kim et al. 2020; Taiaroa et al.; Davidson et al. 2020), and subgenomic RNAs of 7b, which is thought to be translated from subgenomic RNAs of 7a by leaky scanning (Schaecher et al. 2007), were found at relatively low levels. Ribosome profiling using lactimidomycin and harringtonine to identify translation initiation sites predicted translation of ORFs 1ab, S, 3a, E, M, 6, 7a, 7b, 8, N, 9b, and 10, as well as ten novel ORFs overlapping annotated ORFs in another frame, but did not find ORF14 (Finkel et al. 2020). However, such experimental approaches only detect what is present under the specific conditions tested, and thus cannot argue for non-functionality of an ORF due to lack of detection. For example, no peptides were detected for envelope protein E in any previous study (Davidson et al. 2020; Bojkova et al. 2020), even though its function is well-established, making it difficult to reject it, or any hypothetical ORFs, simply due to lack of experimental evidence. Moreover, given the large number of viral RNA molecules in each cell, detection of a transcript, ribosome attachment, or even a translated peptide may simply reflect incidental transcriptional and translational events, rather than adaptive function. For example, only one supporting transcript was found for ORF10 (Kim et al. 2020; Davidson et al. 2020) and the region orthologous to SARS-CoV-1 3b (Davidson et al. 2020), and only two transcripts were found for ORF14 (Davidson et al. 2020), compared to thousands for subgenomic RNAs of other ORFs. (Note: Davidson et al. refer to UniProt ORFs 9b as 9a, and 14 as 9b.)

Another critical research goal is distinguishing which of the many variants that have arisen during the COVID-19 pandemic affect the viral phenotype or its response to therapies. As of this writing, over 17,000 isolates of SARS-CoV-2 have been sequenced, revealing over 1800 variants within the SARS-CoV-2 population (Elbe and Buckland-Merrett 2017; Hadfield et al. 2018). Techniques for distinguishing which of these variants are most likely to have a functional effect can help prioritize experimental and epidemiological studies.

Here, we address these challenge by carrying out a systematic comparative genomics analysis of the SARS-CoV-2 genome in the context of its closely-related complete genomes (**Fig. 1**), in order to to determine which of the uncharacterized ORFs in SARS-CoV-2 code for functional proteins, and to distinguish which SARS-CoV-2 variants are most likely to have functional and potentially medical importance.

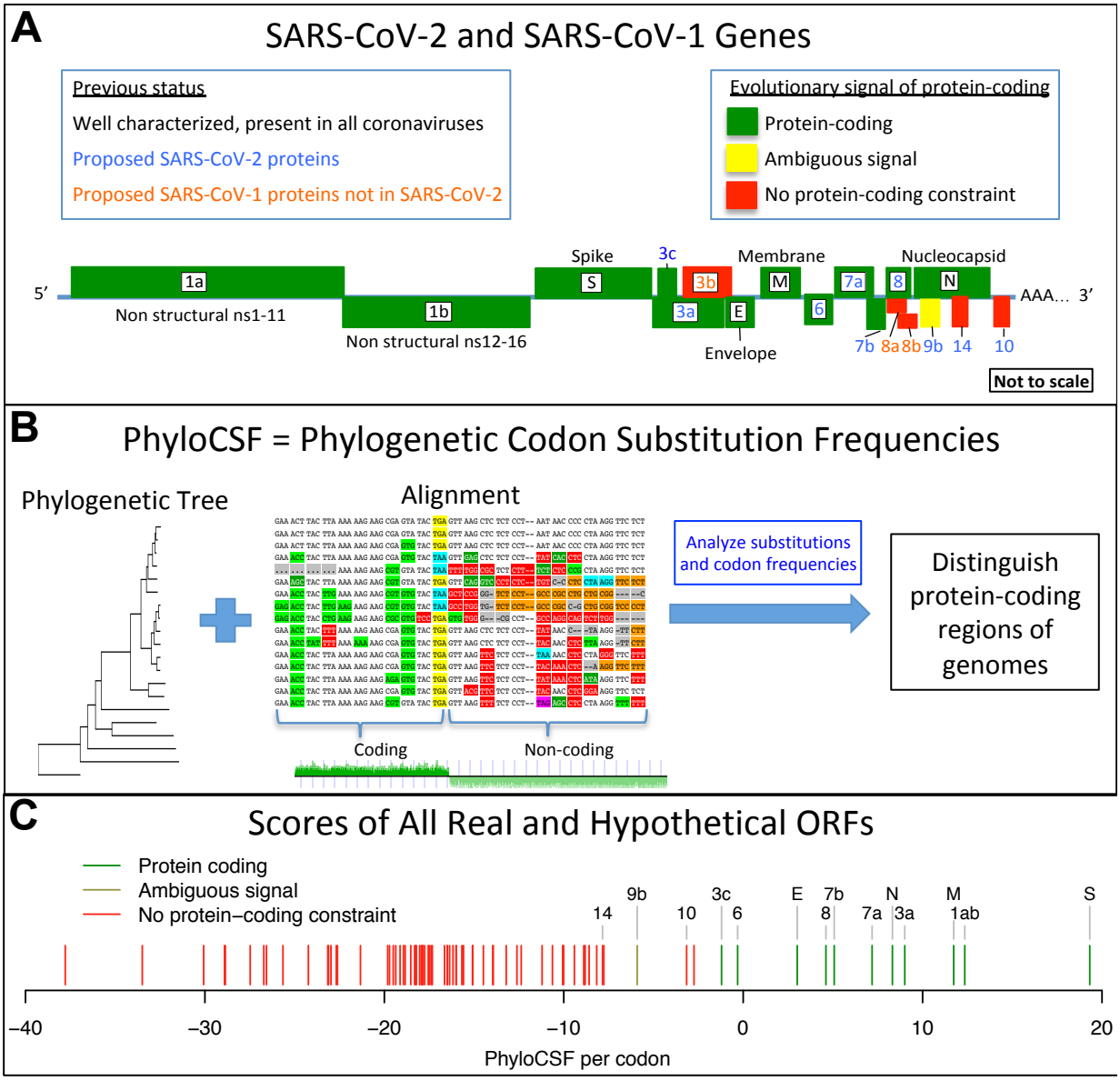


Figure 1. Summary. (A) Graphical representation of annotated SARS-CoV-2 and SARS-CoV-1 genes, with an indication of which ones are supported by the evolutionary evidence. ORFs 1a, 1b, S, 3a, E, M, 6, 7a, 7b, N, recently-proposed 3c, and possibly 9b show evolutionary signatures of being conserved protein-coding regions in both SARS-CoV-2 and SARS-CoV-1, as well as ORF8 in SARS-CoV-2, whereas ORFs 3b, 8a, 8b, 14, and 10 are not supported by the evolutionary evidence. Layout adapted from Fig. 13.4, Human Virology (Oxford et al. 2016). (B) PhyloCSF uses substitutions and codon frequencies in an alignment of genomes at an appropriate evolutionary distance to quantify the evolutionary signatures that distinguish conserved, functional, protein-coding regions. (C) PhyloCSF scores of all annotated and hypothetical AUG-initiated ORFs on the positive strand at least 25 codons long that do not overlap a longer ORF in the same frame.

We select 44 complete and closely-related coronavirus genomes at ideally-suited evolutionary distances (approximately that of mammalian species), generate whole-genome alignments spanning all known genes and hypothetical ORFs, and use them to evaluate protein-coding constraint in all reading frames, and nucleotide-level constraint in synonymous codon positions. We find that five hypothetical ORFs are not conserved protein-coding genes, namely ORFs 10 and 14, and SARS-CoV-1 ORFs 3b, 8a, and 8b, and we confirm protein-coding evolutionary signatures for other hypothetical ORFs (3a, 6, 7a, 7b, 8) and a recently-proposed alternate-frame ORF within 3a; this includes ORF8 despite its yet-unknown function and extremely-rapid evolutionary rate. We also annotate 1394 synonymously-constrained codons within protein-coding regions, which are indicative of overlapping constrained elements that might include dual coding regions, binding sites for RNA-binding-proteins, and RNA structures known to help regulate coronavirus replication, transcription, and translation. We use these protein-level and codon-level annotations to classify 1800 single-nucleotide variants across 17,000 isolates from the current pandemic, yielding insights into mutations that are likely benign vs. those that disrupt evolutionarily-conserved protein-coding or non-coding functions. In particular, we find that several spike-protein variants recently-associated with increased transmission disrupt perfectly-conserved amino-acids, possibly representing novel adaptations to human hosts. These comparative genomics annotations provide a general resource for prioritizing functional variants and strains, for vaccine development and specialization, and for untangling the molecular biology of SARS-CoV-2.

Results

Species selection and alignment of 44 Sarbecovirus genomes

We selected 44 complete Sarbecovirus genomes at an evolutionary distance well-suited for identifying protein-coding genes and non-coding selection within them, consisting of SARS-CoV-2, SARS-CoV-1, and 42 bat coronavirus genomes (**Fig. 2, Supplemental Table S1**). Betacoronavirus genomes outside the Sarbecovirus clade, such as MERS-CoV, are too different from SARS-CoV-2 to be usable for this purpose, and even the closest relative, Hibecovirus Bat Hp-betacoronavirus/Zhejiang2013, shows no detectable homology across ORFs 6, 7a, 7b, and 8. Conversely, among different isolates of the SARS-CoV-2, SARS-CoV-1, and some bat strains, evolutionary distances are too small for reliable evolutionary signatures.

Figure 2. Phylogenetic tree of 44 Sarbecovirus genomes. Left: Phylogenetic tree of a selection of Coronaviridae genomes, including the seven that infect humans (red asterisks). Right: Phylogenetic tree of the 44 Sarbecovirus genomes used in this study. Trees are based on whole-genome alignments and might be different from the history at particular loci, due to recombination.

We created a genome-wide alignment of these 44 genomes, spanning a total phylogenetic branch length of about 3 substitutions per 4-fold degenerate site, comparable to the 29-mammals and 12-flies alignments we previously used for protein-coding gene identification. This rate varies across the SARS-CoV-2 genome from 0.2 in ORF10, 1.2 in E, and 6.5 in S (Supplemental Table S2), though the latter may be inflated due to genomic segments whose histories do not match the whole-genome tree.

Scoring of protein-coding and synonymous constraint

To detect protein-coding evolutionary signatures and distinguish regions that evolved under protein-coding constraint, we previously developed PhyloCSF (Lin et al. 2011a), which compares codon substitutions and frequencies in alignments of closely-related genomes to coding and non-coding evolutionary models trained on whole genome data (**Fig. 1B**), and CodAlignView (I Jungreis, MF Lin, CS Chan, M Kellis 2016) to enable manual curation by visual exploration of the corresponding alignment for substitutions, stop codons, and insertions or deletions indicative of the protein-coding status of a region. These tools have been widely used to identify novel protein-coding regions in human (Lindblad-Toh et al. 2011; Mudge et al. 2019), fly (Lin et al. 2007), and yeast (Lin et al. 2011a), to discover stop-codon readthrough (Jungreis et al. 2016; Lindblad-Toh et al. 2011; Jungreis et al. 2011; Lin et al. 2007; Loughran et al. 2014), and to distinguish protein-coding vs. non-coding genes in human (McCorkindale et al. 2019; Frankish et al. 2019).

We computed PhyloCSF protein-coding scores for every three-nucleotide interval, in all three reading frames of SARS-CoV-2, smoothed using a hidden Markov model, and created tracks for the UCSC Genome Browser quantifying protein-coding evolutionary signatures along the genome, as we previously did for the human genome (Mudge et al. 2019). We also computed an overall PhyloCSF score for each known protein and hypothetical ORF (Supplemental Table S2). We used CodAlignView to create visualizations of the alignment of each ORF, highlighting two signatures of mutations that are tolerated in protein-coding genes across evolutionary time: first, a preference for synonymous substitutions typical of third codon positions and conservative amino acid changes that preserve biophysical properties; second, avoidance of insertions and deletions that are not multiples of 3, as they would disrupt the reading frame of translation, whereas gaps that remove complete codons and preserve the reading frame are more tolerated. We have provided CodAlignView images (**Supplemental Materials**) and links for manual exploration (**Supplemental Table S2**) for each annotated ORF and mature protein.

We also previously developed FRESCo and other software tools for detecting overlapping nucleotide-level constraint within protein-coding regions (Lin et al. 2011b; Sealfon et al. 2015), evidenced by fewer synonymous substitutions, and reflective of overlapping functional elements. Such elements can include dual-coding regions that encode multiple proteins in different reading frames, which are common in viruses with compact genomes (Firth 2014) but also found in other species including human (Lin et al. 2011b; Khan et al. 2020); RNA structures encoded through complementary nucleotide stretches, which are known to play important roles in subgenomic RNA generation and other coronavirus functions; and binding sites for RNA-binding proteins, which can be recognized by virus-encoded proteins or host-encoded proteins, to regulate transcription, processing, and translation of viral mRNAs. FRESCo has been applied to diverse virus species (Sealfon et al. 2015) as well as humans (Khan et al. 2020).

We used FRESCo to calculate the rate of synonymous substitutions in each codon of our alignment, and to recognize synonymous constraint elements (SCEs) within each NCBI-annotated SARS-CoV-2 gene, based on significantly-decreased synonymous rate in 9-codon windows relative to the gene average (**Fig. 3**).

Comparative evidence of protein-coding constraint for nsp proteins and named genes

We found a clear PhyloCSF signal for nsp1-nsp10 and for nsp12-nsp16 (Supplemental Table S2), with a change in the indicated translation reading frame at the known programmed frameshift site (**Fig. 3A**). For the 13-codon nsp11 ORF, the first 9 codons are in the same reading frame as nsp12 (Pol), and the remaining 4 codons are perfectly conserved in Sarbecovirus (**Supplemental Fig. S1A**), but poorly conserved in

betacoronaviruses beyond Sarbecovirus (**Supplemental Fig. S1B**), suggesting that the 13-amino-acid peptide is not performing a conserved function.

The named genes E, M, and N are well-conserved across the 44 Sarbecovirus genomes, with strong overall alignment and very strong PhyloCSF scores, as expected. The S protein shows an unusual evolutionary signal that indicates a history of extremely-rapid evolution, subject to frequent substitutions and recombinations across its phylogeny, resulting in near-zero nucleotide-level conservation scores as measured by phyloP (Pollard et al. 2010) and phastCons (Siepel et al. 2005) over its first half (S1), while the second half (S2) is well-conserved (**Fig. 3a**). However, for protein-coding constraint, both S1 and S2 show very strong PhyloCSF scores, indicating that despite its rapid evolution, S remains strongly selected to preserve a protein-coding function, and highlighting the power of PhyloCSF to recognize protein-coding constraint despite rapid nucleotide evolution.

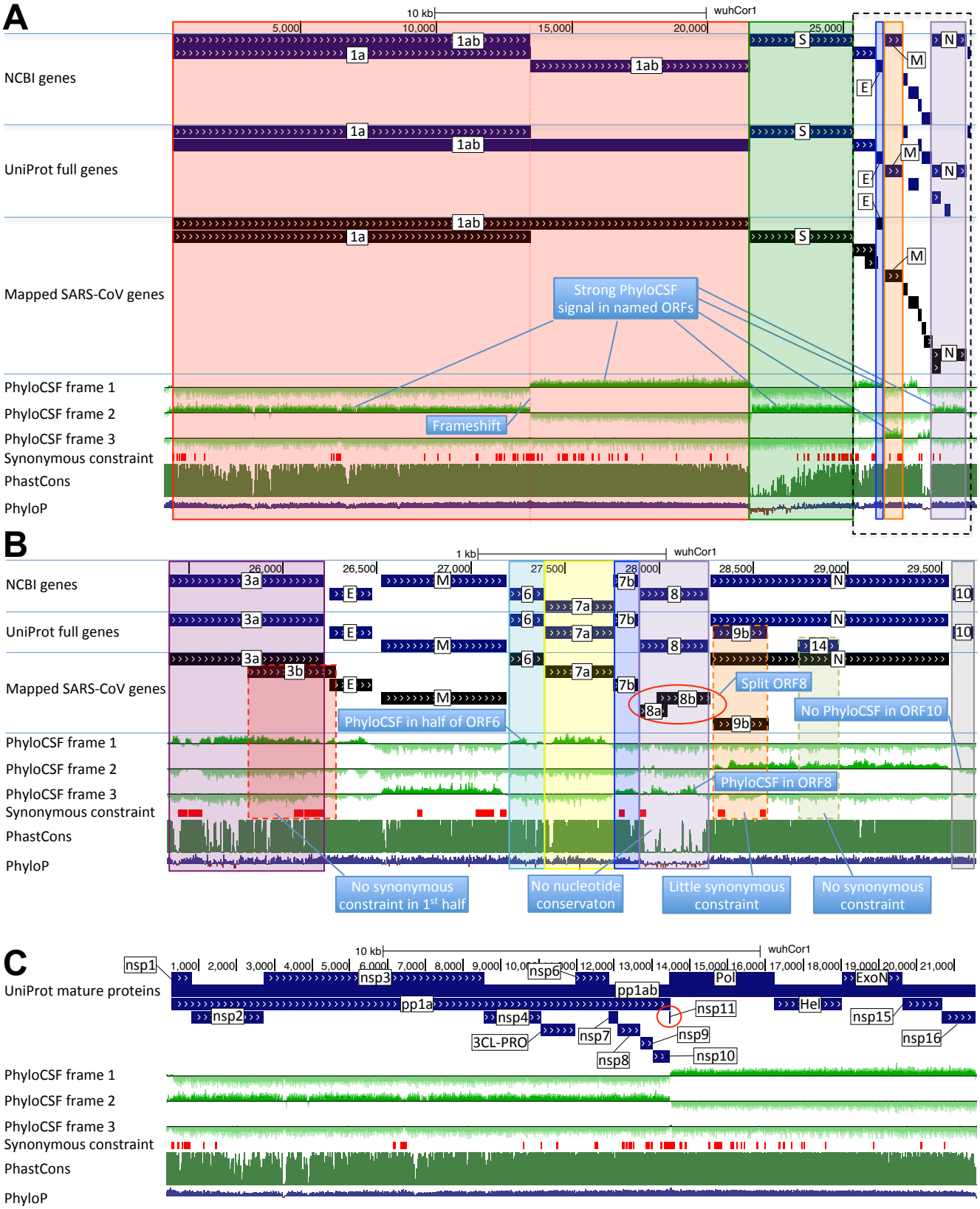


Figure 3. **PhyloCSF signal for SARS-CoV-2 and SARS-CoV-1 ORFs.** UCSC Genome Browser images of SARS-CoV-2 genome. Tracks, from top to bottom, are NCBI genes and UniProt genes for SARS-CoV-2, NCBI genes for SARS-CoV-1 mapped to the SARS-CoV-2 genome, PhyloCSF tracks for each of 3 reading frames, Synonymous Constraint Elements (SCEs), and phastCons and phyloP tracks showing nucleotide-level constraint. Part A shows the entire genome, part B shows the 3' end indicated by the dashed box in part A, and part C

shows the polyprotein at the 5' end of part A. (A) There is a strong PhyloCSF signal in the correct frame for each of the named genes 1ab (polyprotein, red), S (spike, green), E (envelope, blue), M (membrane, orange), and N (nucleocapsid, purple), confirming that they have been under protein-coding constraint in the Sarbecovirus clade. The signal changes frame as expected at the programmed frameshift site in 1ab. There is a strong PhyloCSF signal throughout S despite the lack of nucleotide conservation at the 5' end. (B) There is a clear PhyloCSF signal in the correct frame for unnamed ORFs 3a (dark purple), 7a (yellow), 7b (blue), and 8 (light purple), despite the complete lack of nucleotide conservation in 8. There is a clear signal in the 5' half and 3' quarter of 6 (cyan), but it is weaker in the third quarter of the protein, indicating that this portion has been less constrained. There is no signal for 10 (gray), indicating that it is not a conserved protein-coding region despite high nucleotide conservation. ORFs 9b (dashed orange) and 14 (dashed green) overlap the nucleocapsid phosphoprotein N in an alternate reading frame. If these were conserved coding regions, we would expect to find synonymous constraint through most of the ORF and, possibly, some PhyloCSF signal in the alternate frame, but there is no such PhyloCSF signal in either, and there are no synonymous constraint elements in ORF14. There are two small synonymous constraint elements for 9b, leaving its status as a functional ORF ambiguous. The SARS-CoV-1 annotations also include 3b (dotted red) overlapping 3a in another frame, but most of the ORF is not synonymously constrained and it contains numerous stop codons in other strains so it cannot be a conserved coding region. A frameshifting deletion in 8 occurred in SARS-CoV-1 during the SARS outbreak, creating fragments 8a and 8b (red oval) in some isolates, but there was insufficient evolutionary time for our methods to distinguish if the fragments were still protein-coding. (C) The polyprotein, 1ab, is processed into 16 mature peptides. The PhyloCSF signal shows that all are functional proteins except possibly nsp11 (red circle), which is only 13-amino-acids long and shares its first eight codons with Pol before the latter shifts to a different reading frame.

ORFs 3a, 6, 7a, 7b, and 8 are protein-coding, but ORF10 is a non-coding functional element.

Among the six unnamed ORFs annotated by NCBI (Supplemental Table S2), we found clear positive PhyloCSF scores for 3a, 7a, 7b, and 8, indicating conserved protein-coding regions, functional at the amino-acid level (**Fig. 3**). For ORF6, the first half and last quarter shows strong PhyloCSF signal (**Fig. 3B**), indicating that it encodes a conserved, functional protein, despite a less-constrained intermediate portion, and an overall near-zero average score per codon (-0.3, Fig. 1C).

ORF8 shows near-zero nucleotide-level conservation scores as measured by phyloP and phasCons, and it has no well-established function, suggesting at first glance that it might be non-functional. However, PhyloCSF shows a positive protein-coding signal (average score 4.61 per codon), and long stretches of strong protein-coding conservation, indicating that it produces a functional protein despite its rapid nucleotide-level evolution. The apparent high rate of nucleotide-level evolution in ORF8 (3.9 substitutions per site, 6.2 per 4-fold degenerate site) is in part an artifact of its history of recombination events that result in a different tree from the rest of the genome (**Supplemental Fig. S2**), but even after computing rates in a tree representing the history of ORF8, its rate continues to be very high (2.1 and 3.7, respectively) compared to other ORFs (e.g. 1.1 and 2.8 respectively for ORF1ab when using the whole-genome tree excluding two strains that have no alignment in ORF8, to ensure an apples-to-apples comparison, Supplemental Table S2).

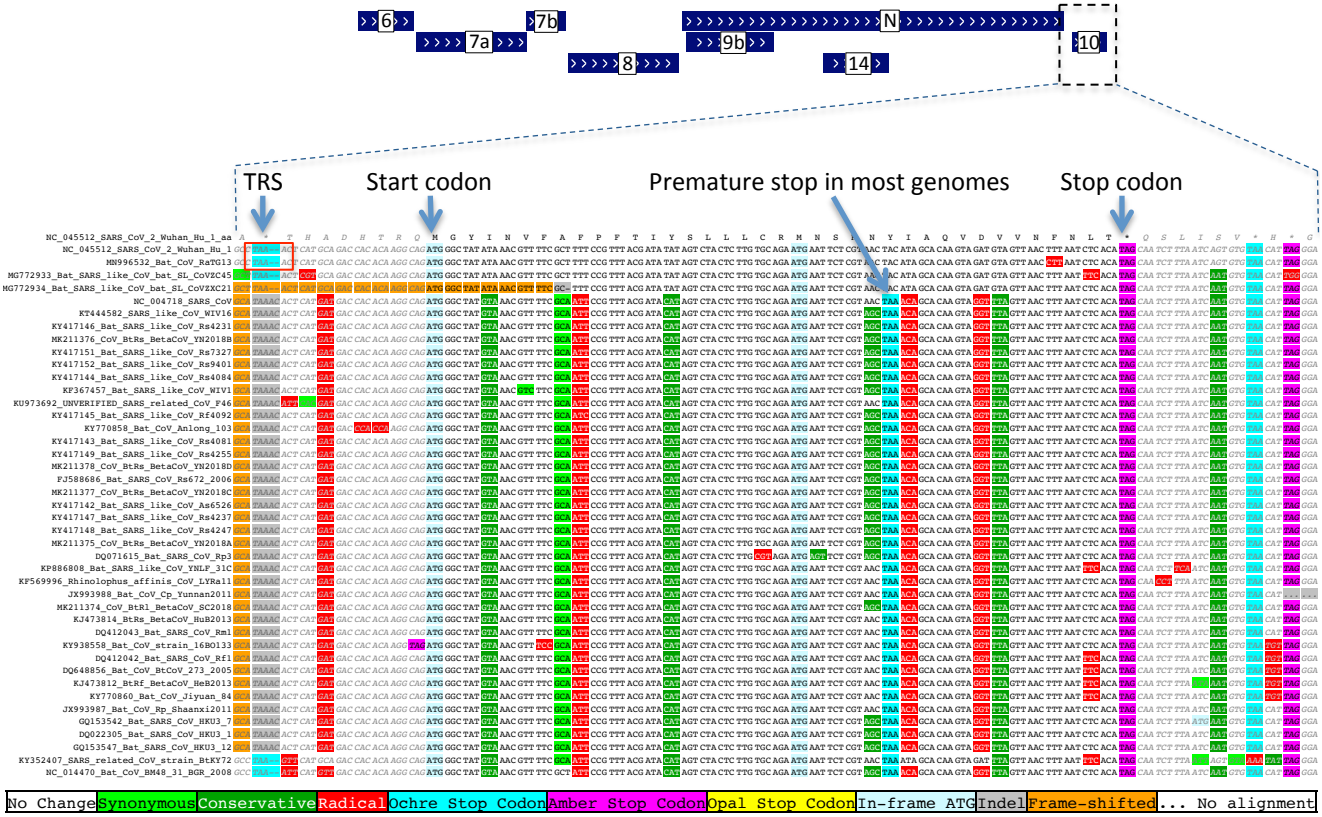


Figure 4. Alignment of ORF10. Alignment of Sarbecovirus genomes at ORF10, including 30 additional nucleotides on each end. Most substitutions are conservative (dark green) or radical (red) amino acid changes, rather than the synonymous (light green) changes expected in protein-coding regions, and there is a premature stop codon (cyan) in most strains, indicating that this is not a conserved protein-coding region. It has extremely high nucleotide-level conservation, which extends beyond the putative ORF in both directions, indicating that this portion of the genome is functionally important even though it does not code for protein. A putative partial transcription-regulatory sequence (TRS) is present only in SARS-CoV-2 and its closest relative, Bat CoV RaTG13, indicating that it is not conserved.

By contrast, ORF10 shows no protein-coding signal anywhere along its length and contains an in-frame premature stop codon in all but four Sarbecovirus genomes, truncating the last third of this already-short (38 amino acid) ORF, indicating that ORF10 does not encode a conserved protein (**Fig. 4**). ORF10 shows near-perfect nucleotide-level conservation that extends beyond the ORF on both sides, as measured by phastCons and phyloP (**Fig. 3B**), indicating that this genomic region is performing some important function despite not coding for protein.

Our conclusion contrasts with recent papers that instead suggested ORF10 is protein-coding. First, a search for transcription-regulatory sequences (TRS) in the original paper reporting the SARS-CoV-2 genome (Wu et al. 2020) found a partial match, with the nucleotides CUAAAC, 22 nucleotides before the start codon of ORF10 (**Fig. 4**); however, this sequence is not conserved, has an intervening ATG, and is only found in SARS-CoV-2 and its closest relative, Bat RaTG13. Indeed, experimental studies using direct-RNA sequencing or proteomics found little or no evidence for expression of an ORF10 subgenomic RNA even in SARS-CoV-2 (Kim et al. 2020; Taiaroa et al.; Davidson et al. 2020; Bojkova et al. 2020), indicating that it is not simply a recent innovation, but possibly a false positive. Second, ribosome profiling data (Finkel et al. 2020) detected footprints within ORF10, leading to a conclusion that it is translated; however, nearly all footprints detected within ORF10 are in either a uORF that overlaps its start codon or a downstream ORF beginning at an interior AUG, which would create a peptide of only 18 amino acids in SARS-CoV-2 (five amino acids in most other Sarbecovirus strains), and the density of footprints within the unique portion of ORF10 is no greater than after its stop codon (**Supplemental Fig. S3A**). Third, a high ratio of nonsynonymous to synonymous substitutions in ORF10 was used as evidence that the ORF is protein-coding and under positive selection for rapid protein evolution (Cagliani et al. 2020); however, this analysis was based on only nine substitutions (one of which was

synonymous), was not statistically significant (nominal p-value>0.18, even without the needed multiple hypothesis correction), only used five closely-related genomes, and excluded a sixth genome that contained a frameshifting indel that would provide strong evidence against protein-coding function if it is not a sequencing error (**Supplemental Fig. S3B**). Overall, the prior evidence is insufficient to argue for protein-coding function for ORF10, and thus we conclude that ORF10 is not protein-coding, given our strong comparative genomics evidence against protein-coding constraint.

ORF14 is not a conserved coding region, and ORF9b is ambiguous

We next investigated the coding potential of two additional hypothetical ORFs annotated by UniProt, ORF 9b (97 amino acids) and ORF14 (73 amino acids), which overlap the nucleocapsid phosphoprotein (N) in a different reading frame. In neither case is there a PhyloCSF signal in the alternate frame (**Fig. 3B, Supplemental Table S2**). While dual coding regions often contain segments having a PhyloCSF signal in the alternate frame, such as those in human *POLG* (Khan et al. 2020), the lack of such signals does not provide a definite negative answer because coding constraint in the main frame alters the pattern of substitutions in the alternate frame, which can depress the PhyloCSF score. Instead, we examined the rate of synonymous substitutions in the reading frame of the nucleocapsid protein, since coding in the alternate frame would be expected to impose overlapping constraint that suppresses synonymous substitutions in the main frame. We find no significant SCEs within ORF14 (**Fig. 5**). Furthermore, its start codon is lost in one strain, and most strains have a stop codon three codons before the ORF14 stop (**Supplemental Fig. S4**). Nor were the subgenomic RNA fragments needed to express ORF14 found in the above-mentioned direct-RNA sequencing experiments (Kim et al. 2020; Taiaroa et al.). We conclude that ORF14 does not encode a functional protein.

The evolutionary evidence for ORF9b is more ambiguous. On the one hand, we do not find significant synonymous constraint in most of the portion of the main frame overlapping 9b, and some segments have synonymous level well above the gene-wide average of the nucleocapsid protein (**Fig. 5**); on the other hand, this region does contain two small SCEs. We note that FRESCo calculates synonymous constraint relative to the gene-wide average, and N has fewer synonymous substitutions per 4-way synonymous site than most of the rest of the genome (Supplemental Table 2), making it more difficult for a constrained region to achieve significance. Both the start and stop codons of 9b are perfectly conserved among our 44 Sarbecovirus strains (**Supplemental Fig. S5**), but conservation of these codons could be due primarily to constraint on the amino acid sequence of the overlapping nucleocapsid protein rather than any constraint on ORF9b itself; indeed, there is only one hypothetical single nucleotide change to the start codon of 9b, ATG->ACG, that would preserve the amino acid sequence of N, and no such changes to its stop codon. There are no premature stop codons in any of the other strains, though again that provides only weak evidence that 9b is coding because 9b is short enough that this could occur by chance. Finally, the start codon of 9b has a strong Kozak context, with A in position -3 and G in position +4, which are believed to be the optimal nucleotides at these positions for ribosomal recognition of the start codon. In contrast, the start codon of the nucleocapsid phosphoprotein, which is only 10 nt 5' of the start codon of 9b, has a weaker Kozak context, with A in position -3 and U in position +4. This leaves open the possibility that the ribosome might initiate translation of 9b from the same subgenomic RNA as N via leaky scanning, making these two proteins in a fixed ratio. If 9b does encode a protein, the high synonymous rate in some overlapping segments of the main frame would indicate that the amino acid sequence encoded by the corresponding segments in 9b are poorly constrained. Proteomics experiments have detected the hypothetical protein product of ORF9b (Davidson et al. 2020; Bojkova et al. 2020), and there is experimental evidence that ORF9b in SARS-CoV-1 localizes to mitochondria and interferes with host cell antiviral response (Shi et al. 2014), so without clear evolutionary evidence one way or the other it seems likely that ORF9b does produce a functional protein, though one with a poorly-conserved amino acid sequence.

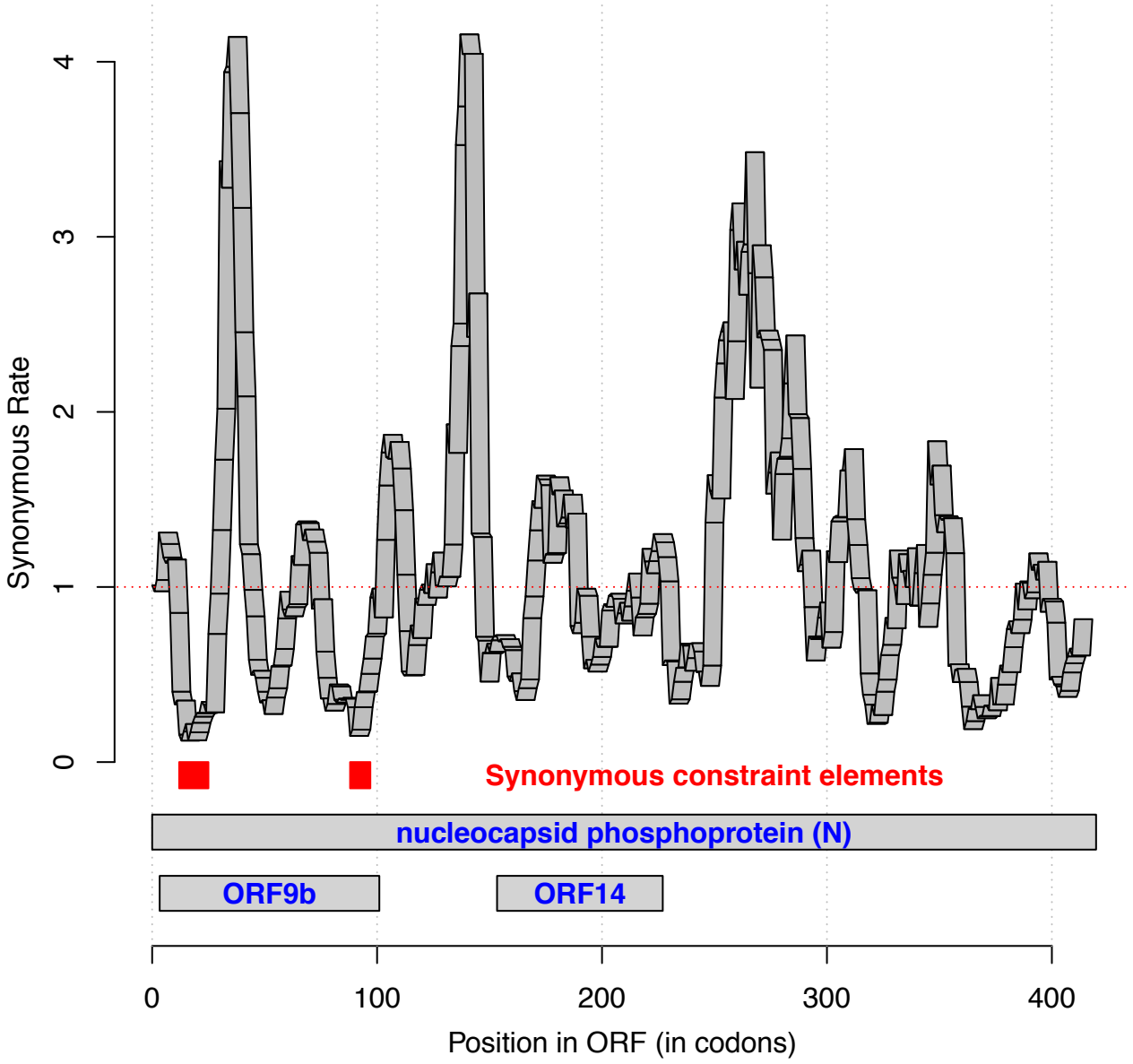


Figure 5. Synonymous Rate in nucleocapsid phosphoprotein. Rate of synonymous substitutions in 9-codon windows within the nucleocapsid phosphoprotein (N), normalized to make the gene-wide average 1. ORFs 9b and 14 (bottom gray rectangles) are hypothetical protein-coding regions within N in a different reading frame. We expect dual coding regions to be synonymously constrained, but there are no significant synonymous constraint elements in ORF14, and only two small ones in ORF9b (red), making it unlikely that ORF14 is a true protein-coding region, and leaving the status of ORF9b ambiguous.

A novel alternate frame protein-coding ORF within ORF3a

We next searched for novel conserved protein-coding regions by scoring all 67 hypothetical AUG-to-stop SARS-CoV-2 ORFs of at least 25 codons that do not overlap an NCBI-annotated ORF in the same frame and that are not contained in a longer ORF in the same frame. None of these had a positive PhyloCSF score, but we investigated the top candidates with the least negative score based on conservation of the start and stop codons, absence of in-frame stop codons and frameshifting indels, and evidence of synonymous constraint in the overlapping coding region.

The candidate with the highest PhyloCSF score per codon (-1.21) is a 41-codon ORF (positions 25457-25579), that overlaps ORF3a in an alternate frame near its 5' end (**Fig. 6**). Although the score is negative, it is 2.57 standard deviations higher than the average over hypothetical non-coding ORFs (mean: -17.9, stdev: 6.5, p = 0.005 under normal approximation, **Fig. 1C**), but closer to the distribution of protein-coding ORFs (mean: 8.03, stdev: 5.55, deviation: -1.67 standard deviations). As this ORF overlaps a known coding gene in an alternate frame, constraint on the known amino acid sequence suppresses synonymous substitutions in the alternate frame, which lowers the PhyloCSF score, so we would expect a lower PhyloCSF score than for non-overlapping protein-coding regions that are subject to the same level of protein-coding constraint. Moreover, the AUG start codon is perfectly conserved except in one strain that has the near-cognate GUG instead, and the stop codon is conserved but with a one-codon extension in SARS-CoV-2 and RaTG13. There are also no in-frame stop codons or indels. Strikingly, this alternate-frame ORF has many synonymous substitutions that are non-synonymous in ORF3a, indicating that this new ORF may be the primary constraint acting in this region, over the corresponding segment of ORF3a. Lastly, 40 of the 41 codons are covered by synonymous constraint elements, and this constraint ends nearly perfectly at the boundaries of the overlapping ORF (**Fig. 6**). Together, these lines of evidence allow us to conclude that this overlapping ORF encodes a conserved, functional protein.

Two previous studies proposed that this new ORF may be protein-coding on the basis of increased synonymous constraint on the overlapping region of ORF3a, initially across 6 very closely-related strains (Cagliani et al. 2020), and subsequently across a broad set of Sarbecovirus strains (Firth 2020), naming it ORF3h (for "Hypothetical") and ORF3a*, respectively. It was predicted to contain a transmembrane domain suggestive of a viroporin (Cagliani et al. 2020), and to be translated from the ORF3a subgenomic RNA via leaky scanning (Firth 2020). However, increased synonymous constraint does not uniquely argue for protein-coding constraint, and could stem from other types of overlapping functional elements. A third study used ribosome footprints to argue this alternate reading frame of ORF3a is translated (Finkel et al. 2020), but a ribosome profiling signal can be due to incidental, non-functional translation; in fact, out of nine candidate novel protein-coding ORFs predicted by this study, eight lack any conservation. By contrast, the well-powered PhyloCSF evidence presented here shows that this ORF has a conserved protein-coding function specifically selected for its amino-acid translation. Given the clear evidence for conserved protein-coding function across Sarbecovirus genomes, including SARS-CoV-2 and SARS-CoV-1, we propose a standard name for this ORF, namely ORF3c, as neither ORF3h, which indicates "hypothetical", nor ORF3a*, which is non-standard, seems appropriate for a deeply-conserved ORF with clear protein-coding function.

The candidate with the next best score (-2.74) is a 32-codon ORF (26183-26278) that overlaps the 3' end of ORF3a and the 5' end of E (**Supplemental Fig. S6A**). Two strains have a frame-shifting 1-base deletion within the ORF, and two others have premature stop codons. None of the substitutions are synonymous. There is high nucleotide-level constraint, but it continues on both sides of the ORF, suggesting it results from something other than translation of the ORF. Overall, this ORF does not show the evolutionary signature of a functional coding sequence. Next in the list is ORF9b which we have already discussed. Fourth is a 31-codon ORF (3207-3299) overlapping ORF1a, having PhyloCSF score -7.77 (**Supplemental Fig. S6B**). Most of the ORF consists of a 75-nt insertion that is only present in SARS-CoV-2, RaTG13, and CoVZC45, and the start and stop codons are missing in CoVZC45, so this is not a conserved coding sequence. Finally, the fifth-ranked candidate is ORF14, which we have already discussed.

The relatively high scores of ORFs 9b and 14 among these 67 hypothetical ORFs are in part an artifact of the low density of substitutions throughout N, which they both overlap. This low density, which is found even in the parts of N that are not in ORFs 9b or 14, decreases the statistical power available to PhyloCSF for distinguishing its coding and noncoding evolutionary models, which compresses the PhyloCSF score towards 0, resulting in a better rank among the negative scores. If we compensate for this by dividing by the maximum-likelihood branch length scale factor computed by PhyloCSF for its coding and non-coding models, ORFs 9b and 14, while still in the top half, move down to the 89th and 79th percentile among the 67 ORFs considered, whereas ORF3c remains the best scoring-candidate (**Supplemental Fig. S7**).

To search for additional novel protein-coding regions, we relaxed our criteria to include ORFs with at least 10

codons, allow non-cognate start codons, and allow ORFs contained within another ORF in the same frame, but we found no additional convincing candidates for conserved protein-coding regions. Because it has been conjectured that translation might occur on the large number of negative-strand genomic and subgenomic RNAs that are intermediates in viral gene expression and replication in positive-strand RNA viruses (Dinan et al. 2020; DeRisi et al. 2019), we also scored ORFs on the negative strand, but again found no convincing candidates. Supplemental Table S4 contains the complete list of ORFs, with scores and other pertinent information.

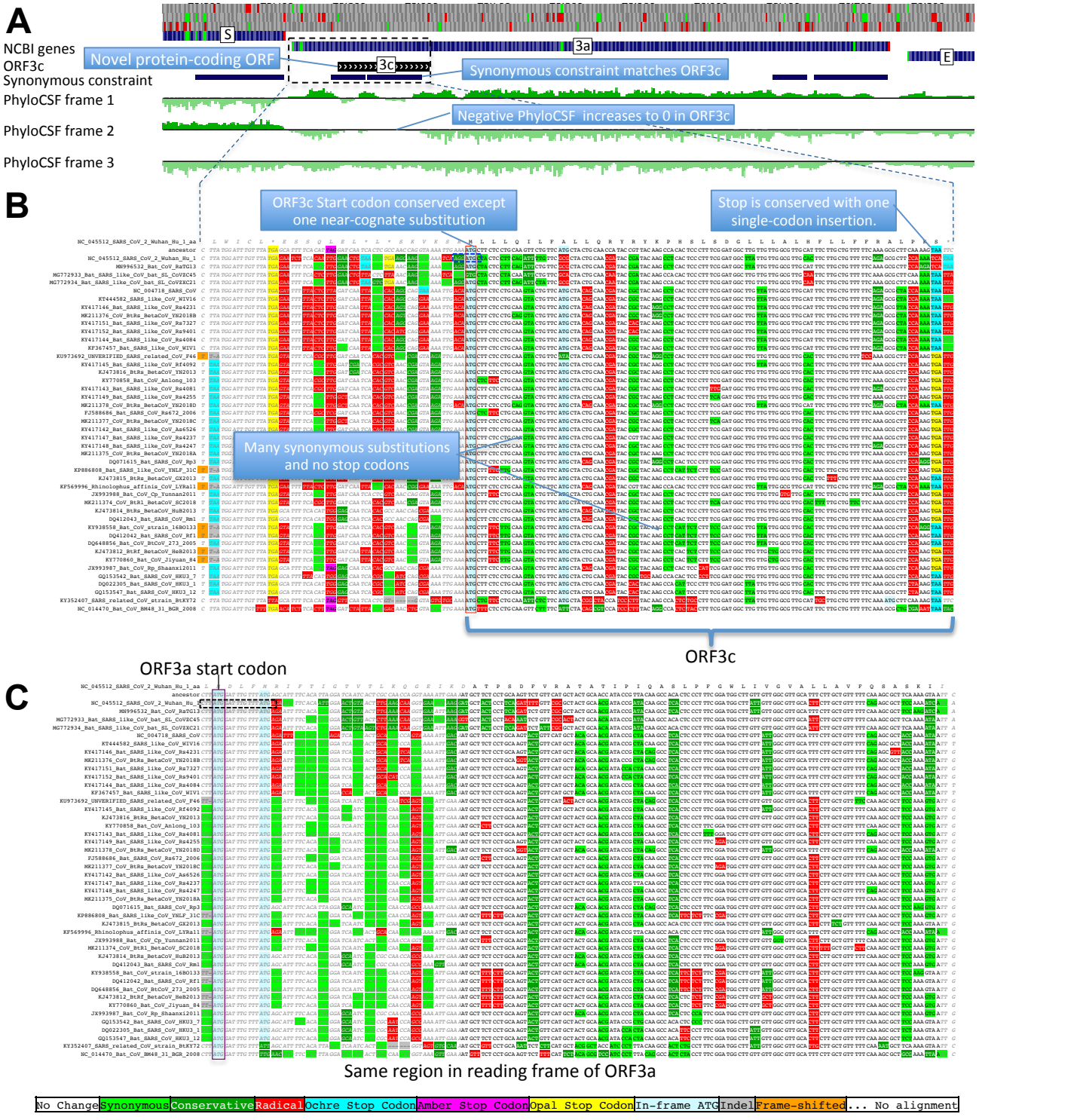


Figure 6. Novel ORF3c. Phylogenetic evidence for an unannotated protein-coding ORF near the 5' end of ORF3a. (A) UCSC Genome Browser image shows ORF3c overlapping ORF3a in a different reading frame. A pair of synonymous constraint elements closely match ORF3c as is expected for a dual coding region. The PhyloCSF signal in the reading frame of ORF3c (frame 2), which is negative for most of ORF3a, is close to 0 within ORF3c, indicating some PhyloCSF signal despite the downward pressure on the PhyloCSF score from constraint in the reading frame of ORF3a. ORF3c might be translated from the ORF3a subgenomic RNA via leaky scanning. (B and C) CodAlignView color-coding of the alignment of the region from 3 nt 5' of the start codon of ORF3a until the stop codon of ORF3c in the reading frame of ORF3c (B) and of ORF3a (C) show that a substantial fraction of substitutions are synonymous (light green) and conservative amino acid changes (dark green) in each frame, as expected in coding regions. The start codon of ORF3c is conserved except in one strain that has the near cognate GUG, and its stop codon is conserved except for a one-codon insertion in two strains, with no intermediate stop codons in any strain.

SARS-CoV-1 3b, 8a, and 8b are not conserved coding genes

We then turned our attention to the three annotated ORFs in SARS-CoV-1 that do not have orthologous ORFs in SARS-CoV-2, namely ORFs 3b, 8a, and 8b. We found that ORF3b, which overlaps 3a, shows poor PhyloCSF protein-coding constraint (score per codon -13.2), and contains numerous stop codons in other strains including SARS-CoV-2, indicating it doesn't have a conserved protein-coding function (**Supplemental Fig. S8**). Finally ORFs 8a and 8b, two fragments of ORF8 that were separated into distinct ORFs during the 2003 SARS outbreak by a 29-nt deletion (**Supplemental Fig. S9**), do not exist as ORFs elsewhere in the Sabercovirus phylogeny, indicating they do not give rise to conserved proteins, and that their previously-reported effect on viral replication (Muth et al. 2018) is likely due to ORF8 loss-of-function rather than 8a/8b gain-of-function.

Sarbecovirus conservation informs analysis of SARS-CoV-2 variants

Finally, we investigated how conservation within the Sarbecovirus clade can help inform our understanding of variation between different isolates of SARS-CoV-2. Since the outbreak of the COVID-19 pandemic, over 1800 single-nucleotide variants (SNVs) have been identified in SARS-CoV-2 isolates. We would expect variants in amino acids or nucleotides that have been highly conserved in the larger clade to be more likely to have a phenotypic effect, so we classified SNVs into five categories according to whether they were intergenic, missense (amino acid changing) in conserved amino acid positions, missense in non-conserved amino acid positions, synonymous in synonymously-constrained codons, or synonymous in synonymously-unconstrained codons (Supplemental Table S3). We defined "conserved" amino acids to be those for which there were no amino acid-changing substitutions in the Sarbecovirus alignment of that codon. We defined codons to be synonymously constrained if they have a low synonymous substitution rate.

To determine if conservation within the Sarbecovirus clade correlates with purifying selection within the SARS-CoV-2 population, we examined the densities of SNVs in conserved and non-conserved positions (**Fig. 7**).

We first calculated the fraction of amino acid positions that were conserved by our definition in each of the mature proteins and hypothetical ORFs (**Fig. 7A**). We observe that more than 83% of amino acids are perfectly conserved in nsp5 (3CL-PRO), nsp7, nsp8, nsp9, nsp10, nsp12 (Pol), nsp13 (Hel), and nsp14 (ExoN), whereas a much lower fraction of amino acids are conserved in nsp1, nsp2, and nsp3. Amino acid conservation in S is high 3' of, and low 5' of, the cleavage site. Amino acid conservation is lower in the unnamed ORFs than the named ones, particularly ORFs 6 and 8. Note that our definition of amino acid conservation does not depend on the phylogenetic tree, so these results are robust even if the tree varies along the genome due to recombination events.

We next calculated the density of missense SNVs among the conserved and non-conserved amino acid positions (**Fig. 7B**), and of synonymous SNVs among synonymously-constrained and unconstrained codons, in each mature protein (**Fig. 7C**). We found that missense SNVs are depleted in conserved amino acid positions (607 SNVs in 6480 conserved positions, 9.4%, versus 535 SNVs in 3264 non-conserved positions, 16.4%, $p < 10^{-10}$) and synonymous SNVs are depleted among synonymously-constrained codons (73 SNVs in 1394 synonymously-constrained codons, 5.2%, versus 555 SNVs in 8350 synonymously-unconstrained codons, 6.6%, binomial $p = 0.029$).

We conclude that conservation in the Sarbecovirus clade at both the amino acid and nucleotide level is associated with purifying selection on SNVs in the SARS-CoV-2 population.

Since SNVs are most likely to have a phenotypic effect if they change a conserved amino acid, we searched for clusters of such SNVs. We found that the region of the nucleocapsid protein encoded by genomic locations 28826 through 28885 is significantly enriched for missense SNVs among its 14 conserved amino acids (14 SNVs, $p < 0.012$ after conservative multiple-hypothesis correction), suggesting that the region has been under positive selection or relaxed purifying selection in SARS-CoV-2 (**Fig. 7D, Supplemental Fig. S10**). There are no other such clusters in the genome that are significantly denser than would be expected by chance. Nor are there any regions that are significantly depleted for missense SNVs in conserved amino acids, which would have indicated regions in which constraint in the Sarbecovirus clade has continued particularly strongly in the

SARS-CoV-2 population; the most depleted regions are 7400-7840 in nsp3 with no missense SNVs among 103 conserved amino acids and 24437-24748 in S2 with no missense SNVs among 99 conserved amino acids ($p = 0.072$ and $p = 0.093$, respectively, without any correction for multiple region lengths searched) (**Supplemental Fig. S11**).

To aid researchers in using our classification of variants, we have created a track hub for the UCSC Genome Browser with each SNV color-coded according to our five categories. The details page for each SNV includes a link to view the alignment of a neighborhood of the SNV using CodAlignView. The track hub also includes tracks showing which codons are conserved at the amino acid and synonymous levels to aid other researchers in classifying SNVs as they are discovered (**Fig. 7D**).

As examples, we analyzed two sets of variants that have been proposed as possibly affecting the viral phenotype. First, we investigated Sarbecovirus conservation of 14 amino acids in the spike protein in which mutations appear to be accumulating in the SARS-CoV-2 population (Korber et al. 2020), namely D614G, L5F, L8V/W, H49Y, Y145H, Q239K, V367F, G476S, V483A, V615I/F, A831V, D839Y/N/E, S943P, P1263L. These are included in the “KorberMutation” column of Supplemental Table S3, with hyperlinks to view the alignment near each of these mutations. Of particular interest is D614G, which has risen in frequency in multiple geographic locations, suggesting that it increases transmissibility. This radical amino-acid change is near the middle of a string of 11 amino acids that are perfectly conserved among our Sarbecovirus genomes (**Fig. 7E**), implying that it would have been deleterious in most of the Sarbecovirus clade; since, to the contrary, it appears to be increasing in the human population, this suggests that it is an adaptation to the human host. Likewise, two others, V615I/F and P1263L are mutations of perfectly conserved amino acids, while A831V is in a highly-conserved region of the protein and its amino acid is conserved in all but the two most distantly-related strains. In contrast, L5F, L8V/W, H49Y, Y145H, Q239K, G476S, and V483A are in amino acids that are not conserved and are in poorly-conserved regions of the protein, so they are less likely to be required for a conserved function. The remaining three are in moderately-conserved contexts with ambiguous interpretation.

Second, we looked at variants from 11 isolates (referred to as ZJU-1 through ZJU-11) that were functionally characterized and found to have different temporal patterns of viral load in-vitro (Yao et al. 2020). Among the 25 loci where at least one of these isolates differs from the reference genome, T27772A is a nonsense mutation that disrupts ORF7b but is present in 7% of the viral RNA in ZJU-11, suggesting that this ORF is not essential for replication. We classified the other 24 according to the evolutionary evidence and found that five are likely to be highly disruptive, another five are somewhat disruptive, four are missense mutations in residues that have been evolutionarily permissive of amino acid changes, and the remaining nine are synonymous in non-synonymously-constrained contexts (Supplemental Table S3). One of the somewhat disruptive mutations is a synonymous change in a 41-codon SCE at the C-terminus of the spike protein, and the other disruptive mutations are missense. Interestingly, one of the highly disruptive mutations, G23607A, is a radical R->Q amino acid change in the polybasic cleavage site of S, which is only present in SARS-CoV-2 (Andersen et al. 2020); it is present in all viral RNA in ZJU-1, whose viral load was near the mean, suggesting that this residue might have little effect on the ability of the virus to gain access to and replicate in cells. The two outliers with unusually high viral load after 24 hours, ZJU-10 and ZJU-11, each have exactly one mutation that we classified as highly disruptive, namely C16114T in ZJU-10 and a trimer substitution TTG->CGA at 27775-27777 in ZJU-11, suggesting that these are the mutations most likely to be responsible for the higher viral load.

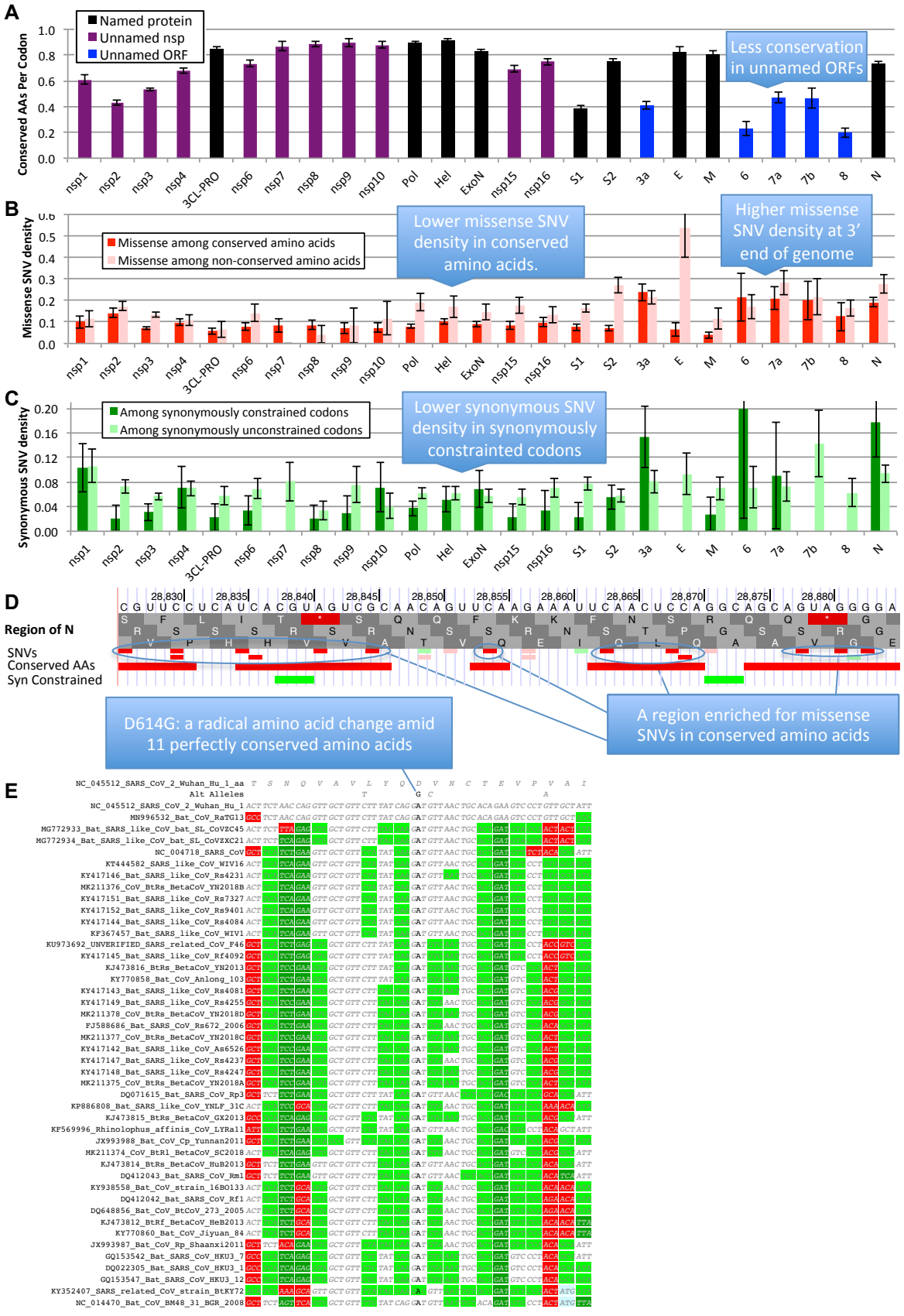


Figure 7. **Single nucleotide variants and conservation.** Error bars indicate standard error of mean. (A) Fraction of amino acids in each mature protein that are perfectly conserved in the Sarbecovirus alignment. We observe that in many products of the polyprotein,

but not all, the vast majority of amino acids are perfectly conserved; that the second part of the S protein is much more conserved than the first part; and that the named proteins are better conserved than the unnamed and hypothetical proteins. (B) Density of amino acid-changing single nucleotide variants (SNVs) among conserved (dark red) amino acid positions is significantly lower than in non-conserved (light red) positions. Both densities are higher near the 3' end of the genome, indicating higher mutation rate or less purifying selection even among amino acids that are perfectly conserved in Sarbecovirus. (C) Density of synonymous SNVs in synonymously-constrained codons (dark green) is significantly lower than among synonymously-unconstrained codons (light green). These results show that conservation in the Sarbecovirus clade at both the amino acid level and nucleotide level is associated with purifying selection on SNVs in the SARS-CoV-2 population. (D) Region in N enriched for missense SNVs in conserved amino acids. UCSC Genome Browser image of our SARS-CoV-2 conservation track hub, which provides information helpful in determining which SNVs are most likely to have a phenotypic effect. SNV conservation track indicates whether SNV is missense in a conserved amino acid (bright red), missense in a non-conserved amino acid (light red), synonymous in a synonymously-constrained codon (bright green), synonymous in a synonymously-unconstrained codon (light green), or noncoding (black, not shown). Other tracks indicate all constrained amino acids and synonymously-constrained codons. This 20 amino acid region of the nucleocapsid protein is significantly enriched for missense SNVs in amino acids that are perfectly conserved among our 44 Sarbecovirus genomes, suggesting positive selection or relaxed purifying selection in SARS-CoV-2. (E) Example of a mutation in a deeply conserved segment of the spike protein. Sarbecovirus alignment near an A to G nucleotide substitution at genomic location 23403 that gives rise to the amino acid change D614G in the spike protein, which has risen in frequency in multiple geographic locations, suggesting that it increases transmissibility. This mutation is near the middle of a string of 11 amino acids that are perfectly conserved among our Sarbecovirus genomes (all substitutions are light green, designating synonymous substitutions), which suggests that the mutation could be an adaptation to the human host.

Discussion

We used comparative genomic methods to determine which of the unnamed ORFs in SARS-CoV-2 and SARS-CoV-1 show the evolutionary signature of conserved, functional, protein-coding regions. We found that SARS-CoV-2 ORFs 3a, 6, 7a, 7b, and 8 have this signature, whereas ORFs 10, and 14 do not, and 9b is ambiguous. We also independently rediscovered a recently proposed novel dual coding region within ORF3a, ORF3c, using different methods, and provide strong evolutionary evidence for its coding potential. In SARS-CoV-1, ORFs 3a, 6, 7a, and 7b have this evolutionary signature, but 3b does not and 9b is again ambiguous, while 8a and 8b are too recent to determine their functional status from evolutionary signatures. We have also classified single nucleotide variants according to their evolutionary constraint, and used this approach to help interpret variants from two studies. These techniques should be applicable to other sets of variants as researchers try to untangle the connection between viral genotype and disease phenotype. Correct protein-coding annotations are essential not only for understanding viral biology, but also for predicting the phenotypic effect of variants, because determining how each variant affects protein sequence is the first step in any such analysis. As an example of the importance of correct annotations, we note that seven variants within ORF3c (T25473C, T25476C, G25494T, G25500A, G25500T, C25539T, C25572T) were classified by nextstrain as synonymous based on their predicted effect on ORF3a, but in fact cause amino acid changes in the ORF3c protein.

Our comparative genomics methods complement experimental approaches by providing a more comprehensive view of conserved function, with the caveat that in some cases, the evolutionarily-conserved function selected over the vast majority of the evolutionary interval studied may have recently changed, and thus evolutionary history may not reflect present state, which is better captured by experimental methods. However, experimental approaches only detect what is present under the specific conditions tested, whereas the comparative genomics approach used here can distinguish functional vs. non-functional regions based on their characteristic patterns of change, or evolutionary signatures, reflecting mutational perturbation experiments over millions of generations that survey conditions experienced by the virus in all hosts throughout the evolutionary history spanned by the genomes compared.

The stark differences between nucleotide-level (phyloP/phastCons) and protein-level (PhyloCSF) constraint in ORF8 and ORF10 highlight the importance of protein-coding evolutionary signatures vs. nucleotide-level constraint. While phyloP and phastConst rely on the *number* of substitutions, PhyloCSF instead relies on the *type* of substitutions, distinguishing those typical of coding vs. non-coding regions, regardless of the total number of substitutions.

Our analyses used a single genome-wide phylogenetic tree, but it is known that there is substantial recombination in Sarbecoviruses, leading to different evolutionary histories for different genomic segments, and segment boundaries have been identified within the S gene and the polyprotein (Wu et al. 2020; Sun et al. 2020; Andersen et al. 2020). PhyloCSF is relatively insensitive to the tree, and in fact an earlier version, CSF, did not make use of the tree. On the other hand, FRESCo relies more heavily on the tree, but it normalizes scores within a gene to the gene-wide average, which limits the effect of an incorrect tree provided that all of the gene has the same evolutionary history. We are not aware of any known recombination points within N or ORF3a, so our conclusions about overlapping reading frames in those genes are unlikely to be affected by this concern.

We identified a 20-amino acid region in the nucleocapsid protein that is significantly enriched for amino acid-changing variants in amino acids that have been conserved throughout the Sarbecovirus clade. Investigation of the effects of these variants on protein structure could yield insights into human adaptation.

Further experimental work will be needed to determine the functions of the unnamed genes and the effects of SARS-CoV-2 variants, which might lead to the identification of weaknesses of the virus. We hope that our conclusions and that the resources we have provided will help guide experimenters to the most fruitful investigations.

Methods

Genomes and Alignments

Genome sequences were obtained from https://www.ncbi.nlm.nih.gov/. The genomes and NCBI annotations for SARS-CoV-2 and SARS-CoV-1 were obtained from the records for accessions NC_045512.2 and NC_004718.3, respectively. The UniProt annotations for SARS-CoV-2 were obtained from the UCSC Genome Browser (Haeussler et al. 2019).

The 44 Sarbecovirus genomes used in this study were selected starting from all betacoronavirus and unclassified coronavirus full genomes listed on ncbi via searches https://www.ncbi.nlm.nih.gov/nuccore/?term=txid694002[Organism:exp] and the same with txid1986197 and txid2664420 on 5-Mar-2020, excluding any that differed from NC_045512.2 in more than 10,000 positions in a pairwise alignment computed using NW-align (Lab 2-Apr-2012), that cutoff being chosen so as to distinguish Sarbecovirus genomes among those that were classified, and removing near duplicates, including all SARS-CoV-1 and SARS-CoV-2 genomes other than the reference. Coronavirus genomes in the left half of Fig. 2 were those listed by https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=11118 on 11-Feb-2020.

The genomes were aligned using clustalo (Sievers and Higgins 2018) with the default parameters. The Phylogenetic tree was calculated using RAxML (Stamatakis 2014) using the GTRCATX model.

PhyloCSF, FRESCo, and other conservation metrics

PhyloCSF (Lin et al. 2011a) was run using the `29mammals` empirical codon matrices but with the Sarbecovirus tree substituted for the mammals tree. Input alignments were extracted from the whole-genome alignment and columns containing a gap in the reference sequence were removed. Browser tracks were created as described previously (Mudge et al. 2019). Scores listed in Supplemental Table S2 were calculated on the local alignment for each ORF or mature protein, excluding the final stop codon, using the default PhyloCSF parameters, including `--strategy=mle`, plus `--debug` in order to get the maximum-likelihood branch length scale factors for the coding and non-coding models. The mean and standard deviation of PhyloCSF-per-codon scores of protein-coding ORFs were calculated using the scores of the NCBI ORFs, excluding ORF1a because it is redundant with ORF1ab and excluding ORF10 because we had already determined it is not protein-coding; the mean and standard deviation for non-coding ORFs were calculated from those in Supplementary Table 4 in the initial subset, excluding ORFs 3h, 9b, and 14, since those were the ones under investigation.protein-coding ORFs were the NCBI ORFs excluding ORF

FRESCo (Sealfon et al. 2015) was run on 9-codon windows in each of the NCBI annotated ORFs. Alignments were extracted for the ORF excluding the final stop codon, and gaps in the reference sequence were removed.

SCEs were found by taking all windows having synonymous rate less than 1 and nominal p-value<10^{-5}, and combining overlapping or adjacent windows. For the variant analysis, FRESCo was also run on 1-codon windows using codon alignments as described previously (Sealfon et al. 2015).

Substitutions per site and per neutral site for each annotated ORF and mature protein were calculated by extracting the alignment column for each site or, respectively, 4-fold degenerate site, from the whole-genome alignment and determining the parsimonious number of substitutions using the whole-genome phylogenetic tree. For columns in which some genomes did not have an aligned nucleotide, the number of substitutions was scaled up by the branch length of the entire tree divided by the branch length of the tree of genomes having an aligned nucleotide in that column

PhastCons and phyloP tracks shown in Fig. 2 are the Comparative Genomics tracks from the UCSC Genome Browser, which were constructed from a multiz (Blanchette et al. 2004) alignment of the list of 44 Sarbecovirus genomes that we supplied to UCSC.

Analysis of Single Nucleotide Variants

Single nucleotide variants were downloaded from the "Nextstrain Vars" track in the UCSC Table Browser on 2020-04-18 at 11:46 AM EDT. We defined an amino acid to be "conserved" if there were no amino acid-changing substitutions in the Sarbecovirus alignment of its codon. We defined codons to be "synonymously constrained" if the p-value for the synonymous rate at that codon calculated by FRESCo using 1-codon windows was less than 0.034, which corresponds to a false discovery rate of 0.125.

To find regions that were significantly enriched for missense SNVs in conserved amino acids, we first defined a null model as follows. For each mature protein, we counted the number of missense SNVs and the number of conserved amino acids and randomly assigned each SNV to a conserved amino acid in the same mature protein, allowing multiplicity. For any positive integer n, we found the largest number of SNVs that had been assigned to any set of n consecutive conserved amino acids within the same mature protein across the whole genome. Doing this 100,000 times gave us a distribution of the number of missense SNVs in the most enriched set of n consecutive conserved amino acids in the genome. Comparing the number of actual missense SNVs in any particular set of n consecutive conserved amino acids to this distribution gave us a nominal p-value for that n. We applied this procedure for each n from 1 to 100 and multiplied the resulting p-values by a Bonferroni correction of 100 to calculate a corrected p-value for a particular region to be significantly enriched. We note that these 100 hypotheses are correlated because enriched regions of different lengths can overlap, so a Bonferroni correction is overly conservative and our reported p-value of 0.012 understates the level of statistical significance. To find significantly depleted regions we applied a similar procedure with every n from 1 to 1000, but did not find any depleted regions with nominal p-value less than 0.05 even without any multiple hypothesis correction.

Miscellaneous

Ribosome footprints shown in Supplemental Fig. S3 are from the track hub at ftp://ftp-igor.weizmann.ac.il/pub/hubSARSRibo.txt (Finkel et al. 2020).

Supplemental Materials:

- **Supplemental Figures S1-S11**
- **Supplemental Table S1.** Tab-separated table with one row for each of 44 Sarbecovirus strains used. Fields are the accession, name used in CodAlignView, and GenBank description.
- **Supplemental Table S2.** Information on all ORFs and mature proteins of SARS-CoV-2 as annotated by UniProt, including:
    - Genomic coordinates in bed-like format (0-based half-open)
    - Names and alternative names for the ORF and protein
    - Number of codons
    - PhyloCSF score per codon
    - Average number of substitutions per site using the whole-genome tree
    - Average number of substitutions per 4-fold degenerate site using the whole-genome tree
    - Average number of substitutions per site using the whole-genome 42-strain tree that excludes the two most distant

strains
    - Average number of substitutions per 4-fold degenerate site using the whole-genome 42-strain tree that excludes the two most distant strains
    - The number and fraction of amino acids that are conserved, codons that are synonymously constrained, and various categories of single nucleotide variants.
    - Excel hyperlink to CodAlignView showing the ORF or mature protein and 5 codons of the neighbor on each side
    - Excel hyperlink to UCSC Genome Browser showing ORF or mature protein and 5 codons of the neighbor on each side
    - UniProt comments on function of the ORF or mature protein
- **Supplemental Table S3.** Information about each of the single nucleotide variants used in this study including position information; number of nextstrain genomes containing the variant and nextstrain's classification ("INFO" field); reference and alternate nucleotide and amino acid; our classification as noncoding, synonymous, or non-synonymous; nonsynonymous rate for nonsynonymous SNVs; synonymous rate, p-value, FDR, localFDR, and containment in SCE for synonymous SNVs; and links to view the SNV in the UCSC Genome Browser and its alignment in CodAlignView with 25 codons of context on each side. For the SNVs associated with the spike-protein variants in Korber et al. we also include the corresponding amino acid variant, our classification, and the other information from Korber et al. Table 1. For the variants from Yao et al., we include the variant and our classification.
- **Supplemental Table S4.** Spreadsheet in tab-separated format listing open reading frames searched for novel coding regions. These consisted of all SARS-CoV-2 ORFs at least 10 codons long, on either strand, beginning with AUG or a near-cognate codon, that do not overlap an NCBI-annotated gene in the same reading frame or the antisense frame (the frame on the opposite strand that shares the 3rd codon position; antisense regions gets artifactually high PhyloCSF scores). Our initial subset consisted of those on the '+' strand, with a canonical (AUG) start codon, at least 25 codons long, that are maximal (i.e, not contained in a longer AUG-initiated ORF in the same frame). ORFs in our initial subset are listed first, in order of decreasing PhyloCSF score per codon, followed by all other ORFs, also in order of decreasing PhyloCSF score per codon. Spreadsheet fields include general information about the ORF, links to view the alignment in CodAlignView with 10 codons on each side for context, links to view the region in the UCSC Genome browser, PhyloCSF score per codon, branch length of strains present in the local alignment as a fraction of total branch length (RelBL), PhyloCSF's branch length scale factors for its coding and noncoding models (RhoC and RhoN), adjusted score consisting of PhyloCSF score per codon divided by the average of RhoC and RhoN, relative branch length of strains conserving the start codon/ stop codon/reading frame, GC content, fraction of the ORF that overlaps Synonymous Constraint Elements, and whether the ORF was reported as translated in the Finkel et al. ribosome profiling experiments.
- Pdf files containing the alignment of each UniProt-annotated SARS-CoV-2 ORF and mature protein with 5 codons of the neighbor on each side, color-coded by CodAlignView for protein-coding evolutionary features.
- Whole-genome alignment of 44-Sarbecovirus genomes in Fasta format.
- Whole-genome phylogenetic tree in Newick format.
- Nextstrain_ncov_global_metadata.tsv: List of authors who contributed genomes to GISAID that were used by nextstrain and UCSC to produce the list of SNVs.

Data Access

The PhyloCSF tracks and FRESCo synonymous constraint elements are available for the SARS-CoV-2/wuhCor1 assembly in the UCSC Genome Browser (Haeussler et al. 2019) using the "PhyloCSF" and "Synonymous Constraint" public track hubs. The alignments and phylogenetic tree are included as supplemental materials. The alignments may be viewed, color coded to indicate protein-coding signatures, using CodAlignView (https://data.broadinstitute.org/compbio1/cav.php) with alignment set wuhCor1_c and chromosome name NC_045512v2.

SARS-CoV-2 single nucleotide variants, color coded by whether they are non-coding, synonymous, or amino acid-changing, and whether they are in conserved codons, as well as indications of which codons are conserved at the amino acid or synonymous level, may be viewed in the UCSC Genome Browser using the track hub at https://data.broadinstitute.org/compbio1/SARS-CoV-2conservation/trackHub/hub.txt. The details page for each SNV includes information about Sarbecovirus conservation and a link to view the alignment of a neighborhood of the SNV in CodAlignView. It is our intention to update this track hub as the list of variants in the UCSC Table Browser is updated.

In this resource, we have augmented variant data made available by UCSC with our own annotations. UCSC data came from nextstrain.org (Hadfield et al. 2018), which was derived from genome sequences deposited in GISAID (Elbe and Buckland-Merrett 2017). Right of use and publication of the underlying sequences is entirely controlled by the authors of the original resource and the contributors of individual sequences, who are acknowledged in the nextstrain metadata file included with supplemental materials. Our analysis provides an additional layer of annotation on their work rather than replicating or replacing it.

Original data usage policy as provided by UCSC:

The data presented here is intended to rapidly disseminate analysis of important pathogens. Unpublished data is included with permission of the data generators, and does not impact their right to publish. Please contact the respective authors (available via the Nextstrain metadata.tsv file) if you intend to carry out further research using their data. Derived data, such as phylogenies, can be downloaded from nextstrain.org (see "DOWNLOAD DATA" link at bottom of page) - please contact the relevant authors where appropriate.

Acknowledgements

We thank all those who have contributed sequences to the GISAID database and those at nextstrain.org and ucsc.edu who have made the variant data available. We would like to thank Jeremy Luban, Robert Garry, and Mark Diekhans for helpful input. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U41HG007234. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Additional support was provided by the Wellcome Trust grant number WT108749/Z/15/Z, NIH grant R01 HG004037.

References

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nat Med*. https://doi.org/10.1038/s41591-020-0820-9.

Baranov PV, Henderson CM, Anderson CB, Gesteland RF, Atkins JF, Howard MT. 2005. Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology* **332**: 498–510.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708–715.

Bojkova D, Klann K, Koch B, Widera M, Krause D, Ciesek S, Cinatl J, Münch C. 2020. SARS-CoV-2 infected host cell proteomics reveal potential therapy targets. *Preprint available at Research Square*.

Cagliani R, Forni D, Clerici M, Sironi M. 2020. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. *Infection, Genetics and Evolution* **83**: 104353. http://dx.doi.org/10.1016/j.meegid.2020.104353.

Davidson AD, Williamson MK, Lewis S, Shoemark D. 2020. Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage …. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.03.22.002204v1.abstract.

DeRisi JL, Huber G, Kistler A, Retallack H, Wilkinson M, Yllanes D. 2019. An exploration of ambigrammatic sequences in narnaviruses. *Sci Rep* **9**: 17982.

Dinan AM, Lukhovitskaya NI, Olendraite I, Firth AE. 2020. A case for a negative-strand coding sequence in a group of positive-sense RNA viruses. *Virus Evol* **6**: veaa007.

Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**: 33–46.

Finkel Y, Mizrahi O, Nachshon A. 2020. The coding capacity of SARS-CoV-2. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.05.07.082909v1.abstract.

Firth AE. 2020. A putative new SARS-CoV protein, 3a*, encoded in an ORF overlapping ORF3a. *bioRxiv* 2020.05.12.088088. https://www.biorxiv.org/content/10.1101/2020.05.12.088088v1.abstract (Accessed May 28, 2020).

Firth AE. 2014. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res* **42**: 12425–12439.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**: 4121–4123.

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. 2019. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**: D853–D858.

I Jungreis, MF Lin, CS Chan, M Kellis. 2016. CodAlignView. *CodAlignView: The Codon Alignment Viewer*. https://data.broadinstitute.org/compbio1/cav.php (Accessed April 30, 2016).

Jungreis I, Chan CS, Waterhouse RM, Fields G, Lin MF, Kellis M. 2016. Evolutionary Dynamics of Abundant Stop Codon Readthrough. *Mol Biol Evol* **33**: 3108–3132.

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res* **21**: 2096–2113.

Khan YA, Jungreis I, Wright JC, Mudge JM, Choudhary JS, Firth AE, Kellis M. 2020. Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. *BMC Genet* **21**: 25.

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.e10.

Korber B, Fischer W, Gnanakaran SG, Yoon H. 2020. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.04.29.069054v1.abstract.

Lab Z. 2-Apr-2012. NW-align. *NW-align*. http://zhanglab.ccmb.med.umich.edu/NW-align (Accessed 23-May-2015).

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–482.

Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St Pierre SE, et al. 2007. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res* **17**: 1823–1836.

Lin MF, Jungreis I, Kellis M. 2011a. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275–82.

Lin MF, Kheradpour P, Washietl S, Parker BJ, Pedersen JS, Kellis M. 2011b. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Research* **21**: 1916–1928. http://dx.doi.org/10.1101/gr.108753.110.

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. *Science Advances* eabb9153.

Loughran G, Chou M-Y, Ivanov IP, Jungreis I, Kellis M, Kiran AM, Baranov PV, Atkins JF. 2014. Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res* **42**: 8928–8938.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**: 565–574.

McCorkindale AL, Wahle P, Werner S, Jungreis I, Menzel P, Shukla CJ, Abreu RLP, Irizarry RA, Meyer IM, Kellis M, et al. 2019. A gene expression atlas of embryonic neurogenesis in Drosophila reveals complex spatiotemporal regulation of lncRNAs. *Development* **146**. http://dx.doi.org/10.1242/dev.175265.

Miller WA, Koev G. 2000. Synthesis of subgenomic RNAs by positive-strand RNA viruses. *Virology* **273**: 1–8.

Mudge JM, Jungreis I, Hunt T, Gonzalez JM, Wright JC, Kay M, Davidson C, Fitzgerald S, Seal R, Tweedie S, et al. 2019. Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res* **29**: 2073–2087.

Muth D, Corman VM, Roth H, Binger T, Dijkman R, Gottula LT, Gloza-Rausch F, Balboni A, Battilani M, Rihtarič D, et al. 2018. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Sci Rep* **8**: 15177.

Oxford JS, Collier LH, Kellam P. 2016. *Human Virology*. Oxford University Press.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**: 110–121.

Schaecher SR, Mackenzie JM, Pekosz A. 2007. The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles. *J Virol* **81**: 718–731.

Sealfon RS, Lin MF, Jungreis I, Wolf MY, Kellis M, Sabeti PC. 2015. FRESCo: finding regions of excess synonymous constraint in diverse viruses. *Genome Biol* **16**: 38.

Shi C-S, Qi H-Y, Boularan C, Huang N-N, Abu-Asab M, Shelhamer JH, Kehrl JH. 2014. SARS-coronavirus open reading frame-9b suppresses innate immunity by targeting mitochondria and the MAVS/TRAF3/TRAF6 signalosome. *J Immunol* **193**: 3080–3089.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **27**: 135–145.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Sun J, He W-T, Wang L, Lai A, Ji X, Zhai X, Li G, Suchard MA, Tian J, Zhou J, et al. 2020. COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. *Trends Mol Med* **0**. https://www.cell.com/trends/molecular-medicine/fulltext/S1471-4914(20)30065-4 (Accessed April 3, 2020).

Taiaroa G, Rawlinson D, Featherstone L, Pitt M, Caly L, Druce J, Purcell D, Harty L, Tran T, Roberts J, et al. Direct RNA sequencing and early evolution of SARS-CoV-2. http://dx.doi.org/10.1101/2020.03.05.976167.

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, et al. 2020. A new

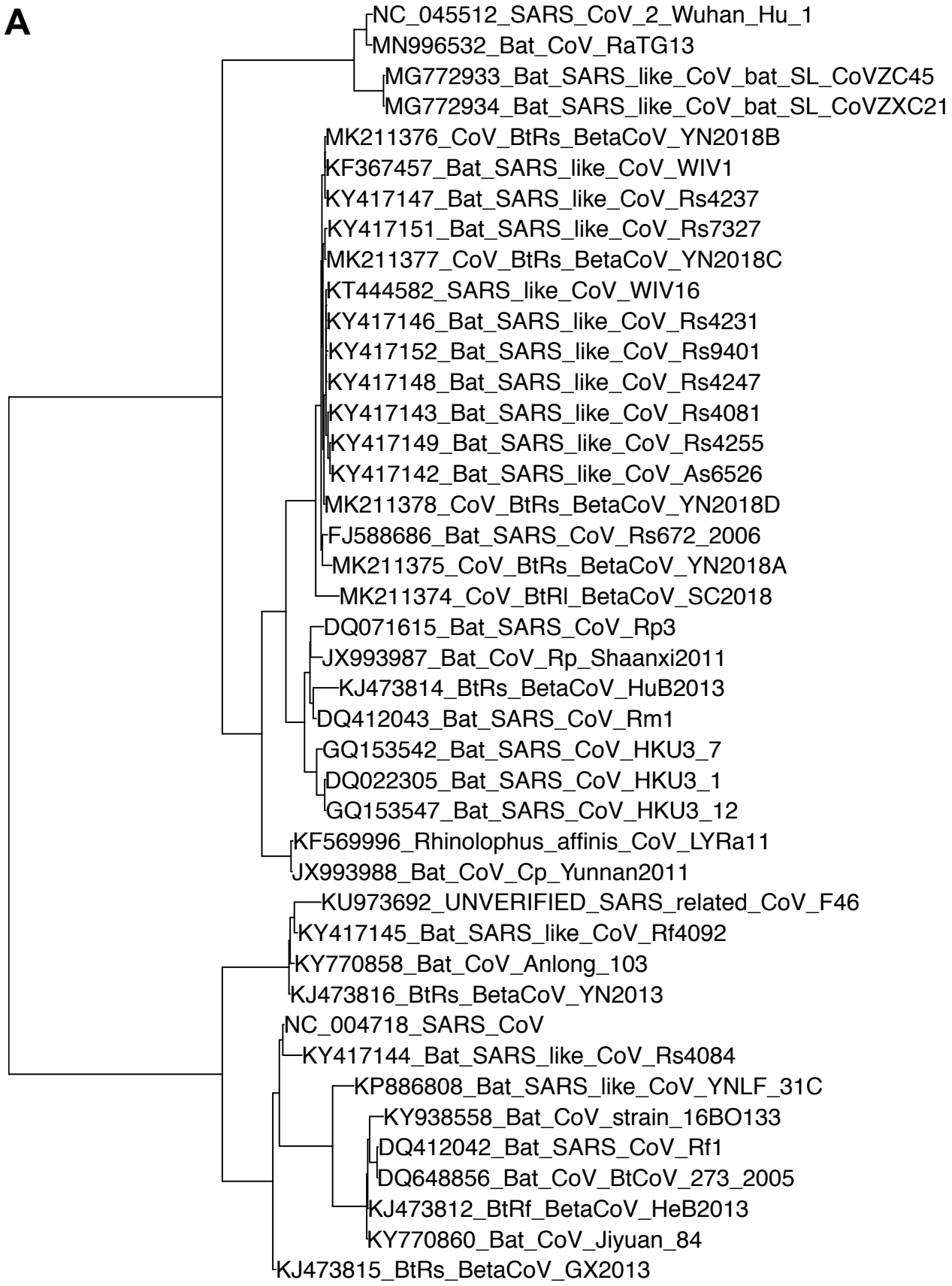
coronavirus associated with human respiratory disease in China. *Nature* **579**: 265–269.

Yao H-P, Lu X, Chen Q, Xu K, Chen Y, Cheng L, Liu F, Wu Z, Wu H, Jin C, et al. 2020. Patient-Derived Mutations Impact Pathogenicity of SARS-CoV-2. https://papers.ssrn.com/abstract=3578153 (Accessed May 9, 2020).

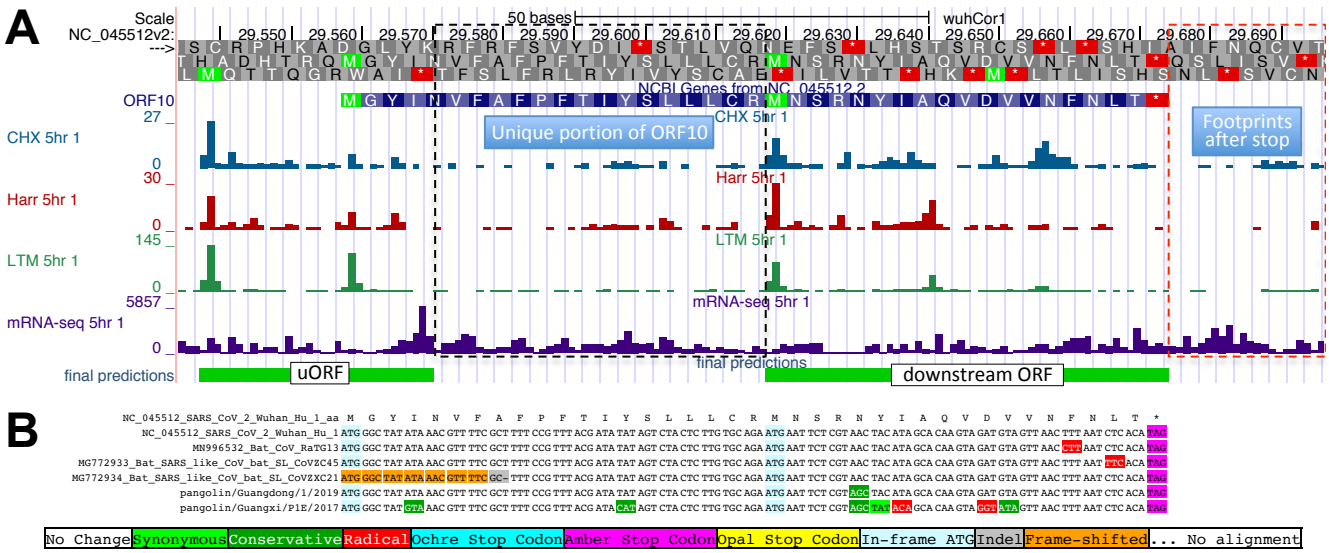
Supplemental Figures

Supplemental Figure S1. **Alignment of nsp11 and frameshift site.** Alignment of nonstructural protein nsp11 and the subsequent 5

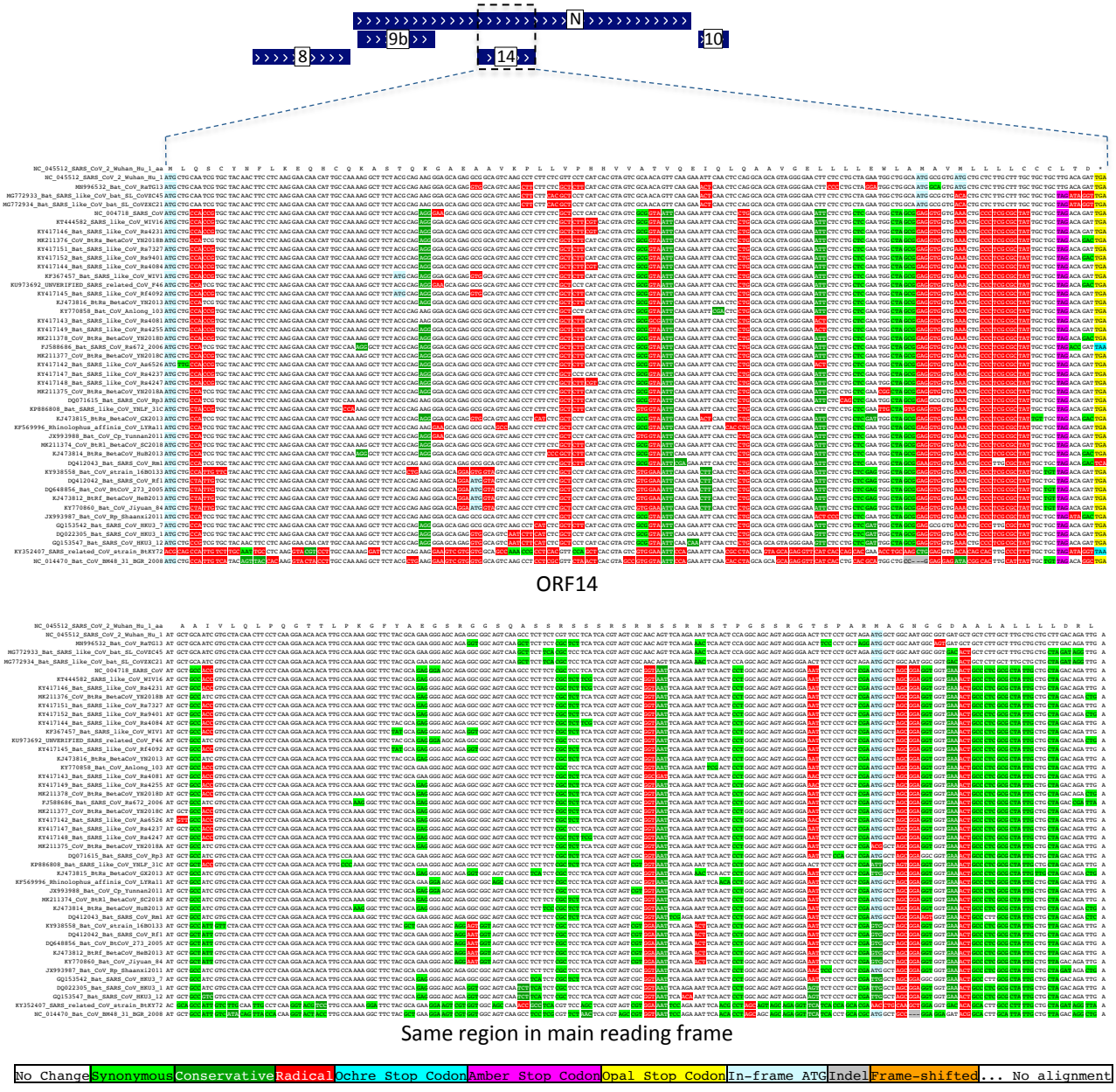
codons in 44 Sarbecoviruses (top) and 52 coronaviridae (bottom). Sarbecoviruses included in the coronaviridae alignment are indicated by green dots. The slippery site of the programmed frameshift (red rectangle) is perfectly conserved in all genomes. The polymerase, Pol, shares the 5' nine codons of nsp11 but then continues 3' of the slippery site in a different reading frame. The four codons 3' of the slippery site are perfectly conserved in Sarbecoviruses, which is consistent with a dual coding region. However, the stop codon of the un-framshifted polyprotein, pp1a, which marks the 3' end of nsp11, is poorly conserved in coronaviridae (cyan, magenta, and yellow stop codons). This, and the fact that nsap11 is only 13 codons long, suggest that it does not produce a functional protein.


A
NC_045512_SARS_CoV_2_Wuhan_Hu_1
MN996532_Bat_CoV_RaTG13
MG772933_Bat_SARS_like_CoV_bat_SL_CoVZC45
MG772934_Bat_SARS_like_CoV_bat_SL_CoVZXC21
MK211376_CoV_BtRs_BetaCoV_YN2018B
KF367457_Bat_SARS_like_CoV_WIV1
KY417147_Bat_SARS_like_CoV_Rs4237
KY417151_Bat_SARS_like_CoV_Rs7327
MK211377_CoV_BtRs_BetaCoV_YN2018C
KT444582_SARS_like_CoV_WIV16
KY417146_Bat_SARS_like_CoV_Rs4231
KY417152_Bat_SARS_like_CoV_Rs9401
KY417148_Bat_SARS_like_CoV_Rs4247
KY417143_Bat_SARS_like_CoV_Rs4081
KY417149_Bat_SARS_like_CoV_Rs4255
KY417142_Bat_SARS_like_CoV_As6526
MK211378_CoV_BtRs_BetaCoV_YN2018D
FJ588686_Bat_SARS_CoV_Rs672_2006
MK211375_CoV_BtRs_BetaCoV_YN2018A
MK211374_CoV_BtRl_BetaCoV_SC2018
DQ071615_Bat_SARS_CoV_Rp3
JX993987_Bat_CoV_Rp_Shaanxi2011
KJ473814_BtRs_BetaCoV_HuB2013
DQ412043_Bat_SARS_CoV_Rm1
GQ153542_Bat_SARS_CoV_HKU3_7
DQ022305_Bat_SARS_CoV_HKU3_1
GQ153547_Bat_SARS_CoV_HKU3_12
KF569996_Rhinolophus_affinis_CoV_LYRa11
JX993988_Bat_CoV_Cp_Yunnan2011
KU973692_UNVERIFIED_SARS_related_CoV_F46
KY417145_Bat_SARS_like_CoV_Rf4092
KY770858_Bat_CoV_Anlong_103
KJ473816_BtRs_BetaCoV_YN2013
NC_004718_SARS_CoV
KY417144_Bat_SARS_like_CoV_Rs4084
KP886808_Bat_SARS_like_CoV_YNLF_31C
KY938558_Bat_CoV_strain_16BO133
DQ412042_Bat_SARS_CoV_Rf1
DQ648856_Bat_CoV_BtCoV_273_2005
KJ473812_BtRf_BetaCoV_HeB2013
KY770860_Bat_CoV_Jiyuan_84
KJ473815_BtRs_BetaCoV_GX2013

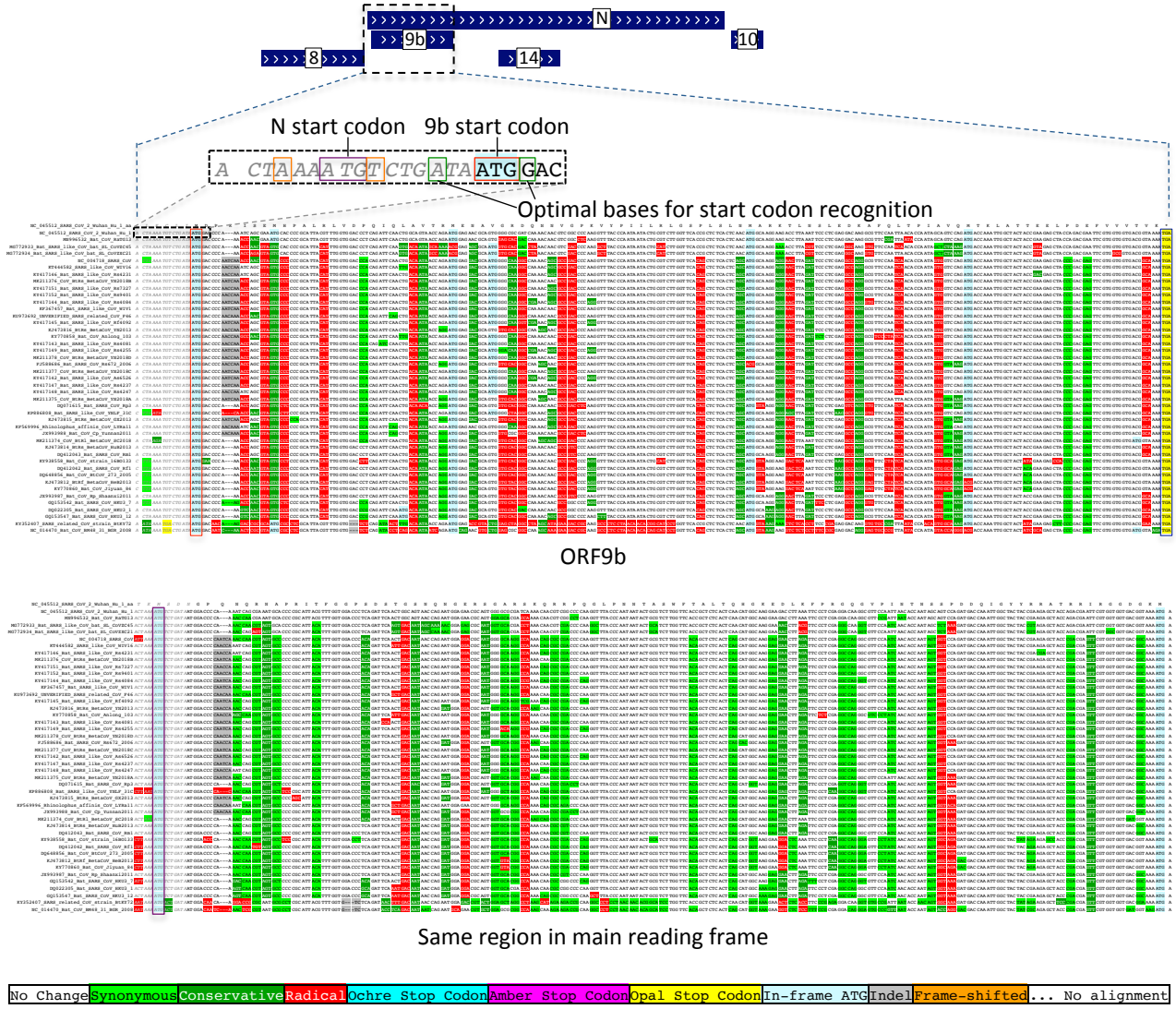
Supplemental Figure S2. ORF8 Phylogeny. (A) Phylogenetic tree computed from the alignment of ORF8, which is quite different from the genome-wide tree in Fig. 2, indicating that there has been a recombination event near this region. (B) Alignment of ORF8 with strains ordered according to the ORF8 tree shows a large number of substitutions, demonstrating that the apparently high evolutionary rate in ORF 8 is not an artifact of recombination.



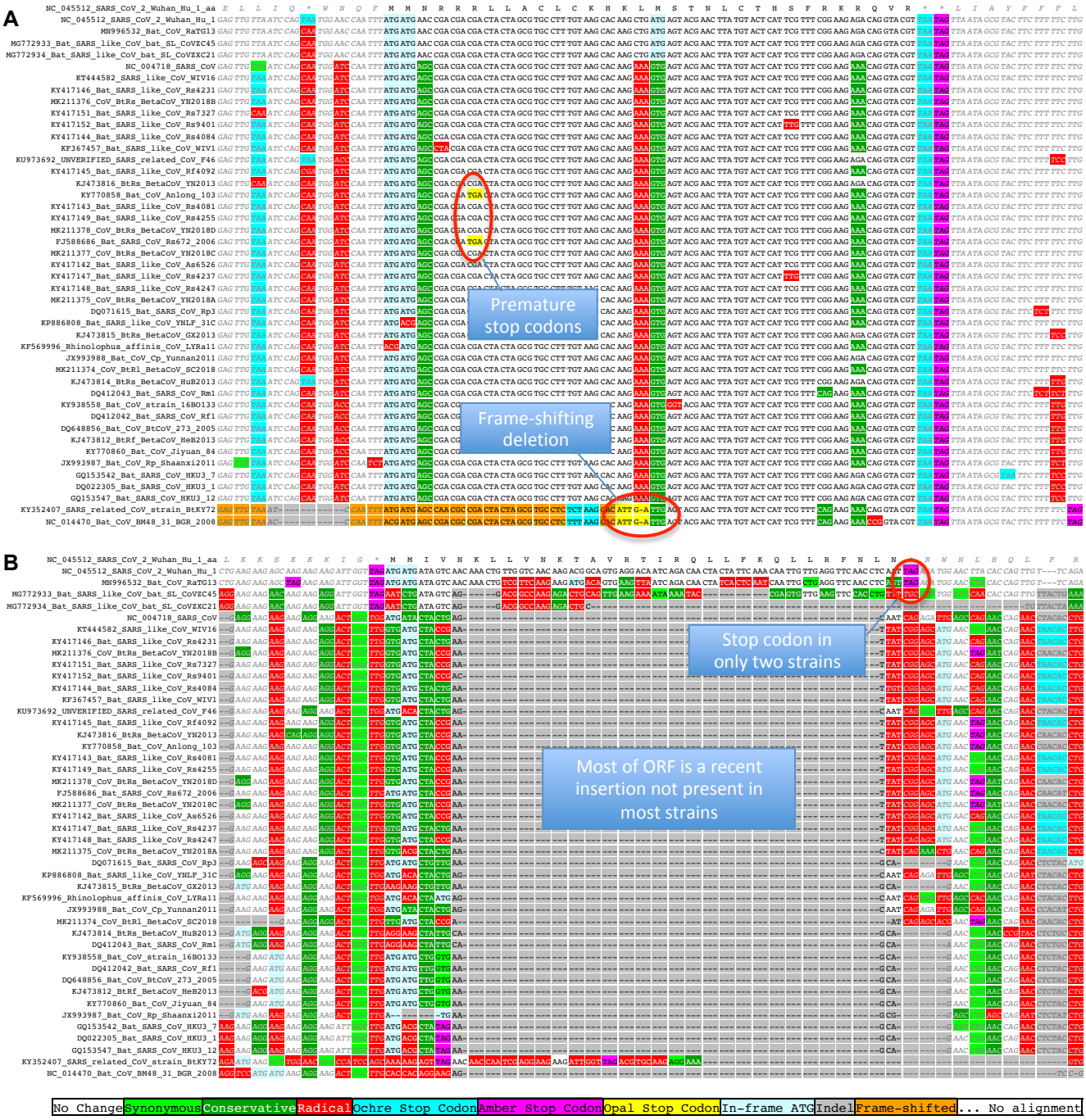
Supplemental Figure S3. Reported evidence that ORF10 is coding. (A) UCSC Genome Browser image of ORF10 and ribosome footprints 5 hours post-infection (Finkel et al. 2020). CHX track shows footprints of actively translating ribosomes from cells treated with cycloheximide. Harr and LTM tracks show footprints from cells treated with harringtonine and lactimidomycin, respectively, to enrich for initiating ribosomes. Final predictions track shows ORFs computationally predicted by Finkel et al. from footprint data. Nearly all footprints within ORF10 are in either the predicted uORF overlapping the ORF10 start codon, or the predicted downstream ORF beginning at an interior AUG, which would create a peptide of only 18 amino acids (only 5 amino acids in all but the four closest strains, since the others have an early stop codon), and the density of footprints in the unique portion of ORF10 (dashed black rectangle) appears to be no greater than the density beyond its stop codon (dashed red rectangle), suggesting they are not due to translation of ORF10. (B) Alignment of ORF10 in SARS-CoV-2, three bat viruses, and two pangolin viruses, used by Cagliani et al. to infer that the ORF is protein-coding and under positive selection due to a high dN/dS ratio (Cagliani et al. 2020). However, the alignment includes only nine substitutions, one of them synonymous. In the hypothetical translation of a non-coding region evolving neutrally, we would expect between two and three of nine substitutions to be synonymous, and the depletion to only one is not statistically significant ($p > 0.18$ even without the necessary multiple hypothesis correction). Furthermore, one sequence was excluded due to a 1-base frameshifting deletion (orange and grey), which, if it is not a sequencing error, would be strong evidence against conserved protein-coding function.



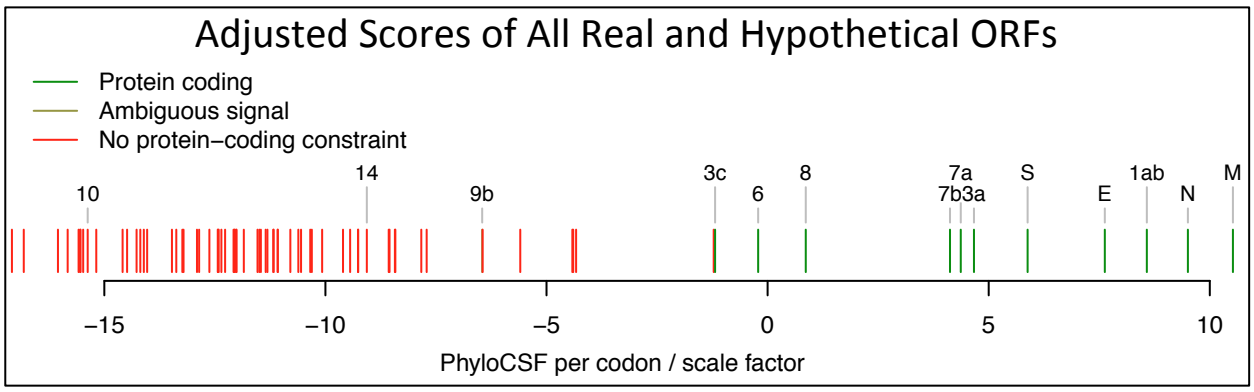
Supplemental Figure S4. Alignment of ORF14. Sarbecovirus alignment of ORF14 (top), which overlaps the nucleocapsid protein N in an alternate frame. The start codon is lost in KY352407_SARS_related_CoV_strain_BtKY72, and most strains have a premature UAG stop codon (magenta) 3 codons before the end. Nearly all substitutions are radical amino acid changes (red). It is unlikely that this ORF encodes a conserved functional protein. Also shown is the same region in the frame of the overlapping nucleocapsid protein (bottom), in which most substitutions are synonymous (light green).



Supplemental Figure S5. Alignment of ORF9b. Sarbecovirus alignment of ORF9b (top), which overlaps the nucleocapsid protein N in an alternate reading frame. Although most substitutions are non-synonymous (red and dark green), the start codon (red box) and stop codon (blue box) are perfectly conserved, and there are no intermediate stop codons in other strains. Bases A and G in positions -3 and +4 (green boxes), respectively, are optimal for ribosomal start codon recognition. The start codon of N (purple box) is 10 nt 5' of start codon of 9b, with less-optimal bases A and T in positions -3 and +4 of this start codon (orange boxes), suggesting that ORF9b could be translated from the same subgenomic RNA as N by leaky scanning. Also shown is the same region in the frame of the overlapping nucleocapsid protein (bottom), in which most substitutions are synonymous (light green). Evolutionary evidence is ambiguous regarding whether 9b is coding.

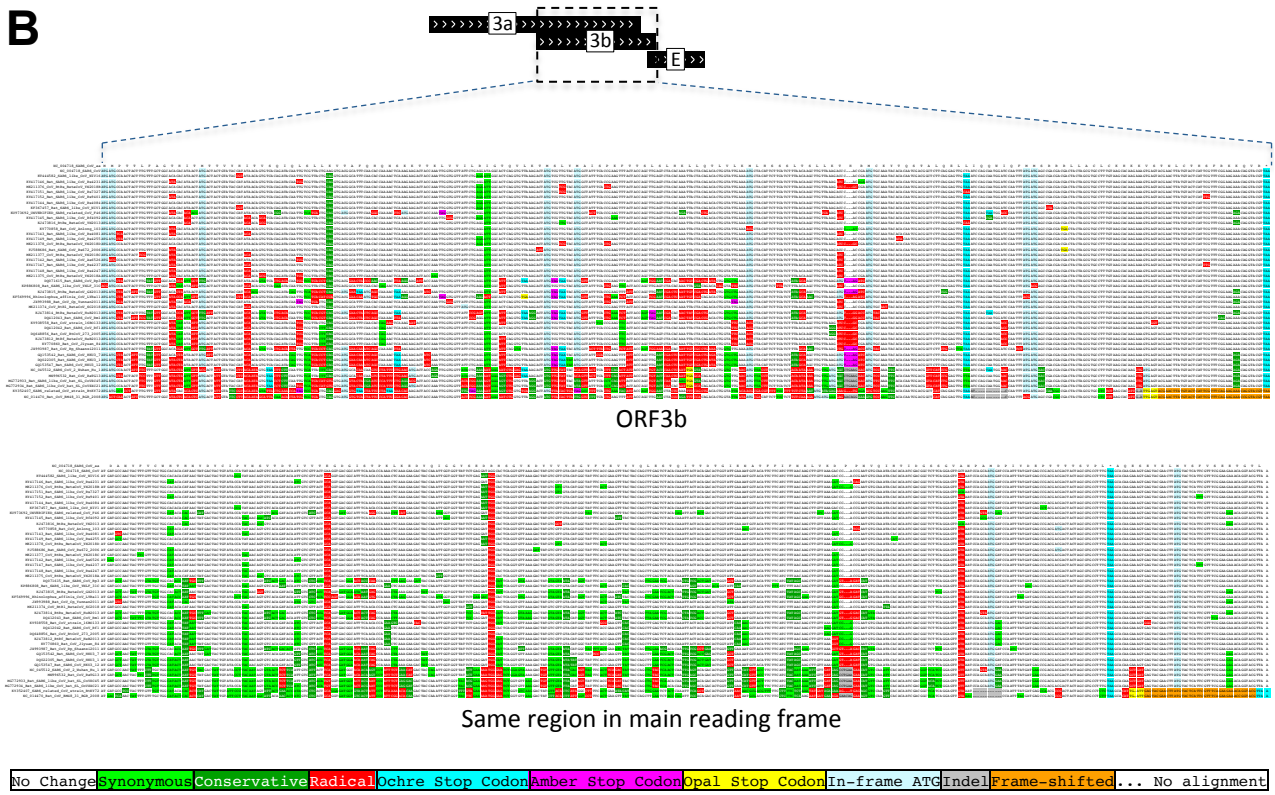


Supplemental Fig S6. Alignments of rejected ORFs. CodAlignView images of ORFs rejected during our search for novel conserved coding regions. (A) 32-codon ORF (26183-26278) that overlaps the 3' end of ORF3a and the 5' end of E with PhyloCSF score -2.74. Two strains have a frame-shifting one-base deletion within the ORF, and two others have premature stop codons. None of the substitutions are synonymous. There is high nucleotide-level constraint, but it continues on both sides of the ORF, suggesting it does not result from translation of the ORF. (B) 31-codon ORF (3207-3299) overlapping ORF1a, with PhyloCSF score -7.77. Most of the ORF consists of a 75-nt insertion that is only present in SARS-CoV-2, RaTG13, and CoVZC45, and the start and stop codons are missing in CoVZC45, so this is not a conserved coding sequence.

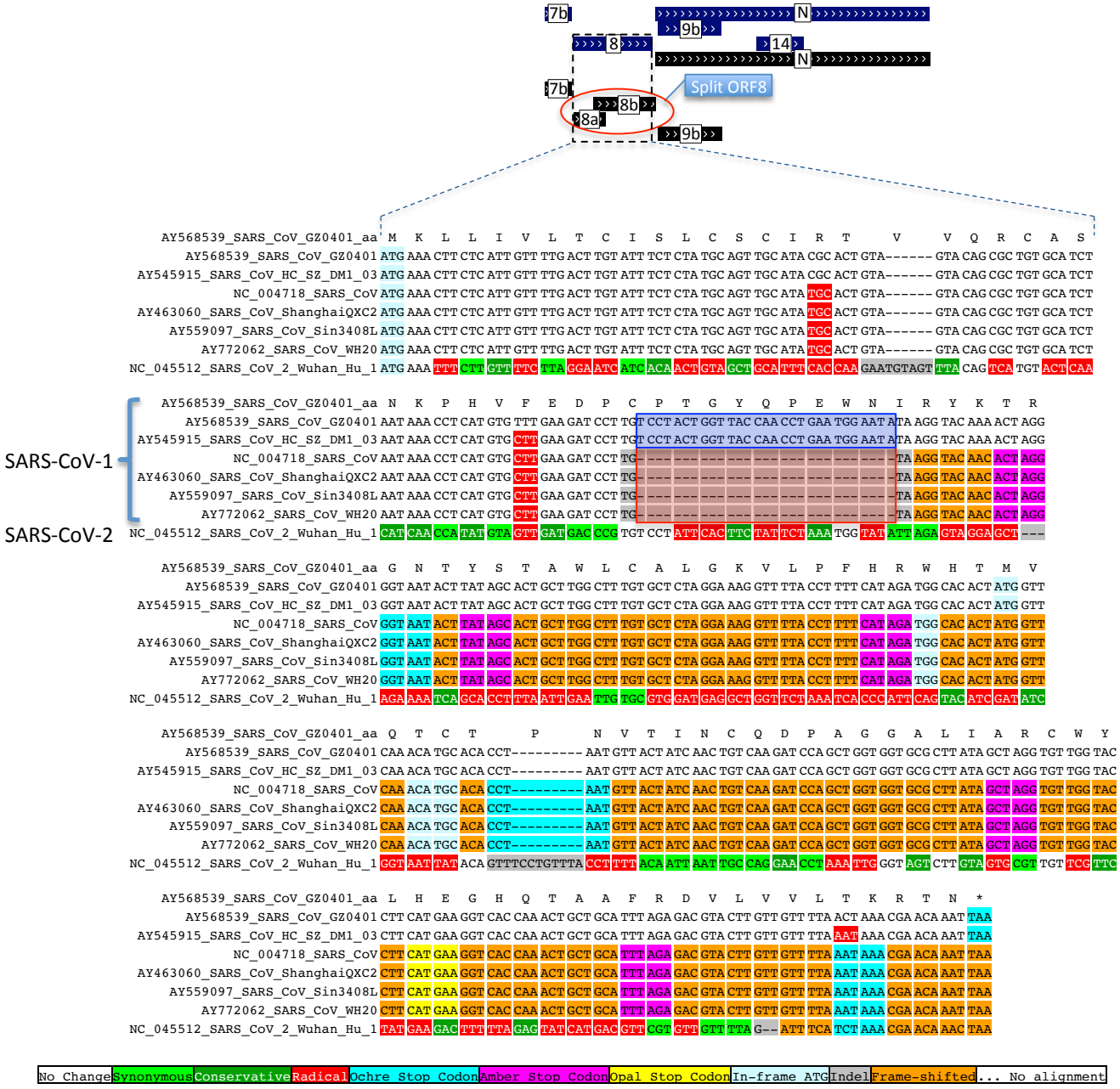


Supplemental Figure S7. Adjusted scores of all ORFs. PhyloCSF score per codon divided by the maximum-likelihood branch length scale factor computed by PhyloCSF for its coding and non-coding models, for all the ORFs in Fig. 1C, namely all annotated and hypothetical AUG-initiated ORFs on the positive strand at least 25 codons long that do not overlap a longer ORF in the same frame. Dividing by this scale factor adjusts for the fact that in regions with a low frequency of substitutions, such as throughout ORFs N and 10, PhyloCSF has less statistical power to distinguish its coding and noncoding evolutionary models, which compresses the PhyloCSF score towards 0, resulting in a better rank among the negative scores. The lower scores of ORFs 10, 14, and 9b with this adjustment show that their relatively high negative scores in Fig. 1C are at least in part an artifact of the low frequency of substitutions in these genomic regions.

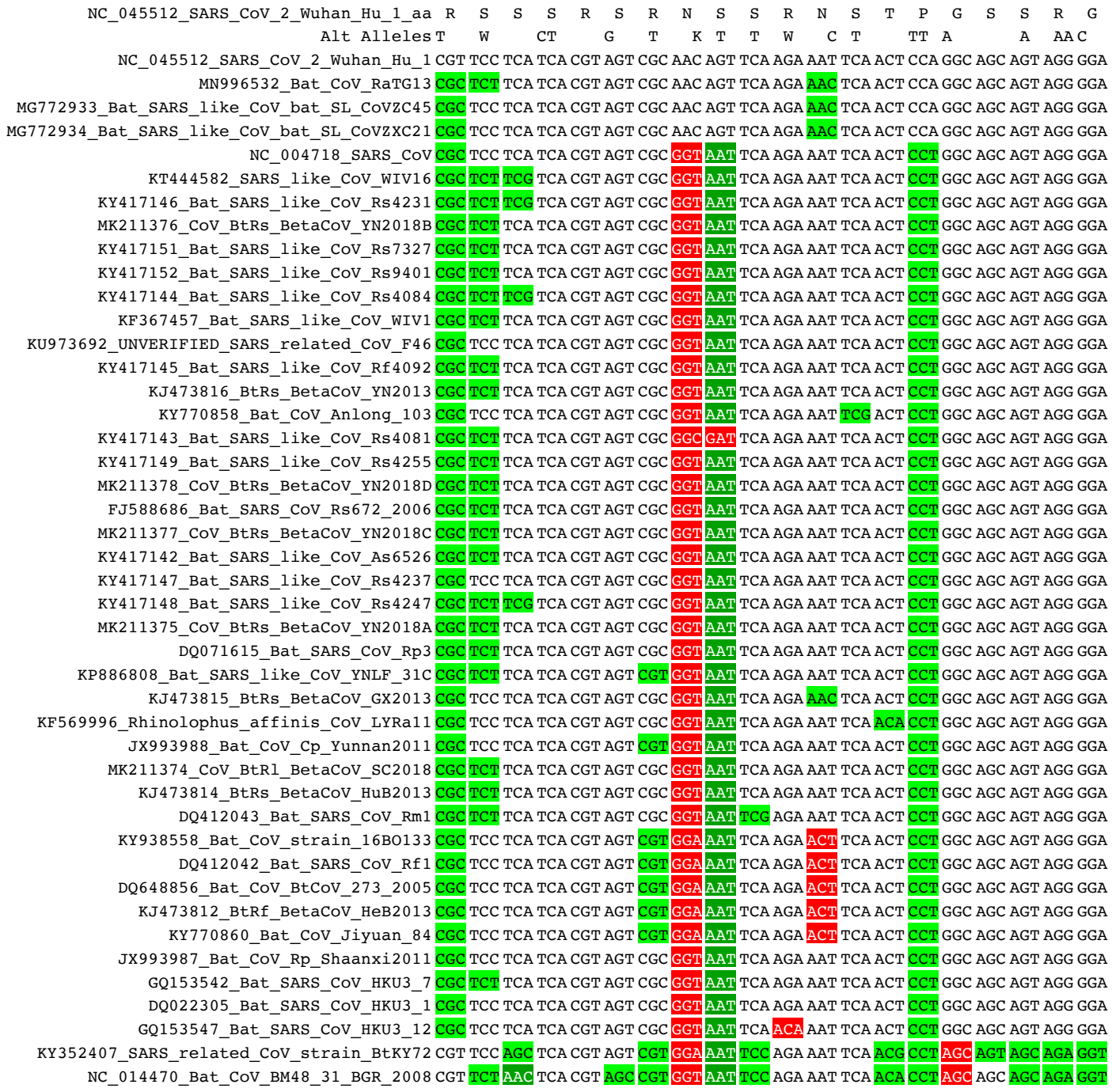

A
3a
3b
E



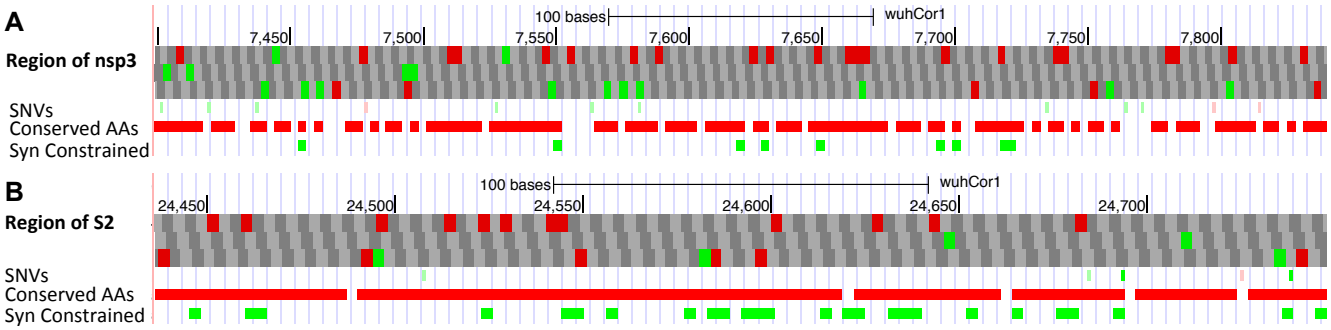
Supplemental Figure S8. SARS-CoV-1 ORF3b. (A) Alignment of SARS-CoV-1 ORF3b in 44 Sarbecoviruses, rearranged so that SARS-CoV-1 and its closest relatives come first. Most of ORF3b overlaps ORF3a in a different frame. There are numerous stop codons in other strains (red ovals), including a stop codon about ¾ of the way through the ORF in most strains including those closest to SARS-CoV-1 (red rectangle), so this cannot be a conserved coding region. (B) ORF3b shown in its reading frame (top) compared to the same region in the reading frame of ORF3a (bottom).



Supplemental Figure S9. ORF8 and 8a. Alignment of ORF8 containing SARS-CoV-2 and six isolates of SARS-CoV-1. A 29-nt deletion is present in four of the SARS-CoV-1 isolates (red box), causing a frameshift leading to an early stop codon, terminating annotated ORF8a in SARS-CoV-1. This deletion is not present in the other two SARS-CoV-1 isolates shown (blue box), or in SARS-CoV-2, which serves as an outgroup, indicating that it occurred within the SARS-CoV-1 strain, presumably during the 2003 SARS outbreak.



Supplemental Figure S10. SNV cluster in N. Alignment of the 20-amino acid region in the nucleocapsid protein that is highly enriched for missense SNVs in perfectly conserved amino acid residues. Alternate alleles are shown in the second row using the standard code for degenerate nucleotides (W indicates the two alternate alleles A or T, and K indicates G or T). There 14 non-synonymous SNVs among the 14 amino acids that are perfectly conserved among the 44 Sarbecovirus genomes (columns with no red or dark green), suggesting positive or relaxed purifying selection.



Supplemental Figure S11. SNV-depleted regions. UCSC Genome Browser images of regions in nsp3 (A) and S2 (B). Most of the amino acids in these regions are conserved (red rectangles in Conserved AAs track), but the only missense SNVs in these regions (light red rectangles in SNVs track) are in non-conserved amino acids (missense SNVS in conserved amino acids would be bright red if present). The lack of missense SNVs in such a large set of conserved amino acid residues could indicate that constraint in the Sarbecovirus clade has continued particularly strongly in the SARS-CoV-2 population. However, although these are the most depleted regions in the genome for missense SNVs in conserved amino acid residues, neither depletion is statistically significant ($p = 0.072$ and $p = 0.093$, respectively, without any correction for multiple region lengths searched).