A previously uncharacterized gene in SARS-CoV-2 illuminates the functional dynamics and evolutionary origins of the COVID-19 pandemic

Chase W. Nelson[1,2,a,*], Zachary Ardern[3,a,*], Tony L. Goldberg[4,5], Chen Meng[6], Chen-Hao Kuo[1], Christina Ludwig[6], Sergios-Orestis Kolokotronis[2,7,8], Xinzhu Wei[9,*]

[1]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan
[2]Institute for Comparative Genomics, American Museum of Natural History, New York, NY, USA
[3]Chair for Microbial Ecology, Technical University of Munich, Freising, Germany
[4]Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, WI, USA
[5]Global Health Institute, University of Wisconsin-Madison, Madison, WI, USA
[6]Bavarian Center for Biomolecular Mass Spectrometry (BayBioMS), Technical University of Munich, Freising, Germany
[7]Department of Epidemiology and Biostatistics, School of Public Health, SUNY Downstate Health Sciences University, Brooklyn, NY, USA
[8]Institute for Genomic Health, SUNY Downstate Health Sciences University, Brooklyn, NY, USA
[9]Departments of Integrative Biology and Statistics, University of California, Berkeley, CA, USA

[a]These authors contributed equally.
*Corresponding authors: cnelson@gate.sinica.edu.tw, zachary.ardern@tum.de, aprilwei@berkeley.edu

Abstract

Understanding and preventing the emergence of novel viruses requires an accurate and comprehensive understanding of their genomes. One under-investigated class of functional genomic elements is overlapping genes (OLGs), which allow a single stretch of nucleotides to encode two distinct proteins in different reading frames. Viral OLGs are common and have been associated with the origins of pandemics, but are still widely overlooked. We investigate *de novo* OLG candidates in SARS-CoV-2 and identify a new gene here named *ORF3c*. *ORF3c* has been documented elsewhere but is unnamed, unannotated, or conflated with *ORF3b* of other SARS-related betacoronaviruses (sarbecoviruses). In fact, *ORF3c* is not homologous to *ORF3b*, as the two genes occupy different genomic positions and reading frames. We find that *ORF3c* exhibits clear evidence of translation from ribosome profiling and important immunological properties. We then conduct an evolutionary analysis of *ORF3c* at three levels: between-species, between-host, and within-host. Specifically, 21 representative sarbecovirus genomes show *ORF3c* is also present in some pangolin-CoVs but not more closely related bat-CoVs; 3,978 SARS-CoV-2 genomes reveal *ORF3c* gained a new stop codon (G25563U) that rose drastically in frequency during the current COVID-19 pandemic; and 401 deeply sequenced samples of SARS-CoV-2 demonstrate the recurrence of this mutation in multiple hosts. Surprisingly, the newly gained *ORF3c* stop codon hitchhiked early with haplotype 241U/3037U/14408U/23403G (Spike-D614G), which appears to drive the European pandemic spread. Our results liken *ORF3c* to other important viral accessory genes recombined, lost, split, or truncated before or during outbreaks, including *ORF3b* and *ORF8* in sarbecoviruses. OLGs deserve considerably more attention, as their rapid evolution may be more important than is currently appreciated in the emergence of zoonotic viruses.

Introduction

The COVID-19 pandemic raises urgent questions about the properties that allow animal viruses to cross species boundaries and spread within humans. Addressing these questions requires an accurate and comprehensive understanding of viral genomes. One frequently overlooked source of novelty is the evolution of new overlapping genes (OLGs) in which an existing protein-coding nucleotide sequence is translated in a new reading frame to produce a distinct additional protein, a phenomenon known as *overprinting*. Such "genes within genes" improve genomic information compression and may offer a major source of genetic novelty (Keese and Gibbs 1992), particularly as frameshifted sequences preserve certain physicochemical properties of proteins (Bartonek et al. 2020). However, OLGs also entail the cost that a single mutation may alter two proteins, complicating sequence analyses. Moreover, genome annotation methods typically miss OLGs, favoring one open reading frame per genomic region (Warren et al. 2010). In SARS-related betacoronaviruses (subgenus *Sarbecovirus*; sarbecoviruses), OLGs are known but remain inconsistently reported. For example, absent or conflicting annotations of *ORF3b*, *ORF9b*, and *ORF9c* persist in SARS-CoV-2 reference genome Wuhan-Hu-1 (NCBI: NC_045512.2) and genomic studies (e.g., Chan et al. 2020; F. Wu et al. 2020), and no overlapping genes within *ORF3a* are displayed in the UCSC SARS-CoV-2 genome browser (Fernandes et al. 2016). Such inconsistencies stymie research, as OLGs may play a key role in the emergence of new viruses. For example, in human immunodeficiency virus-1 (HIV-1), the novel OLG *asp* (within *env*) is actively expressed in human cells (Affram et al. 2019) and is associated with the pandemic M group lineage (Cassan et al. 2016). Similarly, an OLG in SARS-CoV-1, *ORF3b* within *ORF3a*, is sometimes annotated in SARS-CoV-2 even though it contains a premature STOP codon in this virus.

Novel overlapping gene candidates

To identify OLGs within the SARS-CoV-2 genome, we first generated a complete list of candidate ORFs in the Wuhan-Hu-1 reference genome (NCBI: NC_045512.2). Specifically, we used the Schlub et al. codon permutation method (Schlub et al. 2018) to detect unexpectedly long ORFs while controlling for codon usage. One unannotated gene candidate, here named *ORF3c*, scored highly (P=0.0104), exceeding the significance of two known OLGs annotated in Uniprot (*ORF9b* and *ORF9c* [*ORF14*] within *N*; https://viralzone.expasy.org/8996) (Figure 1; Supplement).

ORF3c comprises 58 codons (including STOP) near the beginning of *ORF3a* (Table 1; Figure 2), making it longer than the known genes *ORF7b* (44 codons) and *ORF10* (39 codons) (Supplement). *ORF3c* was discovered independently by Chan et. al (2020) as '*ORF3b*' and Pavesi (2020) as, simply, 'hypothetical protein'. Due to its naming ambiguity and location within *ORF3a*, *ORF3c* has subsequently been conflated with *ORF3b* in multiple studies (Fung et al. 2020; Ge et al. 2020; Gordon et al. 2020; Hachim et al. 2020; Helmy et al. 2020; Yi et al. 2020), an extensively characterized OLG in SARS-CoV-1 and other sarbecoviruses which also overlaps *ORF3a* (McBride and Fielding 2012). In fact, *ORF3c* is unrelated to *ORF3b* as the two genes occupy different reading frames and genomic

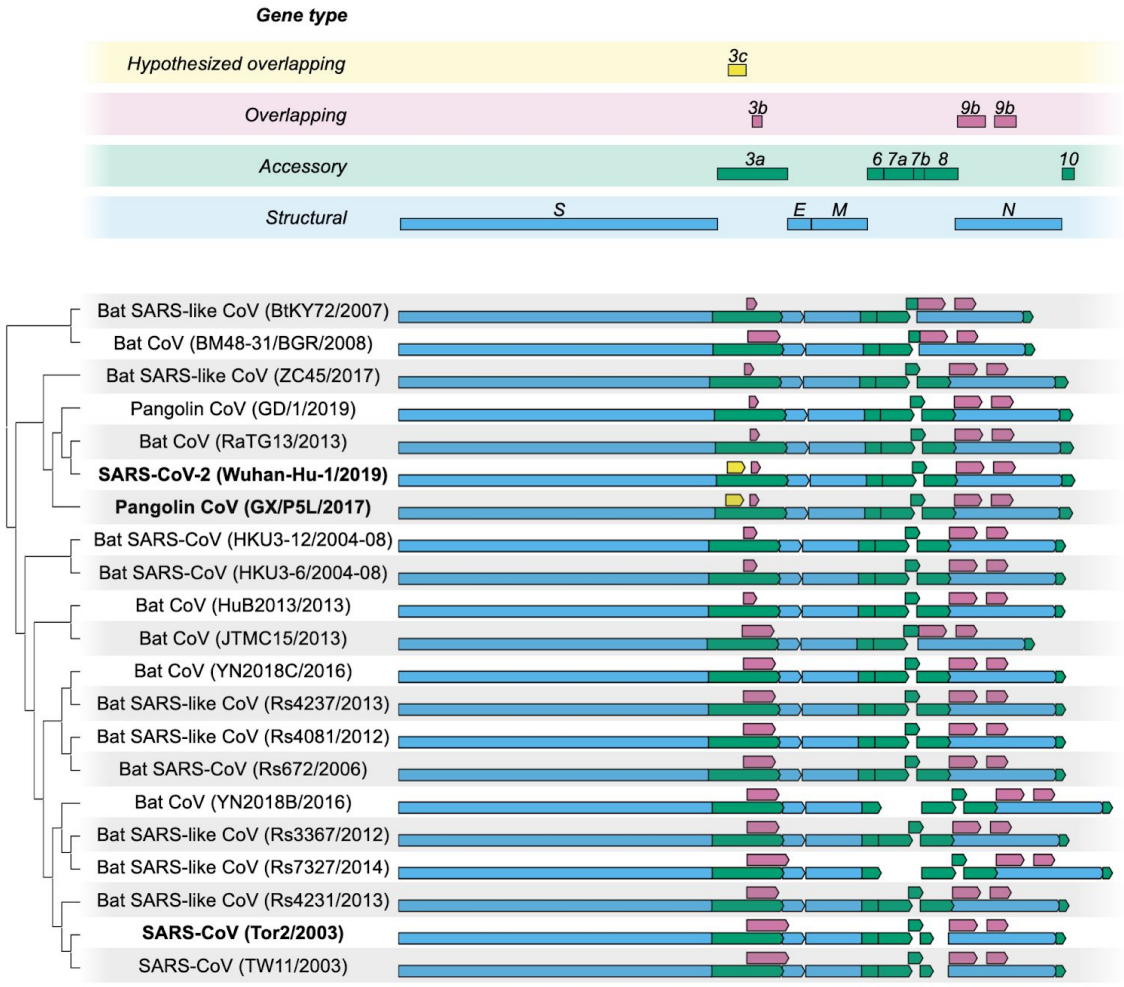


Figure 1. Sarbecovirus gene repertoire and evolutionary relationships. Only genes downstream of *ORF1ab* are shown, beginning with *S* (Spike-encoding). Four types of genes and their relative positions in the SARS-CoV-2 genome are shown on top. Genes are colored by type: hypothesized overlapping (yellow); overlapping (burgundy); accessory (green); and structural (blue). Genes with intact ORFs in each of 21 sarbecovirus genomes are shown on bottom. Positions are relative to each genome, i.e., homologous genes are not precisely aligned. Note that *ORF8* is not novel in SARS-CoV-2 as has been claimed (Chan et al. 2020), and *ORF9b* and *9c* are found throughout sarbecoviruses, though rarely annotated. *ORF3b* is full-length in only 3 sequences (SARS-CoV TW11, SARS-CoV Tor2, and bat-CoV Rs7327), while the remainder fall into two distinct classes having an early or late premature STOP codon (Supplement). *ORF8* is intact in all but 5 sequences: SARS-CoVs TW11 and Tor2, where it has split into *ORF8a* and *ORF8b*; and bat-CoVs BtKY72, BM48-31, and JTMC15, where it is deleted (i.e., only three contiguous green boxes). The full-length version of *ORF3c* is shown in SARS-CoV-2 Wuhan-Hu-1 and pangolin-CoV GX/P5L; however, note that a shorter isoform beginning later has been hypothesized (*ORF3a-iORF2*; Finkel et al. 2020) (Table 1).

Table 1. Nomenclature and reading frames for overlapping gene candidates in SARS-CoV-2 *ORF3a*.

Gene[a]	Reading frame[b]	Genome positions, Wuhan-Hu-1[c]	Description	References
ORF3a	ss11 (reference)	25393-26220 (276 codons)	Ion channel formation and virus release in SARS-CoV-1 infection; host cell apoptosis; triggers inflammation; antagonizes interferon	Lu et al. (2010); Cui et al. (2019)
ORF3h / ORF3a / ORF3a.iORF1*	ss13	25457-25582 (42 codons)	Predicted similarity to viroporin; overlaps codons 22-64 of *ORF3a*	Cagliani et al. (2020); Firth (2020); Finkel (2020); Pavesi (2020) conflates it with *ORF3b*
ORF3c	ss12	25524-25697 (58 codons)	Binds STOML2 mitochondrial protein (Gordon et al. 2020); short form contains a predicted signal peptide (Finkel et al. 2020); may contribute to differences between SARS-CoV-1 and SARS-CoV-2 in immune response as a unique antigenic target (Hachim et al. 2020); interferon antagonism has not been demonstrated; aligned to *ORF3b* by Chan et al. but is not homologous; overlaps codons 44-102 of *ORF3a*	Present study; Chan et al. (2020) and citing studies refer to it as *ORF3b* (Gordon et al. 2020; Hachim et al. 2020; etc.); Pavesi (2020) refers to it as 'hypothetical protein'
ORF3a-iORF2 / ORF3c-short	ss12	25596-25697 (34 codons)	*ORF3c*, but excluding the 24-codon upstream region harboring the majority of premature STOP codons in SARS-CoV-2; contains a predicted signal peptide (Finkel et al. 2020); overlaps codons 68-102 of *ORF3a*	Finkel et al. (2020)
ORF3a-short	ss11 (reference)	25765-26220 (152 codons)	Evidence of separate expression from 3a; has also been conflated with *ORF3b*; equivalent to codons 124-276 of *ORF3a*	Davidson et al. (2020) and pers. comm.
ORF3b	ss13	25814-26281 (ORFs at 25814-82, 25910-84, 26072-170, and 26183-281; i.e., 23, 25, 33, and 33 codons)[d]	Truncated in SARS-CoV-2; functions as interferon antagonist in SARS-related viruses; may contribute to differences between SARS-CoV-1 and SARS-CoV-2 in immune response, including asymptomatic phase; although aligned to *ORF3c* by Chan et al., is not homologous; overlaps codons 141-276 of *ORF3a*	Konno et al. (2020) (2020) claim functionality of first (23-codon) ORF in SARS-CoV-2

[a]Genes are listed by start site from 5' (top) to 3' (bottom).

[a]Nomenclature as described in Nelson et al. (2020): ss=sense-sense (same strand); ss12=codon position 1 of the reference frame overlaps codon position 2 of the overlapping frame on the same strand; ss13=codon position 1 of the reference frame overlaps codon position 3 of the overlapping frame on the same strand.

[c]Positions and numbers of codons include the STOP codons.

[d]The SARS-CoV-2 region homologous to SARS-CoV-1 *ORF3b* contains 4 premature STOP codons, resulting in the presence of four distinct ORFs (AUG-to-STOP); see Supplement.

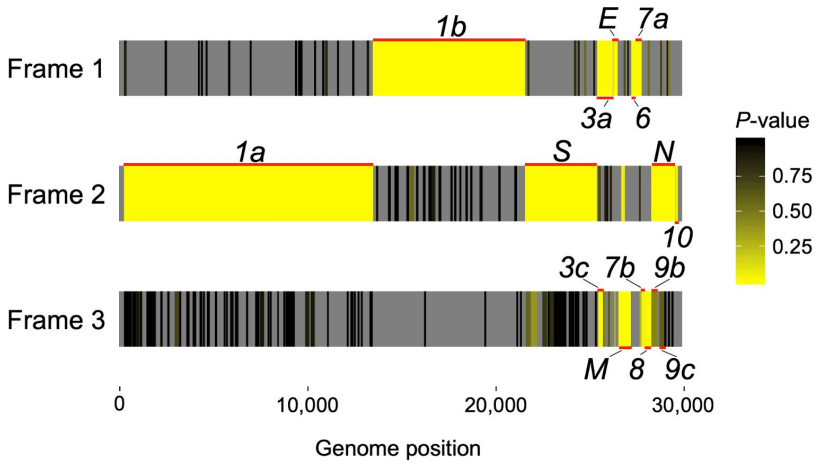


Figure 2. Codon permutation analysis to identify candidate overlapping genes in all three forward-sense reading frames. Known genes and the hypothesized *ORF3c* are indicated with horizontal red lines. Reading frames 1, 2, and 3 refer to start sites of frames beginning at position 1, 2, or 3 of the Wuhan-Hu-1 reference genome, respectively, with genome coordinates shown at the bottom (Supplement). Yellow indicates low *P*-values (natural logarithm scale), while gray indicates absence of an ORF longer than 30 codons (not tested).

positions within *ORF3a*. Specifically, *ORF3c* ends 39 codons upstream of the SARS-CoV-2 genome region homologous to *ORF3b*, where the same start site encodes only 23 codons (A. Wu et al. 2020) (Table 1; Figures 1 and 2; Supplement). It is also distinct from other OLGs hypothesized within *ORF3a* (Table 1), and an independent sequence composition analysis predicts *ORF3c* over the alternative candidate *ORF3a**/*ORF3h* (Pavesi 2020). Thus, *ORF3c* putatively encodes a novel protein not present in other sarbecoviruses, and the absence of full-length *ORF3b* in SARS-CoV-2 distinguishes it from SARS-CoV-1 (Figure 1). In contrast, *ORF3b* plays a central role in SARS-CoV-1 immune interactions and its absence or truncation in SARS-CoV-2 may be immunologically important (Konno et al. 2020; Yuen et al. 2020).

ORF3c molecular biology and expression

To assess expression of *ORF3c*, we re-analyzed the ribosome profiling (Ribo-seq) data of Finkel et al. (2020), who report a shorter isoform of *ORF3c* beginning within codon 68 of *ORF3a* (*ORF3a-iORF2*). Results for samples with ribosomes stalled by lactimidomycin and harringtonine reveal a clear peak at the start site of the full-length *ORF3c* (longer isoform), similar to the start site read distribution observed for annotated genes (Figure 3). This suggests *ORF3c* is actively translated. Referring to *ORF3c* as *ORF3b*, Gordon et al. (2020) demonstrate that stable protein expression can occur and that 3c interacts with the mitochondrial protein STOML2. However, the resolution of mass spectrometry is too low to detect short proteins (*ORF3c*, *ORF9c*, *ORF10*, or other OLG candidates Table 1). In four publicly available SARS-CoV-2 mass spectrometry datasets, signals for *ORF3c* are above a 1% false-discovery threshold (Bezstarosti et al. 2020; Bojkova et al. 2020; Davidson et al. 2020; PRIDE Project PXD018581; Methods). Despite that, structural prediction of the *ORF3c*

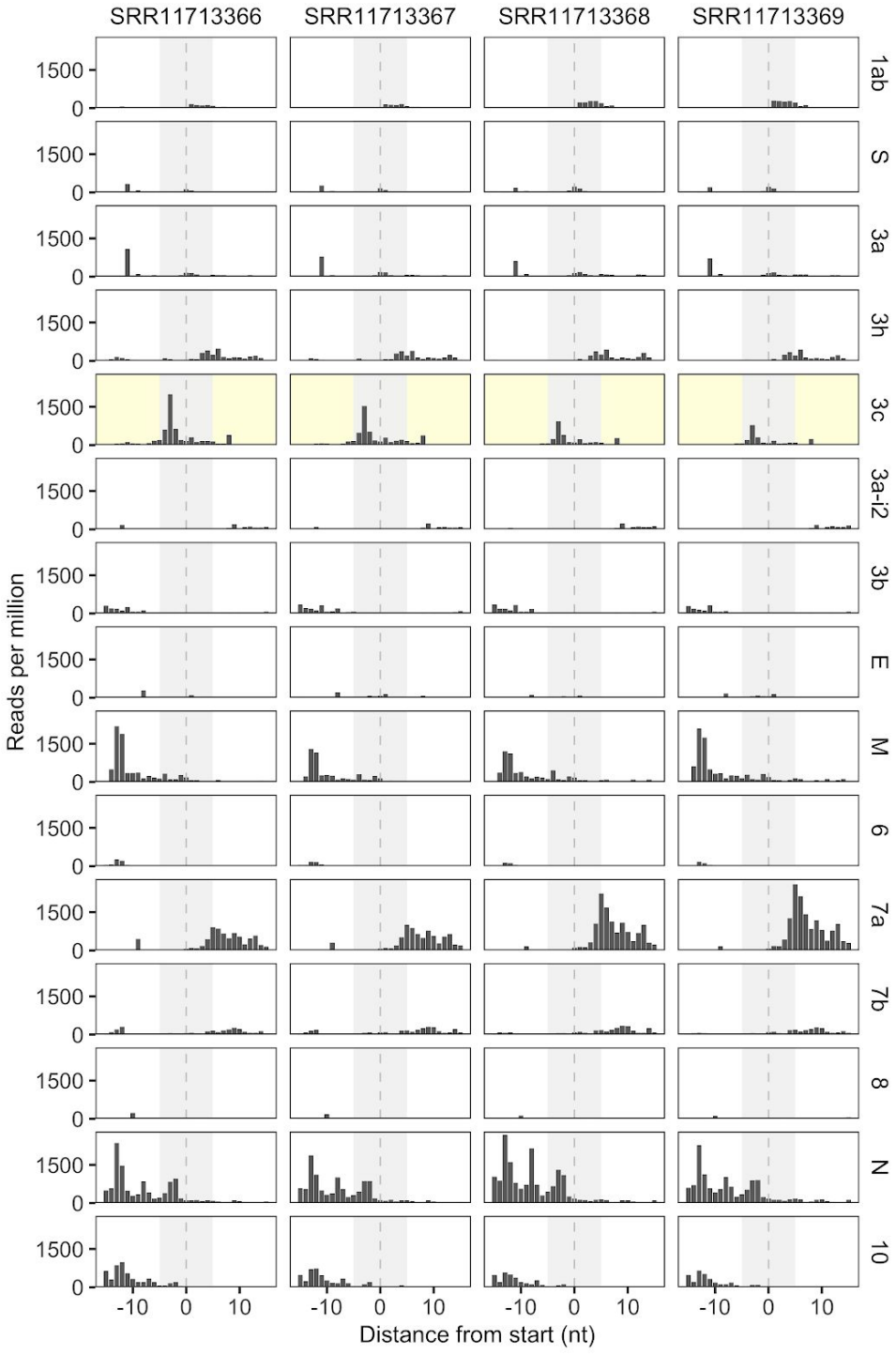


Figure 3. **Ribosome profiling re-analysis of *ORF3c* expression in four public ribosome-stalled datasets from Finkel et al. (2020).** Ribosome accumulation at the start site is a key signature of translation, emphasized in ribosome-stalled samples. *ORF3c* (yellow) shows a clear signature of ribosome accumulation, measured in reads per million mapped reads in the sample, at its hypothesized start site (vertical dashed line), exceeding start site accumulation for the two other long hypothesized ORFs overlapping *ORF3a*, namely *ORF3h* (upstream, different reading frame; Cagliani et al. 2020) and *ORF3a-iORF2* (downstream, same open reading frame but shorter; Finkel et al. 2020). Results for known genes are shown for comparison with low (e.g., *ORF8*, third from bottom) and high (*N*, second from bottom) levels of expression.

protein suggests α-helices connected with coils and an overall fold model that matches known protein structures (e.g., Protein Data Bank ID: 2WB7, 6A93) with borderline confidence (TM-score<0.514) (SFigure 1). Finally, the proteins encoded by *ORF3c* (referred to as *ORF3b*), *ORF8*, and *N* elicit the strongest antibody responses observed in COVID-19 patient sera, with *ORF3c* sufficient to accurately diagnose in the majority of COVID-19 cases (Hachim et al. 2020), providing further strong evidence of expression.

To further investigate the immunological properties of *ORF3c*, we predicted linear T-cell epitope candidates for each 9-mer of the SARS-CoV-2 proteome using NetMHCPan (Jurtz et al. 2017) to estimate MHC class I binding affinity for representative HLA alleles (Sidney et al. 2008). The lowest predicted epitope density occurs in *ORF3c*, the only gene significantly depleted compared to both short unannotated ORFs (*P*=0.019; two-sided percentile) and randomized peptides (*P*=0.044; permutation tests), followed by *ORF8* and *N* (Figure 4). Thus, the three peptides eliciting the strongest antibody (B-cell epitope) responses in SARS-CoV-2 are also predicted to contain the lowest T-cell epitope density, *ORF3c* among them. This suggests the action of selective pressures on *ORF3c* that would only be possible if its protein is produced *in situ*. Taken together, these results provide strong evidence for expression of *ORF3c*.

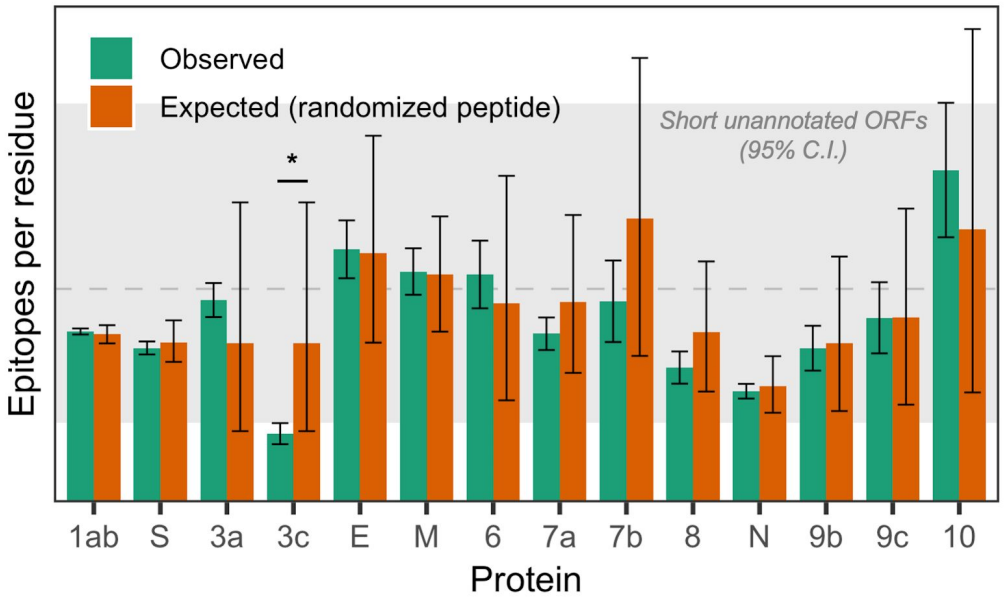


Figure 4. **Predicted T-cell epitope density per gene**. Mean number of predicted 9-amino acid epitopes per residue for each SARS-CoV-2 protein (green bars), calculated as the number of epitopes overlapping each amino acid position divided by protein length. Two sets of negative controls were used: (1) products of *n*=103 short unannotated (putatively nonfunctional) ORFs present in the SARS-CoV-2 genome, representing the result expected for real ORFs that have been evolving in the genome without functional constraint; and (2) *n*=1,000 randomized peptides generated from each protein by randomly sampling its amino acids with replacement (orange bars), representing the result expected for ORFs encoding the same amino acid content whose precise sequence has not been subjected to an evolutionary history (Supplement). Error bars show 95% confidence intervals. For nonfunctional ORFs, the horizontal gray dotted line shows the mean number of epitopes per residue, and the gray shaded region shows a 95% confidence interval. **P*=0.019, two-sided percentile for short unannotated ORFs; *P*=0.044, permutation test for randomized peptides.

ORF3c taxonomic range

To assess the origin of *ORF3c* and its conservation within and among host taxa, we created an alignment of 21 sarbecovirus genomes from Lam et al. (2020), limiting to those with an annotated *ORF1ab* and no frameshift mutations in the core genes *ORF1ab*, *S*, *ORF3a*, *E*, *M*, *ORF7a*, *ORF7b*, or *N* (Supplement). Among the sarbecoviruses, all core genes are intact (i.e., no mid-sequence STOP) in all sequences, with the exception of *ORF3c*, *ORF3b*, and *ORF8*. *ORF3c* is intact in only 2 sequences: SARS-CoV-2 Wuhan-Hu-1 and pangolin-CoVs from Guangxi (GX/P5L) (Figure 5). *ORF3b* is intact in only 3 SARS-CoV-1 sequences: SARS-CoV TW11, SARS-CoV Tor2, and bat-CoV Rs7327, with the remainder falling into two distinct groups sharing an early or late STOP codon, respectively (Supplement). Finally, *ORF8* is intact in all but 5 sequences, where it contains premature STOPs or large-scale deletions (Figure 1).

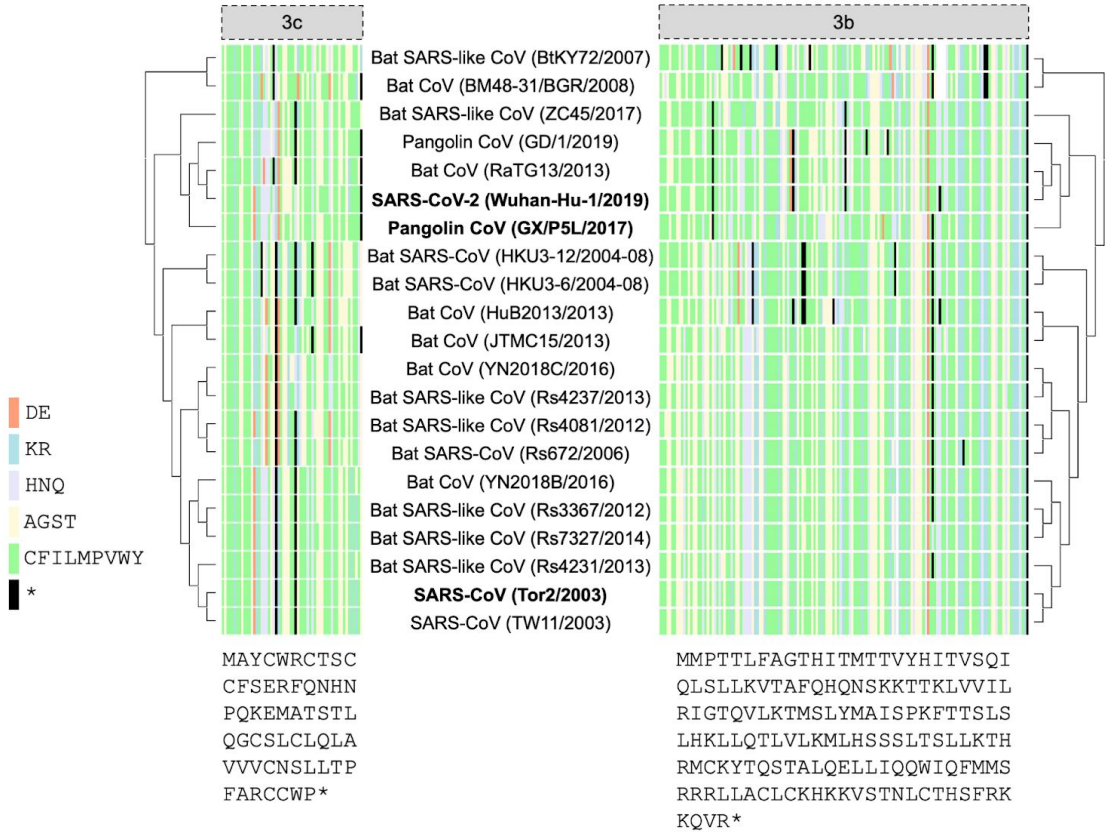


Figure 5. Amino acid variation in proteins encoded by *ORF3c* and *ORF3b* across sarbecoviruses. Amino acid alignments of 3c and 3b show their sequence conservation. Black lines indicate STOP codons in *ORF3c* and *ORF3b*, showing their restricted taxonomic ranges. Intact *ORF3c* is restricted to SARS-CoV-2 and pangolin-CoV GX/P5L, whereas *ORF3b* is found throughout the sarbecoviruses, but truncated early in most genomes outside of SARS-CoV-1. Sequences show the 3c residues of SARS-CoV-2 Wuhan-Hu-1 (57aa; NCBI=NC_045512.2; bottom left) and the 3b residues of SARS-CoV Tor2 (154aa; NCBI=NC_004718.3; bottom right).

The presence of intact *ORF3c* homologs among host species suggests possible functional conservation. However, the taxonomic distribution of this intact ORF is incongruent with whole-genome phylogenies in that *ORF3c* is present in Guangxi pangolin-CoVs (GX/P5L; more distantly related to SARS-CoV-2) but absent from Guangdong pangolin-CoVs (GD/1; more closely related to SARS-CoV-2) (Figure 5), confirmed by the alignment of Boni et al. (2020). Further, phylogenies built on *ORF3a* are also incongruent with whole-genome phylogenies, and *ORF3c* contains two STOP codons in the closely related bat-CoV NY02 (data not shown). These observations are likely due to the presence of recombination breakpoints in *ORF3a* near *ORF3c* (Boni et al. 2020; Rehman et al. 2020). Thus, recombination, convergence, or recurrent loss played a role in the origin or taxonomic distribution of *ORF3c*.

Between-species divergence

To examine natural selection on *ORF3c*, we measured diversity at three evolutionary levels: between-species (*Sarbecovirus*), between-host (human SARS-CoV-2), and within-host (human SARS-CoV-2). At each level, we inferred selection by estimating mean pairwise nonsynonymous (amino acid changing) and synonymous (not amino acid changing) nucleotide divergence (*d*; between sarbecoviruses) or diversity (π; within SARS-CoV-2) among all sequenced genomes at each level. Importantly, we combined standard (non-OLG) methods (Nei and Gojobori 1986; Nelson et al. 2015) with a new method tailored for OLGs, which we previously used to detect purifying selection on the *asp* OLG in HIV-1 (Nelson et al. 2020).

For between-species analyses, we utilized the aforementioned alignment of 21 sarbecovirus genomes unless otherwise noted. At this and all hierarchical evolutionary levels, the strongest signals of purifying selection are consistently observed in the non-OLG regions of *N* (nucleocapsid-encoding gene, which is also the most highly expressed gene (Methods, Figure 6; SFigure 2). Thus, the non-OLG regions of *N* experience disproportionately low rates of nonsynonymous change, evidencing strict functional constraint. Note that this signal can be missed if non-OLG methods are applied to *N* without accounting for its internal OLGs, *ORF9b* and *ORF9c* (e.g., P=0.0268 vs. 0.411, excluding vs. including OLG regions at the between-host level; Supplement). On the other hand, significant purifying selection is not observed at the between-species level for any gene not detected by our proteomic analysis (*ORF3c*, *ORF9c*, and *ORF10*) (Figure 6; Supplement), showing that highly expressed genes tend to exhibit the greatest functional constraint.

Comparing Wuhan-Hu-1 to pangolin-CoV GX/P5L, *ORF3c* shows d_N/d_S=0.14 (P=0.264), whereas inclusion of a third allele found in pangolin-CoV GX/P4L results in d_N/d_S=0.43 (P=0.488) (Figure 6; Supplement). Additionally, one of two possible changes synonymous in both genes is observed, but only one of 245 possible changes nonsynonymous in both genes is observed (P=0.0162, Fisher's Exact Test) (Supplement). As this evidence is suggestive of constraint, we performed sliding windows of d_N/d_S across the length of *ORF3a* to check whether potential purifying selection is specific to the expected host species and genome positions. Indeed, pairwise comparisons of each sequence to SARS-CoV-2 reveal purifying selection that is highly specific to the reading frame, genome positions, and

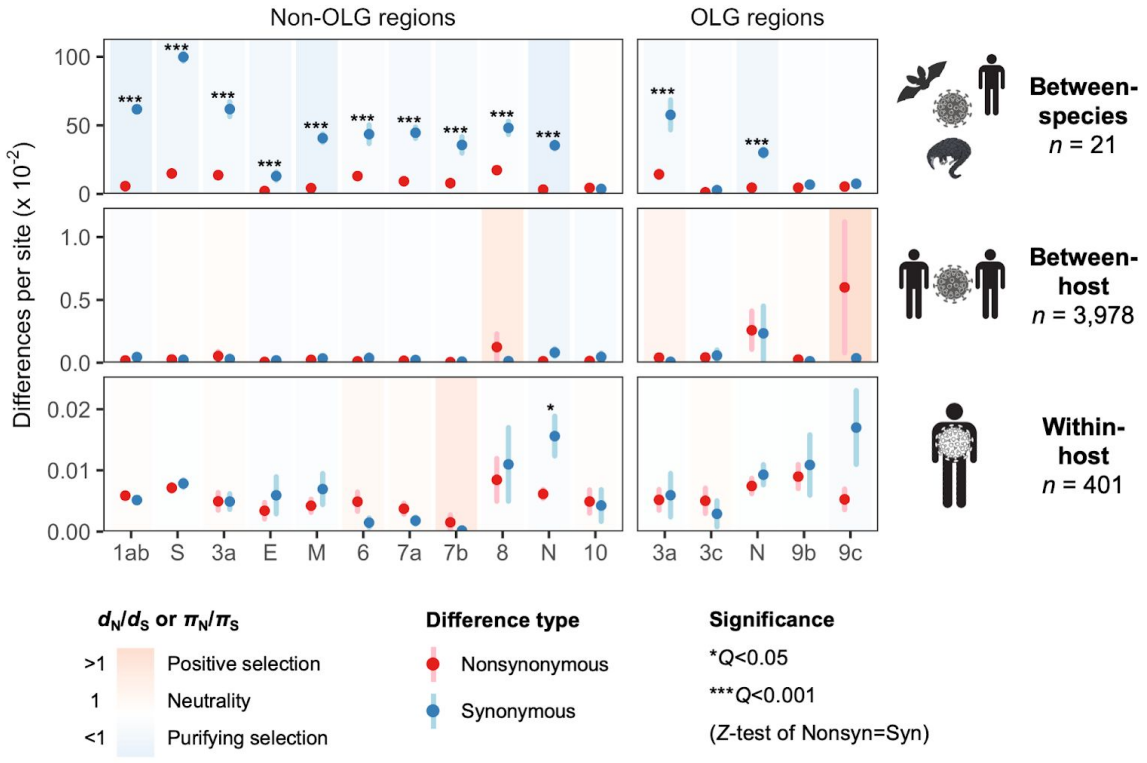


Figure 6. Natural selection analysis of nucleotide differences at three evolutionary levels. Nucleotide differences were analyzed at three levels: between-species divergence (*d*), between-host diversity (π; consensus-level), and within-host diversity (π; deep sequencing). Each gene/level is shaded according to the ratio of mean nonsynonymous to synonymous differences per site to indicate purifying selection ($d_N/d_S<1$ or $\pi_N/\pi_S<1$; blue) or positive selection ($d_N/d_S>1$ or $\pi_N/\pi_S>1$; red). For each gene, sequences were only included in the between-species analysis if a complete, intact ORF (no STOPs) was present. Genes containing a second overlapping gene (OLG) in a different frame were analyzed separately for non-OLG and OLG regions using SNPGenie and OLGenie, respectively. The short overlap between *ORF1a* and *ORF1b* (*nsp11* and *nsp12*) was excluded from analysis. Error bars represent the standard error of mean pairwise differences, estimated using 10,000 bootstrap replicates (codon unit). Significance (*Q*) refers to a Benjamini-Hochberg false-discovery rate correction after *Z*-tests of the hypothesis that d_N-d_S=0 or π_N-π_S=0, evaluated using 10,000 bootstrap replicates (codon unit). See Methods for further details.

between-species comparison where *ORF3c* is intact (SARS-CoV-2 vs. pangolin-CoV GX/P5L) (Figure 7, left). This signal is independent of whether STOP codons are present, so its consilience with the only open ORF in this region across sarbecoviruses is remarkable. The contrastive signal is also similar to that observed for known OLGs *ORF3b* in comparisons to SARS-CoV-1 (Figure 7, right) and *ORF9b* and *ORF9c* in both viruses (SFigure 3).

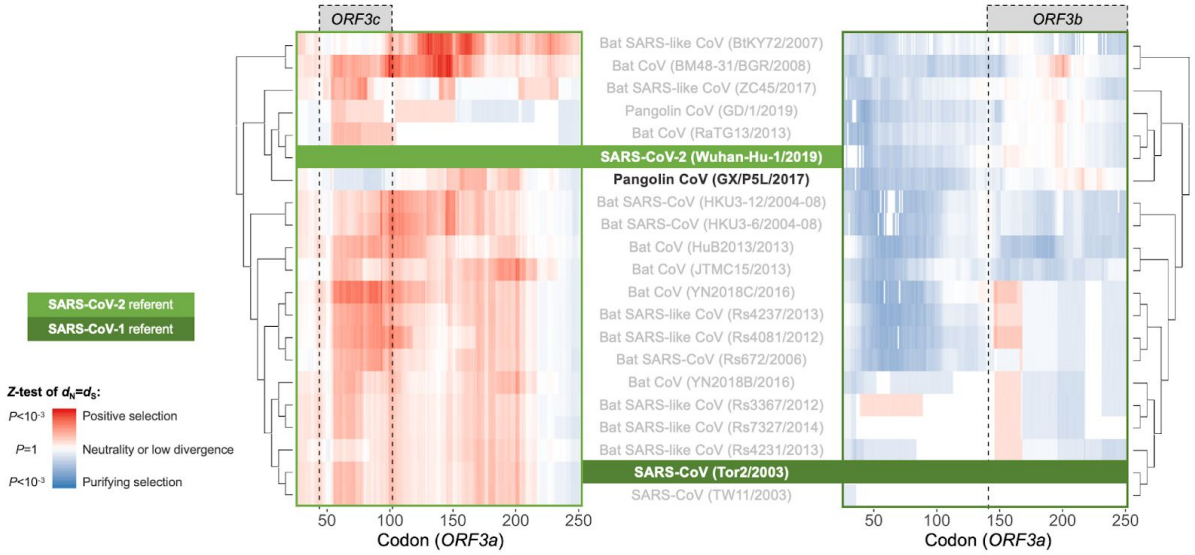


Figure 7. Between-species sliding window analysis of natural selection on overlapping frames of *ORF3a*. Pairwise analysis of selection across sarbecoviruses using OLGenie (OLG-appropriate d_N/d_S values). On the left-hand side, the ss12 frame (see Table 1 footnotes) of each sarbecovirus genome is compared with this frame in SARS-CoV-2, showing some evidence for purifying selection in the *ORF3c* region when the pangolin-CoV GX/P5L sequence is compared to SARS-CoV-2. On the right-hand side, this analysis is repeated for ss13, this time with respect to SARS-CoV-1, where *ORF3b* in ss13 is functional. Here it is seen that there is constraint in this frame across much of the gene, across sarbecoviruses.

Between-host evolution and pandemic spread

We obtained *n*=3,978 human SARS-CoV-2 consensus sequences from GISAID, limiting to whole-genome high-coverage sequences lacking indels in coding regions (accessed April 10, 2020; Supplement). Between-host diversity was sufficient to detect marginally significant purifying selection across all genes (π_N/π_S=0.50, *P*=0.0613, *Z*-test; SFigure 4A-C) but not individual genes (Figure 6). Thus, we instead investigated single mutations over time. One high-frequency mutation denoted *ORF3c*-LOF (*ORF3c*-loss-of-function) causes a STOP codon in *ORF3c* (3c-E14*) but a nonsynonymous change in *ORF3a* (3a-Q57H). This mutation increases in frequency over time in multiple locations (G25563U; SFigure 4D), raising the possibility that it experiences natural selection on *ORF3a*, *ORF3c*, or both. This variant is also not observed in any other sarbecovirus included in our analysis (Figure 5; Supplement), where the most common variant causing an *ORF3c* STOP is instead synonymous in *ORF3a* (C25614U).

With respect to *ORF3a*, *ORF3c*-LOF (G25563U) has been identified as a strong candidate for positive selection for its effect as *ORF3a*-Q57H (Kosakovsky-Pond 2020). However, temporal allele frequency trajectories (SFigure 4D) and similar signals from phylogenetic branch tests are susceptible to ascertainment bias (e.g., preferential sequencing of imported infections and uneven geographic sampling) and stochastic error (e.g., small sample sizes). Thus, we performed an independent assessment to partially account for these confounding factors. We first constructed the mutational pathway leading from the SARS-CoV-2

haplotype collected in December 2019 to the haplotype carrying *ORF3c*-LOF (G25563U). This pathway involves five mutations (C241U, C3037U, C14408U, A23403G, G25563U), constituting five observed haplotypes (EP–3 → EP–2 → EP → EP+1 → EP+1+LOF, shown in Table 2). Here, EP is suggested to have driven the European Pandemic (detected in German patient #4, footnote 3 of Table 2; Forster et al. 2020; Rothe et al. 2020); EP–3 is the Wuhan founder haplotype; and +LOF refers to *ORF3c*-LOF. We then documented the frequencies and earliest collection date of each haplotype (Table 2) to determine whether *ORF3c*-LOF occurred early on the EP background.

Surprisingly, despite its expected predominance in Europe due to founder effects, the EP haplotype is extremely rare. By contrast, haplotypes with one additional mutation (C14408U) on the EP background are common in Europe, with *ORF3c*-LOF occurring very early on this background to create EP+1+LOF from EP+1. Neither of these two haplotypes is observed in China (Table 2), suggesting that they arose in Europe subsequent to the arrival of the EP haplotype in February. Thus, we further partitioned the samples into two groups, corresponding to countries with or without early (January) samples ("early founder" and "late founder", respectively) (Figure 8). In the early founder group, EP–3 is the first haplotype detected in all countries but Germany, consistent with most early COVID-19 cases being related to travel from Wuhan. As implies that genotypes EP–3 and EP had longer to spread in the early founder group, it is surprising that their spread is dwarfed by the increase of EP+1 and EP+1+LOF starting in late February. This turnover is most obvious in the late founder group, where multiple haplotypes are detected in a narrow time window, and the number of cumulative samples is always dominated by EP+1 and EP+1+LOF. Thus, the quick spread of *ORF3c*-LOF seems to be caused by its linkage with another driver, either C14408U (+1 variant) or a subsequent variant(s) occuring on the EP+1+LOF background (Discussion). These observations highlight the necessity of empirically evaluating the effects of *ORF3c*-LOF, linked variants, and their interactions with Spike-D614G (A23403G).

Within-host diversity and mutational bias

For within-host analyses, we obtained *n*=401 high-depth (>50-fold coverage) human SARS-CoV-2 samples from the Sequence Read Archive. Within human hosts, 42% of SNPs passed our false-discovery rate criterion (Methods), with a median minor allele frequency of 2% (21 reads; 1,344 depth). The non-OLG regions of *N* again show significant purifying selection (π_N/π_S=0.39; *Q*=0.0477), but *ORF3c* remains non-significant (π_N/π_S=1.73; *Q*=0.701) (SFigure 4A, middle). We also examined 6 high-depth samples of pangolin-CoVs from Guangxi, but no conclusions could be drawn (e.g., due to low quality; Methods; Supplement).

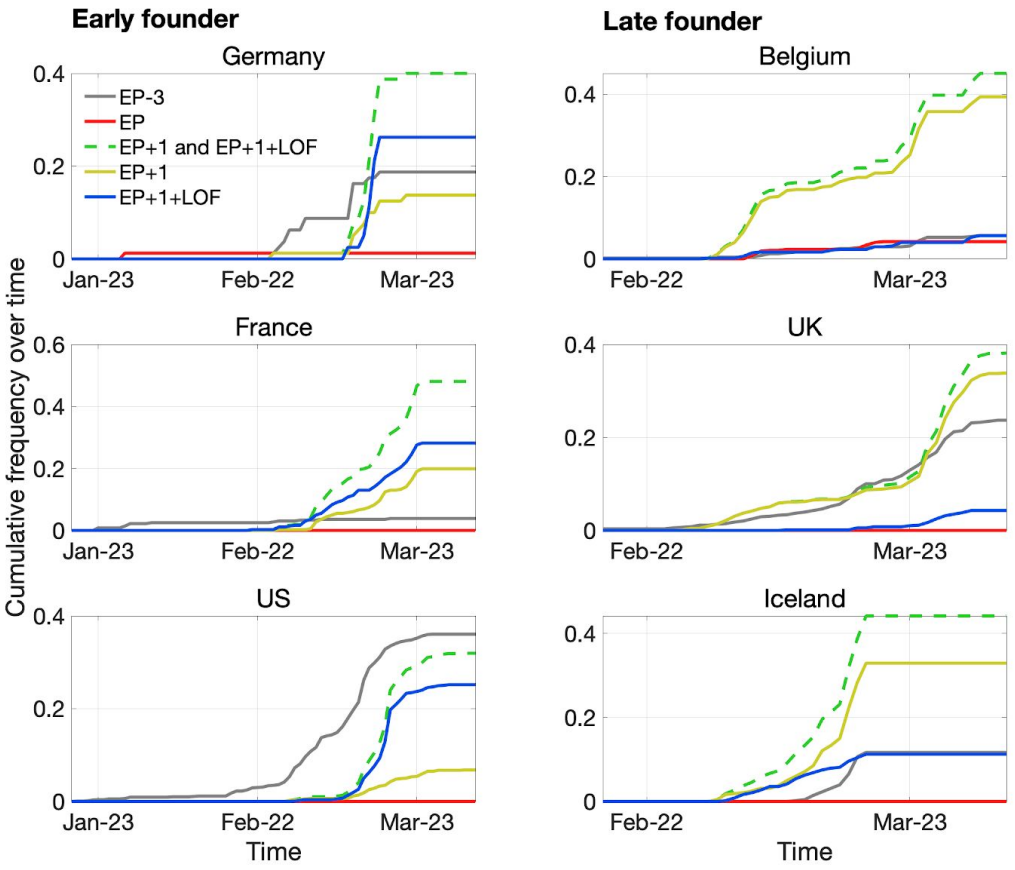


Figure 8. Pandemic spread of EP+1 haplotype and the hitchhiking of *ORF3c*-LOF. Cumulative frequencies of haplotypes in samples from Germany and five other countries with the most abundant sequence data. Countries are grouped into early founder (left) and late founder (right) based on the presence of absence of SARS-CoV-2 samples from January, respectively. In the early founder group, EP–3 (gray) is observed much earlier than other haplotypes in France and the US, and EP (red) is observed early in Germany, giving them the advantage of a founder effect. However, neither EP nor EP–3 dominate later spread. Instead, EP+1 (yellow) and EP+1+LOF (blue) increase much faster despite their later occurrence in these countries. In the late founder group, multiple haplotypes occur at almost the same time, but EP-3 and EP spread slower. The green dashed line denotes the combined frequencies of EP+1 and EP+1+LOF (yellow and blue, respectively).

Table 2. The mutational pathway to European pandemic founder haplotypes[+]

Ancestral allele Coordinate Derived allele	EP–3	EP–2	EP [3]	EP+1 [3]	EP+1+LOF
C241U	0	0	1	1	1
C3037U	0	0	1	1	1
C14408U	0	0	0	1	1
A23403G[4]	0	1	1	1	1
G25563U	0	0	0	0	1
Earliest collection [*]	24-Dec	7-Feb	28-Jan	20-Feb	21-Feb
Earliest location [*]	Wuhan	Wuhan	Munich (Shanghai) [4]	Lombardy	Hauts de France
Occurrence in China	233	1	1 (2) [4]	0	0
Occurrence in Europe	458	0	21	1153	310
Occurrence in Italy	1	0	0	27	0
Occurrence in Germany	15	0	1	11	21
Occurrence in Belgium	27	1	20	187	27
Occurrence in UK	210	0	0	338	38
Occurrence in Iceland	56	0	0	212	54
Occurrence in France	14		0	72	102
Occurrence in US	467	0	0	88 [2]	326 [2]
Occurrence in GISAID [1]	1610	2	22	1455	752

[+] The haplotypes are here defined by five variants in the table, and other variants with lower frequency on these backgrounds are ignored.

[*] The earliest collection location and time are highly subject to collection and submission bias and do not necessarily reflect where the mutation/haplotype first occurs.

[1] These numbers are based on 3853 samples from Dec 24 to Apr 1 at the time of GISAID accession that passed both our quality control procedure for alignment (based on missingness) and for this particular analysis (no ambiguous genotype calls among the five SNPs in this table) unless otherwise stated.

[2] There is likely a testing bias in the US, as EP+1+LOF haplotype was often detected in Washington, and EP+1 haplotype was not.

[3] This EP haplotype is first detected in German patient #4, a documented “founder” for coronavirus spread in Germany. However, neither EP haplotype nor EP+1 haplotype was detectable between 28-Jan and 20-Feb, but they immediately became a major haplotype when EP+1 became detectable. The failure of detecting these two haplotypes during the three weeks could potentially be explained by ascertainment bias, e.g., lack of testing for travel-independent cases.

[4] This Shanghai sample EPI_ISL_416327 has 1.32% of poly-N and failed our quality control process, added here since it is potentially relevant to the origin of EP haplotype. If this sample is included, the EP haplotype is observed in China and Shanghai twice.

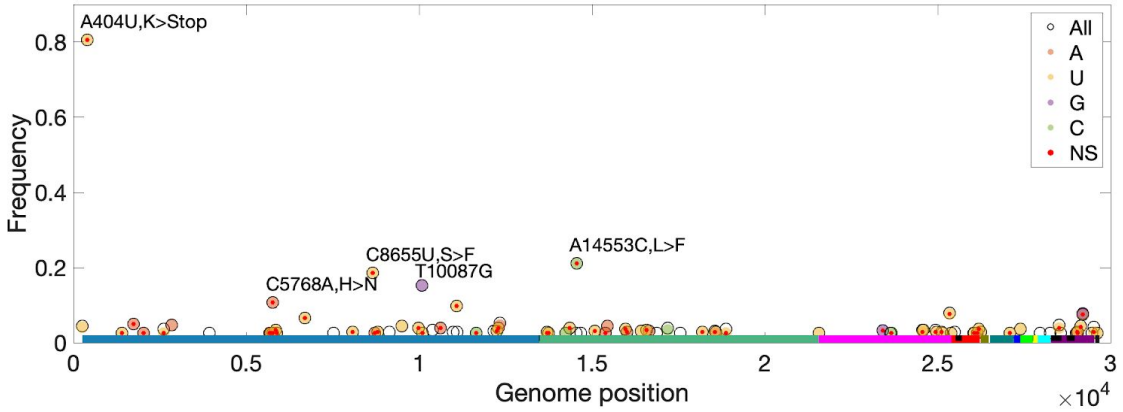


Figure 9. High-frequency within-host mutations. Mutations with a minor allele frequency of ≥0.025 in one or more of *n*=401 samples are shown. Only variants where the major allele matches the Wuhan-Hu-1 genome were considered. Each locus has up to three possible single-nucleotide derived alleles compared to the reference background. Open circles (black outlines) show the pooled frequencies of all minor alleles ("All"), while solid circles (color fill) show the frequencies of individual derived alleles. For most sites, only one derived mutation type (e.g., C→U) was observed across all samples. Precluding co-infection by multiple genotypes and sequencing errors, derived mutations occurring in more than one sample (y axis) must be identical by state but not descent (i.e., recurrent). Genome positions are plotted on the x-axis, with distinct genes shown in different colors and overlapping genes shown as a black blocks within reference genes. Nonsynonymous and nonsense mutations ("NS") are indicated with a red dot.

Our within-host analysis allowed the detection of mutations recurring in multiple samples, which might indicate mutational pressure or selective advantage. Precluding co-infection by multiple genotypes (coinfection rate x^2 is negligible when infection rate x is small), derived mutations occurring in more than one sample must be identical by state but not descent (i.e., recurrent). Limiting to 220 samples where the major allele was also ancestral (Wuhan-Hu-1 genotype), *ORF3c*-LOF is observed as a minor allele in two samples (SRR11410536 and SRR11479046 at frequencies of 0.0190 and 0.0463, respectively). This frequency of recurrent mutation (2 of 220) is high but not unusual, as 1.76% of genomic changes have an equal or higher derived allele frequency. In addition, no recurrent mutations with frequency >2.5% (Figure 9) occur in *ORF3c*. However, we note that a small number of genomic loci exhibit high rates of recurrent mutations, with five mutations observed in >10% of host samples (Methods; Figure 9, SFigure 5). Surprisingly, another STOP mutation (A404U; *nsp1*-L47*) is observed in the majority of samples (Figure 9), unexplainable by mutational bias. As NSP1 promotes host mRNA degradation and suppresses host protein synthesis in SARS-CoV-1 (Kamitani et al. 2006), its full-length form likely plays a similar role in SARS-CoV-2, and deactivated *nsp1* (A404U) may be under frequency dependent selection within-host.

Discussion

Our analyses provide strong evidence that SARS-CoV-2 contains a third overlapping gene, *ORF3c*, that has not been consistently identified or fully analyzed before this study. The annotation of a newly emerged virus is difficult, and OLGs tend to be less carefully documented than non-OLGs, for example, *ORF9b* and *ORF9c* are still not annotated in the most used reference genome, Wuhan-Hu-1 (NCBI: NC_045512.2). This difficulty is exacerbated for SARS-CoV-2 by the highly dynamic process of frequent gains and losses of accessory genes across the *Sarbecovirus* subgenus. Therefore, *de novo* and homology-based annotation are both essential, followed by careful expression analyses using multi-omic data and evolutionary analyses within and between species. In particular, we emphasize the importance of using whole-gene or genome alignments when inferring homology for both OLGs and non-OLGs, taking into account genome positions and all reading frames. Unfortunately, in the case of SARS-CoV-2, the lack of such inspections has led to mis-annotation and a domino effect. For example, homology between *ORF3b* (*Sarbecovirus*) and *ORF3c* (SARS-CoV-2) has been implied (Chan et al. 2020) and repeated (Table 1). This has led to unwarranted inferences of shared functionality (e.g., “Orf3b [*ORF3c*] is shown to be an interferon antagonist and is involved in pathogenesis”; Gordon et al. 2020) and subsequent claims of homology between *ORF3b* and other putative OLGs within ORF3a of SARS-CoV-2, e.g., *ORF3h/3a** (on the basis of a shared reading from despite having no shared genomic positions; Pavesi 2020). Given the speed of growth of SARS-CoV-2 literature, it is likely this mistake will be further promulgated. We therefore provide a detailed annotation of Wuhan-Hu-1 protein-coding genes and codons in Supplement, respectively, as a resource for future studies.

Our study highlights the highly dynamic process of frequent gains and losses of accessory genes across the *Sarbecovirus* subgenus, with the greatest functional constraint observed for the most highly expressed genes (SFigure 2). Indeed, while many or all accessory genes may be dispensable for viruses in cell culture, they often play an important role in natural hosts (Forni et al. 2017), and their loss may represent a key step in adaptation to new hosts after crossing a species barrier (Gorbalenya et al. 2006). For example, the absence of full-length *ORF3b* in SARS-CoV-2 has received attention from few authors (e.g., Lokugamage et al. 2020), even though it plays a central role in SARS-CoV-1 infection and early immune interactions as an interferon antagonist (Kopecky-Bromberg et al. 2007), with effects modulated by ORF length (Zhou et al. 2012). *ORF3b* is central in SARS-CoV-1 infection and early immune interactions and its absence or truncation in SARS-CoV-2 may be immunologically important (Yuen et al. 2020), e.g., in the suppression of type I interferon induction (Konno et al. 2020). Furthermore, the apparent presence of the *ORF3c* coincident with the inferred entry of SARS-CoV-2 into humans from a hitherto undetermined reservoir host suggests that this gene may be functionally relevant for the emergent properties of SARS-CoV-2, analogous to *asp* for HIV-1-M (Cassan et al. 2016).

Excluding OLGs when studying regions with OLG can lead to erroneous detection of natural selection (or lack thereof). For example, a synonymous variant in one reading frame is very likely to be nonsynonymous in a second overlapping frame. As a result, purifying selection against the nonsynonymous effect in the second frame will lower d_S (raise d_N/d_S) in the first

frame, increasing the likelihood of positive selection misinference (Holmes et al. 2006; Nelson et al. 2020). Such errors could, in turn, lead to mischaracterization of the genetic contributions of OLG loci to important viral properties such as incidence and persistence. One potential consequence is misguided countermeasure efforts, e.g., failure to detect functionally conserved or immunologically important regions. Finally, although only *ORF3c* is discussed in this study, other OLG candidates were also assessed at the between-host level, of which one shows evidence of translation in ribosome profiling and purifying selection (π_N/π_S=0.22, *P*=0.0278; *S-iORF2* in Finkel et al. 2020) (Table 1).

Our comprehensive evolutionary analysis of the SARS-CoV-2 genome demonstrates that many genes are under relaxed purifying selection, consistent with the exponential growth of the virus and consequent relaxation of selection (Gazave et al. 2013). At the between-host level, nucleotide diversity increases somewhat over the period 2019/12/24-2020/03/31 of the COVID-19 pandemic, tracking the number of locations sampled, while the π_N/π_S ratio remains relatively constant at 0.46 (±0.030 SEM) (SFigure 4). Other genes differ in the strength and direction of selection at the between- and within-host levels, suggesting a shift in function or importance over time. *ORF3c* and *ORF8* are both among the youngest genes in SARS-CoV-2, taxonomically restricted to a subset of betacoronaviruses (Cui et al. 2019), and *ORF8* exhibits relatively high levels of nonsynonymous change between isolates (between-host π_N/π_S ratios) (Figure 6) and frequent insertions and deletions among sarbecoviruses (Figure 1; Supplement). High between-host π_N/π_S was also observed in SARS-CoV-1 *ORF8*, perhaps due to a relaxation of purifying selection upon entry into civet cats or humans (Forni et al. 2017). However, *ORF3c* and *ORF8* both exhibit strong antibody (B-cell epitope) responses (Finkel et al. 2020) and predicted T-cell epitope depletion (Figure 4) in SARS-CoV-2. This highlights the important connection between evolutionary and immunologic processes (Daugherty and Malik 2012), as antigenic peptides allow immune detection and may impose a fitness cost for the virus. The loss or truncation of these genes may share an immunological basis and deserves further attention.

Although mutational bias marginally favors the recurrence of *ORF3c*-LOF (within-host analysis), the quick expansion of this mutation and its haplotype during this pandemic is puzzelsome (between-host analysis). One potential explanation for the slower spread of EP–3 is a sampling policy bias in case isolation; in most countries, testing and quarantine enforcement were preferentially applied to travellers who recently visited Wuhan, which may have led to selective detection, isolation, quarantining, and tracing of EP–3 and EP haplotypes. Because mutations occurring within Europe (e.g., C14408U and G25563U) are not from intercontinental travelers, we expected they would contribute more to community-acquired infections, particularly as testing biases might have provided an opportunity for them to spread in Europe. However, the EP+1 and EP+1+LOF haplotypes also grow faster in the late founder group, where it is unclear which haplotype was more travel related. Because G25563U simultaneously creates a nonsynonymous change (*ORF3a*-Q57H) and a loss of function (*ORF3c*-LOF), it could influence the spread of SARS-CoV-2 through selection on either change. Despite that, the quick spread of G25563U seems to be caused by its early occurrence in linkage with the +1 variant (C14408U causes RdRp-P323L), suggesting that this mutation is the real driver and the increase in *ORF3c*-LOF is due to hitchhiking. The spread of EP+1 and EP+1+LOF but not EP is

unexpected, as EP was the earliest haplotype and carried Spike-D614G (A23403G), a variant with predicted functional relevance (Bhattacharyya et al. 2020). Thus, it is possible that the +1 mutation (C14408U) acts synergistically with D614G or other mutations (5'UTR-C241U, *nsp3*-C3037U) unique to the EP background, causing differences among the haplotypes in infection rate, disease rate, hospitalization rate, latent period, transmission rate, or other symptoms. The faster spread of EP+1+LOF than EP+1 in the early founder countries but not late founder countries (p=0.0312) also requires explanation. These observations highlight the necessity of empirically evaluating the effects of 3c-LOF (G25563U), 3a-Q57H (G25563U), RdRp-P323L (C14408U), and their interactions with Spike-D614G (A23403G). Lastly, because we excluded minor alleles with frequencies <2.5%, it is likely that the spread *ORF3c*-LOF or other haplotypes is further assisted or hindered by subsequent mutations (Supplement).

Our study has several limitations. Short peptides can be difficult to detect using mass spectrometry methods, and the second half of 3c does not contain any potential targets. Thus, we were unable to discriminate 3c, 9c, or 10 from noise even in two high-quality datasets, a limitation likely to be true of any proteomics dataset for SARS-CoV-2. With respect to between-host diversity, we focused on relatively abundant consensus-level sequence data; however, this approach can miss important variation (Holmes 2009), stressing the importance of deeply sequenced within-host samples, sequenced with technology appropriate for calling within-host variants. As we use Wuhan-Hu-1 for read mapping and remove duplicate reads, reference bias could potentially affect our within-host results (Degner et al. 2009). We detected natural selection using counting methods that examine all pairwise comparisons between or within specific groups of sequences, which may have less power than methods that trace changes over a phylogeny. However, this approach is robust to errors in phylogenetic and ancestral sequence reconstruction, and to artifacts due to linkage or recombination (Hughes et al. 2006; Nelson and Hughes 2015). Additionally, although our method for measuring selection in OLGs does not explicitly account for mutation bias, benchmarking with other viruses suggests detection of purifying selection is conservative (Nelson et al. 2020). Finally, given multiple recombination breakpoints in *ORF3a* and the relative paucity of sequence data for viruses closely related to SARS-CoV-2, our analysis could not differentiate between convergence, recombination, or recurrent loss in the origin of *ORF3c*.

In conclusion, OLGs are an important part of viral biology that deserve more attention. We document several lines of evidence for the expression and functionality of a novel OLG in SARS-CoV-2, here named *ORF3c*, and compare it to other hypothesized OLG candidates in *ORF3a*. Finally, we provide a detailed annotation of the SARS-CoV-2 genome and highlight mutations of potential relevance to the within- and between-host evolution of SARS-CoV-2 as a resource for future studies.

Author Contributions

X.W. conceived the study. C.W.N., Z.A., and X.W. designed the study. T.G., S.-O.K., and X.W., advised on the study. C.W.N. and Z.A. obtained and processed data. C.W.N., Z.A., C-H.K., M.C., C.L., S.-O.K., and X.W., analyzed data. C.W.N., Z.A., S.-O.K., and X.W. conceived, discussed, and illustrated figures. All authors discussed and interpreted results. C.W.N., Z.A., T.G., S.-O.K., and X.W. wrote the first draft. All authors read, discussed, and revised the manuscript.

Acknowledgements

This work was funded by an Academia Sinica Postdoctoral Research fellowship (to C.W.N.; P.I. Wen-Hsiung Li); funding from the Bavarian State Government and National Philanthropic Trust (to Z.A.; P.I. Siegfried Scherer); NSF IOS grants #1755370 and #1758800 (to S.-O.K.). The authors thank the originating and submitting laboratories who kindly uploaded SARS-CoV-2 sequences to the GISAID EpiFlu™ Database for public access (Supplement), and the GISAID platform. The authors thank Maciej F. Boni, Reed A. Cartwright, John Flynn, Kyle Friend, Dan Graur, Robert S. Harbert, Cheryl Hayashi, Niloufar Kavian, Kin-Hang (Raven) Kok, Wen-Hsiung Li, Ming-Hsueh Lin, Meiyeh Lu, David A. Matthews, Lisa Mirabello, Apurva Narechania, Felix Li Jin, Priya Moorjani, Montgomery Slatkin, Yun S. Song, and attendees of the UC Berkeley popgen journal club for useful information and discussion; and special thanks to Priya Moorjani, Jacob Tennessen, Montgomery Slatkin, Jianzhi George Zhang, Meredith Yeager, Michael Dean, Hongxiang Zheng for commenting on earlier versions of this draft.

References

Affram Y, Zapata JC, Gholizadeh Z, Tolbert WD, Zhou W, Iglesias-Ussel MD, Pazgier M, Ray K, Latinovic OS, Romerio F. 2019. The HIV-1 Antisense Protein ASP Is a Transmembrane Protein of the Cell Surface and an Integral Protein of the Viral Envelope. *J Virol* 93:e00574-19.

Bartonek L, Braun D, Zagrovic B. 2020. Frameshifting preserves key physicochemical properties of proteins. *Proc Natl Acad Sci USA* 117:5907–5912.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57:289–300.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29:1165–1188.

Bhattacharyya C, Das C, Ghosh A, Singh AK, Mukherjee S, Majumder PP, Basu A, Biswas NK. 2020. Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of *TMPRSS2* and *MX1* Genes. Genomics Available from: http://biorxiv.org/lookup/doi/10.1101/2020.05.04.075911

Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry B, Castoe T, Rambaut A, Robertson DL. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Evolutionary Biology Available from: http://biorxiv.org/lookup/doi/10.1101/2020.03.30.015008

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol* 55:756–768.

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.

Cassan E, Arigon-Chifolleau A-M, Mesnard J-M, Gross A, Gascuel O. 2016. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. *Proc Natl Acad Sci USA* 113:11537–11542.

Chan JF-W, Kok K-H, Zhu Z, Chu H, To KK-W, Yuan S, Yuen K-Y. 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerging Microbes & Infections* 9:221–236.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26:1367–1372.

Cui J, Li F, Shi Z-L. 2019. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology* 17:181–192.

Daugherty MD, Malik HS. 2012. Rules of Engagement: Molecular Insights from Host-Virus Arms Races. *Annual Review of Genetics* 46:677–700.

Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom K, Zambon M, Ellis J, Lewis PA, Hiscox JA, et al. 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 induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. Microbiology Available from: http://biorxiv.org/lookup/doi/10.1101/2020.03.22.002204

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25:3207–3212.

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research* 15:330–340.

Ewens WJ, Grant GR. 2001. Statistical Methods in Bioinformatics. New York: Springer-Verlag

Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, Hernandez RD, Frankel AD. 2016. Functional Segregation of Overlapping Genes in HIV. *Cell* 167:1762–1773.

Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Yahalom-Ronen Y, Tamir H, Achdout H, Melamed S, Weiss S, Israely T, et al. 2020. The coding capacity of SARS-CoV-2. Microbiology Available from: http://biorxiv.org/lookup/doi/10.1101/2020.05.07.082909

Firth AE. 2020. A putative new SARS-CoV protein, 3a*, encoded in an ORF overlapping ORF3a. Bioinformatics Available from: http://biorxiv.org/lookup/doi/10.1101/2020.05.12.088088

Forni D, Cagliani R, Clerici M, Sironi M. 2017. Molecular Evolution of Human Coronavirus Genomes. *Trends in Microbiology* 25:35–48.

Forster P, Forster L, Renfrew C, Forster M. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci USA*:202004999.

Fung S-Y, Yuen K-S, Ye Z-W, Chan C-P, Jin D-Y. 2020. A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses. *Emerging Microbes & Infections* 9:558–570.

Gazave E, Chang D, Clark AG, Keinan A. 2013. Population Growth Inflates the Per-Individual Number of Deleterious Mutations and Reduces Their Mean Effect. *Genetics* 195:969–978.

Ge H, Wang X, Yuan X, Xiao G, Wang C, Deng T, Yuan Q, Xiao X. 2020. The epidemiology and clinical information about COVID-19. *European Journal of Clinical Microbiology & Infectious Diseases* 39:1011–1019.

Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. 2006. Nidovirales: Evolving the largest RNA virus genome. *Virus Research* 117:17–37.

Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL, et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* [Internet]. Available from: http://www.nature.com/articles/s41586-020-2286-9

Hachim A, Kavian N, Cohen CA, Chin AW, Chu DK, Mok CK, Tsang OT, Yeung YC, Perera RA, Poon LL, et al. 2020. Beyond the Spike: identification of viral targets of the antibody responses to SARS-CoV-2 in COVID-19 patients. medRxiv Available from: https://doi.org/10.1101/2020.04.30.20085670

Helmy YA, Fawzy M, Elaswad A, Sobieh A, Kenney SP, Shehata AA. 2020. The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis,

Treatment, and Control. *JCM* 9:1225.
Holmes EC. 2009. The Evolution and Emergence of RNA Viruses. New York: Oxford University Press
Holmes EC, Lipman DJ, Zamarin D, Yewdell JW. 2006. Comment on “Large-Scale Sequence Analysis of Avian Influenza Isolates.” *Science* 313:1573b–1573b.
Hughes AL, Friedman R, Glenn NL. 2006. The Future of Data Analysis in Evolutionary Genomics. *Current Genomics* 7:227–234.
Jukes TH, Cantor CR. 1969. Evolution of Protein Molecules. In: Munro HN, editor. Mammalian Protein Metabolism. New York: Academic Press. p. 21–132. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9781483232119500097
Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J.I.* 199:3360–3368.
Kamitani W, Narayanan K, Huang C, Lokugamage K, Ikegami T, Ito N, Kubo H, Makino S. 2006. Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proceedings of the National Academy of Sciences* 103:12885–12890.
Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30:772–780.
Keese PK, Gibbs A. 1992. Origins of genes: “big bang” or continuous creation? *Proceedings of the National Academy of Sciences* 89:9489–9493.
Konno Y, Kimura I, Uriu K, Fukushi M, Irie T, Koyanagi Y, Nakagawa S, Sato K. 2020. SARS-CoV-2 ORF3b is a potent interferon antagonist whose activity is further increased by a naturally occurring elongation variant. Microbiology Available from: http://biorxiv.org/lookup/doi/10.1101/2020.05.11.088179
Kopecky-Bromberg SA, Martínez-Sobrido L, Frieman M, Baric RA, Palese P. 2007. Severe Acute Respiratory Syndrome Coronavirus Open Reading Frame (ORF) 3b, ORF 6, and Nucleocapsid Proteins Function as Interferon Antagonists. *JVI* 81:548–557.
Kosakovsky-Pond SL. 2020. Natural selection analysis of SARS-CoV-2/COVID-19. *usegalaxy* [Internet]. Available from: https://covid19.galaxyproject.org/evolution/
Lam TT-Y, Shum MH-H, Zhu H-C, Tong Y-G, Ni X-B, Liao Y-S, Wei W, Cheung WY-M, Li W-J, Li L-F, et al. 2020. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* [Internet] in press. Available from: http://www.nature.com/articles/s41586-020-2169-0
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359.
Langmead B, Wilks C, Antonescu V, Charles R. 2019. Scaling read aligners to hundreds of threads on general-purpose processors.Hancock J, editor. *Bioinformatics* 35:421–432.
Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30:3276–3278.
Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges.Prlic A, editor. *PLoS Comput Biol* 9:e1003118.
Lokugamage KG, Hage A, Schindewolf C, Rajsbaum R, Menachery VD. 2020. SARS-CoV-2 is sensitive to type I interferon pretreatment. Microbiology Available from: http://biorxiv.org/lookup/doi/10.1101/2020.03.07.982264
Lu W, Xu K, Sun B. 2010. SARS Accessory Proteins ORF3a and 9b and Their Functional Analysis. In: Lal SK, editor. Molecular Biology of the SARS-Coronavirus. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 167–175. Available from: https://doi.org/10.1007/978-3-642-03683-5_11
McBride R, Fielding B. 2012. The Role of Severe Acute Respiratory Syndrome (SARS)-Coronavirus Accessory Proteins in Virus Pathogenesis. *Viruses* 4:2902–2923.
McKinney W. 2010. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. Vol. 445. p. 51–56.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3:418–426.
Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York, NY: Oxford University Press
Nelson CW, Ardern Z, Wei X. 2020. OLGenie: Estimating Natural Selection to Predict Functional Overlapping Genes. *Molecular Biology and Evolution* in press:msaa087.
Nelson CW, Hughes AL. 2015. Within-host nucleotide diversity of virus populations: Insights from

next-generation sequencing. *Infection, Genetics and Evolution* 30:1–7.
Nelson CW, Moncla LH, Hughes AL. 2015. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics* 31:3709–3711.
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32:268–274.
Pavesi A. 2020. New insights into the evolutionary features of viral overlapping genes by discriminant analysis. *Virology* 546:51–66.
R Core Team. 2018. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing Available from: https://www.R-project.org/
Rehman S ur, Shafique L, Ihsan A, Liu Q. 2020. Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2. *Pathogens* 9:240.
Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, Zimmer T, Thiel V, Janke C, Guggemos W, et al. 2020. Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. *N Engl J Med* 382:970–971.
Schlub TE, Buchmann JP, Holmes EC. 2018. A simple method to detect candidate overlapping genes in viruses using single genome sequences. *Molecular Biology and Evolution* 35:2572–2581.
Sidney J, Peters B, Frahm N, Brander C, Sette A. 2008. HLA class I supertypes: a revised and updated classification. *BMC Immunology* 9:1.
Soubrier J, Steel M, Lee MSY, Der Sarkissian C, Guindon S, Ho SYW, Cooper A. 2012. The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. *Molecular Biology and Evolution* 29:3345–3358.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34:W609–W612.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* 17:57–86.
Warren AS, Archuleta J, Feng W, Setubal JC. 2010. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* 11:131.
Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research* 40:11189–11201.
Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J, et al. 2020. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host & Microbe* 27:325–328.
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.
Yang Z. 1995. A Space-Time Process Model for the Evolution of DNA Sequences. *Genetics* 139:993–1005.
Yi Y, Lagniton PNP, Ye S, Li E, Xu R-H. 2020. COVID-19: what has been learned and to be learned about the novel coronavirus disease. *Int. J. Biol. Sci.* 16:1753–1766.
Yuen K-S, Ye Z-W, Fung S-Y, Chan C-P, Jin D-Y. 2020. SARS-CoV-2 and COVID-19: The most important research questions. *Cell Biosci* 10:40.
Zhou P, Li H, Wang H, Wang L-F, Shi Z. 2012. Bat severe acute respiratory syndrome-like coronavirus ORF3b homologues display different interferon antagonist activities. *Journal of General Virology* 93:275–281.

Methods

Genomic features and coordinates

All genome coordinates are given with respect to reference sequence Wuhan-Hu-1 (NCBI: NC_045512.2; GISAID: EPI_ISL_402125) unless otherwise noted. SARS-CoV-1 genome coordinates are given with respect to reference sequence Tor2 (NC_004718.3). SARS-CoV-2 Uniprot peptides were obtained from https://viralzone.expasy.org/8996, where *ORF9c* is referred to as *ORF14*. Nucleotide sequences were translated using R::Biostrings (Lawrence et al. 2013), Biopython (Cock et al. 2009), or SNPGenie (Nelson et al. 2015). Alignments were viewed and edited in AliView v1.20 (Larsson 2014). To identify OLGs using the Schlub et al. codon permutation method (Schlub et al. 2018), all 12 ORFs annotated in the Wuhan-Hu-1 reference genome were used as a reference (NCBI=NC_045512.2; *ORF1a*, *ORF1b*, *S*, *ORF3a*, *E*, *M*, *ORF6*, *ORF7a*, *ORF7b*, *ORF8*, *N*, and *ORF10*).

SARS-CoV-2 genome data and alignments

SARS-CoV-2 genome sequences were obtained from GISAID on April 10, 2020 (Supplement). Whole genomes were aligned using MAFFT v7.455 (Katoh and Standley 2013), and subsequently discarded if they contained internal gaps (-) >900 nt from either terminus, a length sufficient to exclude sequences with insertions or deletions (indels) in coding regions. Coding regions were identified using exact or partial homology to SARS-CoV-2 or SARS-CoV-1 annotations.

Sarbecovirus genome data and alignments

SARS-CoV-related genome IDs were obtained from Lam et al. (2020) and downloaded from GenBank or GISAID. Only genotype Wuhan-Hu-1 was used to represent SARS-CoV-2. Except for pangolin-specific analyses, only genotypes GX/P5L and GD/1 were used to represent pangolin-COVs; GD/1 was chosen as a representative because the other lacks the *S* gene and contains 27.76% Ns, while GX/P5L was chosen because it is one of two high-coverage sequences derived from lung tissue that also contains no Ns. Other sequences were excluded if they lacked an annotated *ORF1ab* with a ribosomal slippage, or contained a frameshift indel in any gene, leaving 21 sequences for analysis (Supplement). To produce whole-genome alignments, we first aligned all sequences using MAFFT. Then, coding regions were identified using exact or partial sequence identity to SARS-CoV-2 or SARS-CoV-1 annotations, translated, and aligned at the amino acid level using ProbCons v1.12 (Do et al. 2005). The longest gene was used in the case of OLGs. Amino acid alignments were then imposed on the coding sequence of each gene using PAL2NAL v14 (Suyama et al. 2006: 2) to maintain complete codons. Finally, whole genomes were manually shifted to match the codon alignments in AliView. Codon breaks were preferentially resolved to align S/Q/T at 3337-3339 and L with T/I at 3343-3345 because of biochemical similarity. This preserved all nucleotides of each genome while concurrently incorporating codon-aware alignments.

Phylogenetic analysis and ancestral sequence reconstruction

Phylogenetic relationships among isolates were explored using maximum likelihood phylogenetic inference, as implemented in IQ-tree (Nguyen et al. 2015). The generalized time-reversible (GTR; Tavaré 1986) and non-reversible (asymmetric substitution matrix; Boussau and Gouy 2006) were contrasted based on their logLik value, while accounting for among-site rate heterogeneity using discrete rate categories modeled by the Γ distribution (Yang 1995) and the FreeRate model (Soubrier et al. 2012).

Proteomics analysis

iBAQ values (proportional estimates of the molar protein quantity of a protein in a given sample, allowing relative quantitative comparisons) were computed using the Max-Quant software (Cox and Mann 2008) as the sum of all peptide intensities per proteins divided by the number of theoretical peptides per protein.

Ribo-seq analysis

Ribo-seq datasets with accession numbers SRR117133166, SRR117133167, SRR117133168, and SRR117133169 (Finkel et al. 2020) were downloaded from the Sequence Read Archive. These samples comprised the data for ribosomes stalled with either lactimidomycin or harringtonine, with the Vero E6 cells harvested at 24hours post infection, and had higher sequence coverage depth than other samples, allowing for reliable start determination. They were mapped to the Wuhan-Hu-1 reference genome with the sequenced strain's mutations, as listed in Finkel et al. (2020). Mapping used Bowtie2 (Langmead et al. 2019) local alignment, with a seed length of 20 and up to one mismatch allowed. Mapped reads within 15 nucleotides of each putative start site were then counted and counts plotted using the 5' most mapped position of each read.

NetMHCPan T-cell epitope analysis

Viral protein sequences were analyzed using 9-mer substrings in NetMHCpan4.0 (Jurtz et al. 2017). Twelve (12) HLA supertype representative were used in the analysis: HLA-A*01:01 (A1), HLA-A*02:01 (A2), HLA-A*03:01 (A3), HLA-A*24:02 (A24), HLA-A*26:01 (A26), HLA-B*07:02 (B7), HLA-B*08:01 (B8), HLA-B*27:05 (B27), HLA-B*39:01 (B39), HLA-B*40:01 (B44), HLA-B*58:01 (B58), and HLA-B*15:01 (B62). NetMHCpan4.0 returns percentile ranks that characterize a peptide's likelihood of antigen presentation compared to a set of random natural peptides. We employed the suggested threshold of 2% to determine potential presented peptides, and 0.5% to identify strong MHC binder. Both strong and weak binders were considered predicted epitopes.

Statistical tests on synonymous and nonsynonymous rate

Statistical and data analyses and visualization were carried out in R v3.5.2 (R Core Team 2018) (libraries: boot, RColorBrewer, scales, tidyverse), Python (BioPython, pandas) (McKinney 2010), Excel, Google Sheets, and PowerPoint. Colors were explored using Coolors (https://coolors.co). Copyright-free images of a bat, human, and pangolin were obtained from Pixabay (https://pixabay.com). Only two-sided *P*-values are reported for statistical tests. For known and putative OLGs, the d_N/d_S (π_N/π_S) ratio was estimated using d_{NN}/d_{SN} (π_{NN}/π_{SN}) for the reference frame and or d_{NN}/d_{NS} (π_{NN}/π_{NS}) for the alternate frame (ss12 or ss13), because the number of SS (synonymous/synonymous) sites was insufficient to estimate d_{SS} (π_{SS}). Unless otherwise noted, the null hypothesis of d_N-d_S=0 (π_N-π_S=0) was evaluated using both *Z* and achieved significance level (ASL) tests (Nei and Kumar 2000) with 10,000 and 1,000 bootstrap replicates for genes and sliding windows, respectively, using individual codons (alignment columns) as the resampling unit (Nei and Kumar 2000). For ASL, *P*-values of 0 were reported as the lowest non-zero value possible given the number of bootstrap replicates. Benjamini-Hochberg (Benjamini and Hochberg 1995) or Benjamini-Yekutieli (Benjamini and Yekutieli 2001) false-discovery rate corrections (*Q*-values) were used for genes (independent regions) and sliding windows (contiguous overlapping regions), respectively.

Between-species analyses

Because the uncorrected *d* value often exceeded 0.1 in between-species comparisons, a Jukes-Cantor correction (Jukes and Cantor 1969) was applied to d_N and d_S estimates. For each ORF, sequences were only used to estimate d_N/d_S if a complete, intact ORF (no STOPs) was present. Additionally, the following codons were excluded from analysis: codons 1-13 of *E*, which overlap *ORF3b* in SARS-CoV-related taxa; codons 62-64 of *ORF6*, which follow an early STOP in some taxa; and codons 72-74 of *ORF9c*, which following an early STOP in some taxa.

Between-host analyses

To quantify the diversity and evenness of sample locations, we quantified their entropy as $-\sum p*\ln(p)$, where *p* is the number of distinct (unique) locations or countries reported for a given window (Ewens and Grant 2001).

Cumulative haplotype frequency

We define haplotypes along the mutational pathway using all five high DAF mutations from Wuhan-Hu-1 to *ORF3c*-LOF, and subsequent mutations after *ORF3c*-LOF are ignored in the haplotype analysis. Samples with missing data at any of the five loci are removed from the haplotype analysis. We calculate the cumulative haplotype frequency of each haplotype in Germany (where EP haplotype is a documented founder) and five other countries with the most abundant samples by the time of data accession. The cumulative frequency is calculated as the total number of occurrences of each haplotype collected by each day divided by the total number of samples from the same country. Countries are subsequently divided into early founders and late founders to investigate founder effects. Early founder countries tend to have more than a few samples from January, and late founders tend to have samples collected only after mid-February.

Within-host diversity

For within-host analyses, we obtained *n*=401 high-depth (at least 50-fold mean coverage) human SARS-CoV-2 samples sequenced with Illumina technology, from the Sequence Read Archive. Only Illumina samples were used as some Nanopore samples exhibited apparent systematic bias in calling putative intrahost SNPs, and this technology has also been shown to be unsuitable for intra-host analysis (Grubaugh et al. 2019). Reads were trimmed with BBDUK (Bushnell B. 2017. BBTools. https://jgi.doe.gov/data-and-tools/bbtools/), and mapped against the Wuhan-Hu-1 reference sequence using Bowtie2 (Langmead and Salzberg 2012) with local alignment, seed length 20, and up to 1 mismatch. SNPs were called using the LoFreq (Wilm et al. 2012) variant caller from mapped reads with sequencing quality and MAPQ both at least 30. Only single-end or the first read in a pair of paired-end reads were used. Variants were dynamically filtered based on each site's coverage using a binomial cutoff to ensure a false-discovery rate of ≤1 within-host variant in our study (401 samples), assuming a mean sequencing error rate of 0.2% (Schirmer et al. 2016).

To estimate π, numbers of nonsynonymous and synonymous differences and sites were first calculated individually for each of the 401 samples using SNPGenie (Nelson et al. 2015) (https://github.com/chasewnelson/SNPGenie). Next, average within-host numbers of differences and sites were calculated for each codon by taking the mean across all samples. For example, if a particular codon contained nonsynonymous differences in two of 401 samples, with the two samples exhibiting mean numbers of 0.01 and 0.002 pairwise differences per site, this codon was considered to exhibit a mean of (0.01+0.002)/401=0.0000299 pairwise differences per site across all samples. These codon means were then treated as independent units of observation during bootstrapping.

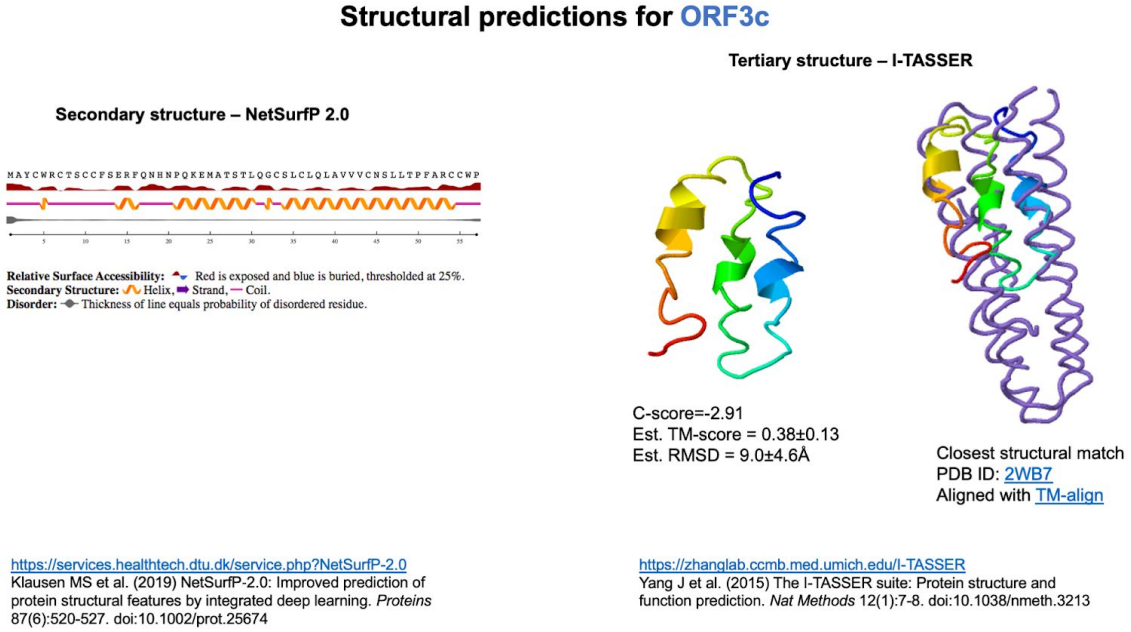
Pangolin samples examined refer to Sequence Read Archive records SRR11093266, SRR11093267, SRR11093268, SRR11093269, SRR11093270, SRR11093271. Only 179 single

nucleotide variants were called prior to our FDR filtering, and samples SRR11093271 and SRR11093270 were discarded entirely due to low mapping quality. We also note that after our quality filtering, 4 samples contain consensus alleles that do not match their reference sequence (at GISAID): P1E, P4L, P5E, and P5L (Supplement).

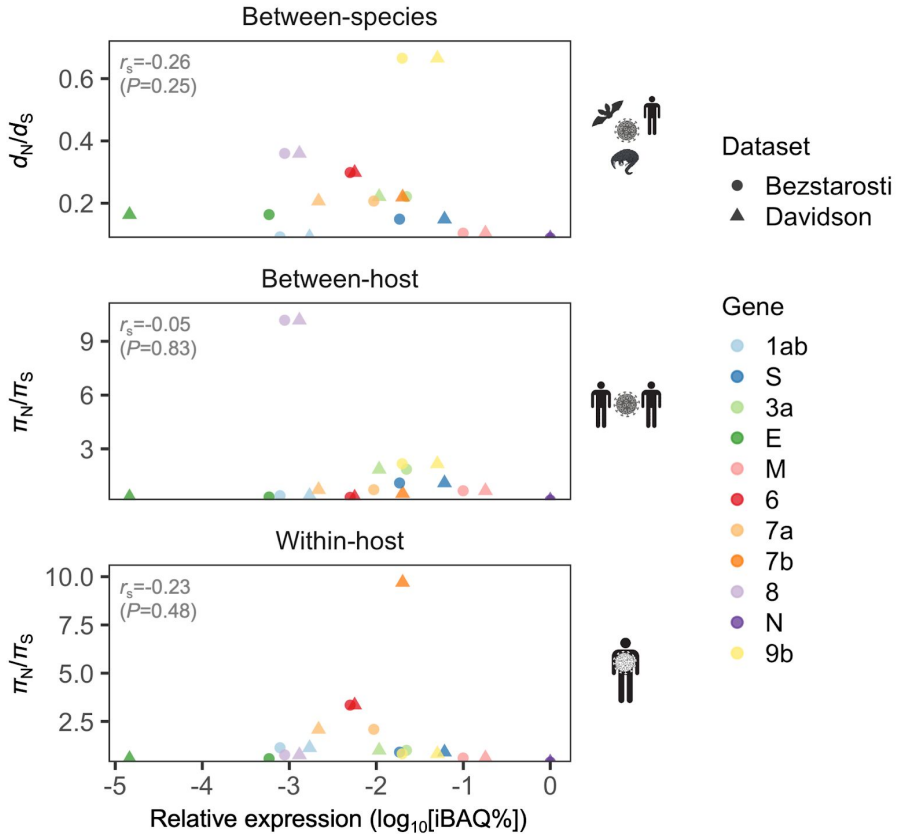
Within-host recurrent mutations analyses

We assume that each host was infected by a small number of viruses of the same genotype. Under this assumption, the minor allele of each segregating site within-host is either due to genotyping and sequencing artifacts or due to new mutations during or post replication. Because there are very few loci with high frequency derived allele between-host, and because the Wuhan-Hu-1 genome is used as the reference in mapping, we here only consider within-host mutations against the reference background. There are four possible bases at each locus, A, U, G, and C, and three possible mutational directions against the Wuhan-Hu-1 reference genome. For each locus, we calculate the number of samples with reference allele as $N = N_1 + N_2$, where N_1 is the number of samples that all reads mapped to the Wuhan reference allele, and N_2 is the number of samples that the Wuhan reference allele is major allele. Out of N_2, we calculated the number of samples carrying each of the three possible non-reference alleles, as N_A, N_U, N_G, N_C. If A is one of the observed non-reference alleles, we would calculate the frequency of A as $p_A = N_A/N$. If the reference allele is U, we calculate p_A, p_G, p_C, and $p_{All} = p_A + p_G + p_C$. A larger frequency indicates the derived allele is observed in many samples. The Derived Allele Frequency (DAF) within-host is calculated as the total number of reads mapped to the observed minor allele divided by the total number of reads mapped to the locus. If all reads are mapped to the Wuhan-Hu-1 reference allele, then the DAF = 0. There are five mutations that occur in more than 10% of the samples, four of which are nonsynonymous, for which we plotted their DAF within-host. For this analysis, we did not apply the per-site FDR cutoff, thus a DAF=0 is equivalent to the absence of reads mapped to the mutation, after reads are filtered by sequence quality, mapping quality and LoFreq's default significance threshold (P-value = 0.010000).

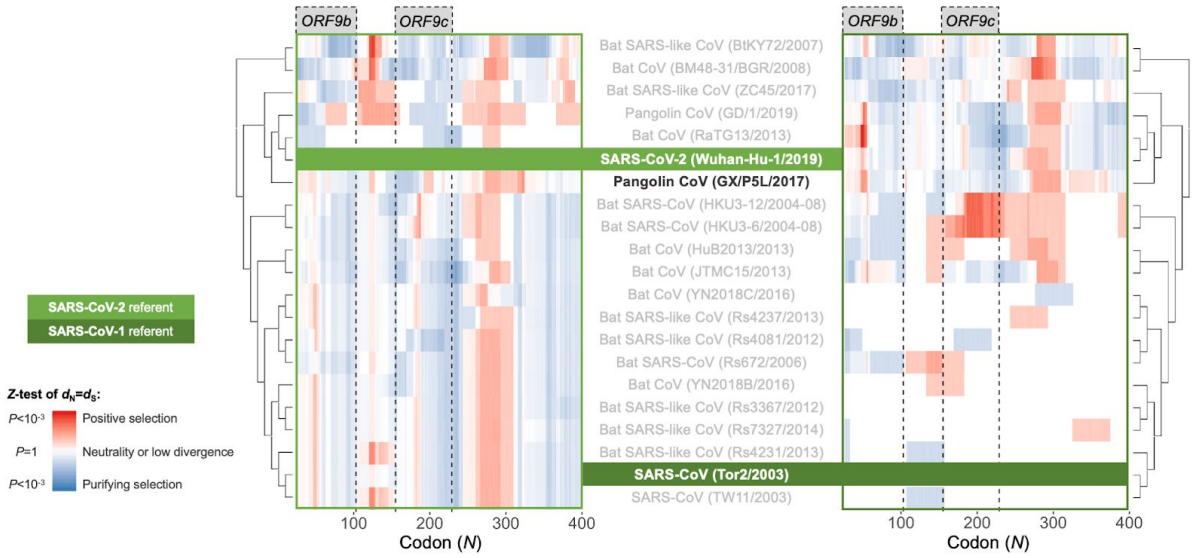
Supplementary Figures



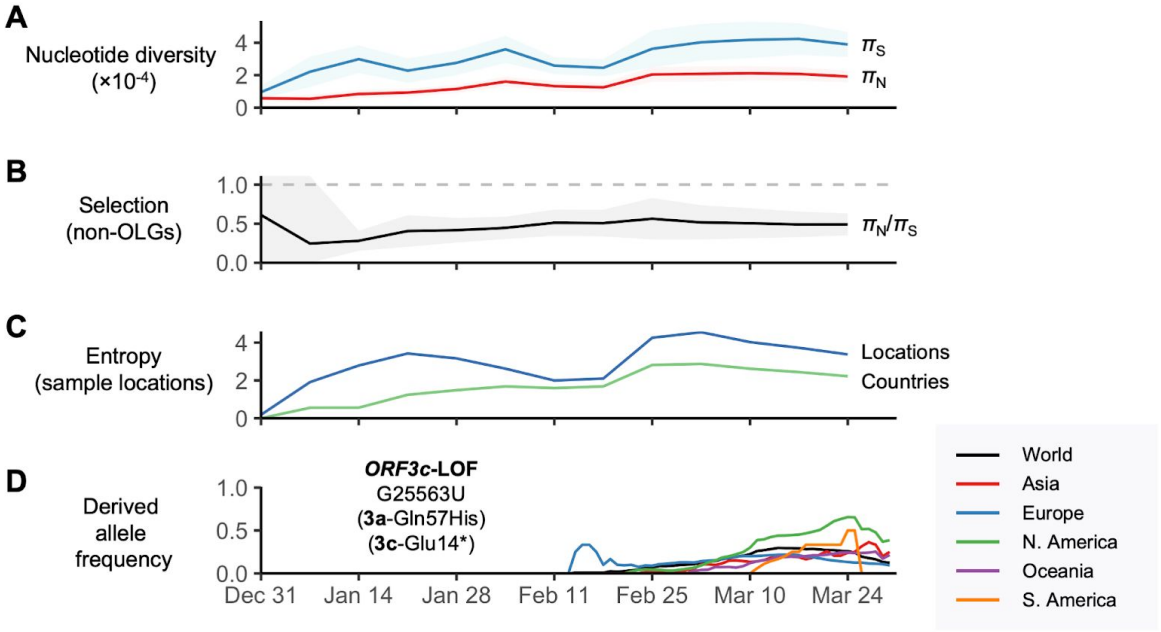
Supplementary Figure 1. Structural prediction for 3c protein. Independent computational modeling predictions of α-helices at the secondary (left inset, carried out in NetSurf v2) and tertiary structure levels (right inset, carried out in I-TASSER). Folding concordance with the closest protein structure match is shown (rightmost inset, aligned with TM-align). For explanation of shown metrics, see https://zhanglab.ccmb.med.umich.edu/I-TASSER. Chan et al. (2020) also predict a fold with α-helices (Raven Kok, pers. comm.).



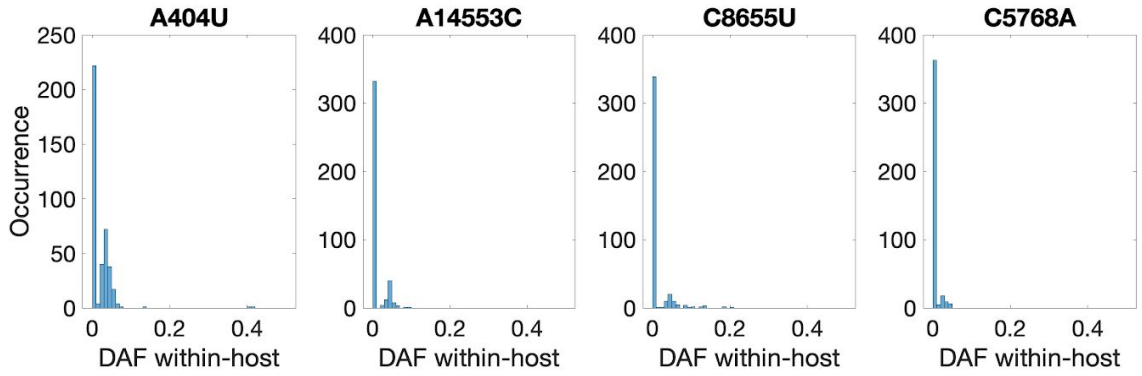
Supplementary Figure 2. Correlation between natural selection and protein expression. A weak negative Spearman correlation between the ratio of changes in amino acids to synonymous changes (d_N/d_S) and protein expression levels is observed across all three evolutionary levels: between species, between hosts, and within hosts. For each evolutionary level (panel), a given gene (color) has only one selection value (y axis) but two values of expression (x axis) from independent datasets (shape). Selection is calculated either as d_N/d_S or, for the overlapping gene *ORF9b*, in terms of the OLG-appropriate measure d_{NN}/d_{NS}.



Supplementary Figure 3. Between-species sliding window of genes overlapping *N*. Pairwise OLGenie analysis of the N gene across sarbecoviruses, in the ss13 reading frame. Each genome was compared with SARS-CoV-2 (left hand side) and SARS-CoV (right hand side plot). Methods as for Figure 7.



Supplementary Figure 4. SARS-CoV-2 between-host nucleotide diversity and allele frequencies as a function of time. Nonsynonymous (π_N) and synonymous (π_S) nucleotide diversity, π_N/π_S, diversity of sampling locations, and allele frequencies as a function of time for human SARS-CoV-2 (GISAID data). Results show sliding windows of 14 days (step size=7 days) representing 13 time points since the first GISAID sample was collected (EPI_ISL_402123 on 12/24/2019). Regions with overlapping genes were excluded (*ORF3a*/*ORF3c*, *N*/*ORF9b*, and *N*/*ORF9c*). Shaded regions show standard error of the mean (10,000 bootstrap replicates, codon unit). The horizontal dotted gray line denotes the π_N/π_S ratio expected under neutrality (1.0). Entropy in of sampling locations was defined as $-\sum p^*\ln(p)$, where p is the number of distinct (unique) locations or countries reported for a given window (Ewens and Grant 2001). Observed (sampled) allele frequency trajectories are colored by continents with sufficient sample sizes (Supplement).



Supplementary Figure 5. Recurrent nonsynonymous mutations within multiple human hosts. Histogram of the derived allele frequency of the four most common recurrent with-host protein-coding mutations across 401 samples (bin width=0.01). Mutation A404U introduces a premature stop codon in *nsp1*, whereas the remainder are nonsynonymous.