Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function

Alejandro Berrio[1], Valerie Gartner[1,2] and Gregory A. Wray[1,3]

[1] Department of Biology, Duke University, Durham, NC, USA
[2] University Program in Genetics and Genomics, Duke University, Durham, NC, USA
[3] Center for Genomic and Computational Biology, Duke University, Durham, NC, USA

ABSTRACT

Background: The emergence of a novel coronavirus (SARS-CoV-2) associated with severe acute respiratory disease (COVID-19) has prompted efforts to understand the genetic basis for its unique characteristics and its jump from non-primate hosts to humans. Tests for positive selection can identify apparently nonrandom patterns of mutation accumulation within genomes, highlighting regions where molecular function may have changed during the origin of a species. Several recent studies of the SARS-CoV-2 genome have identified signals of conservation and positive selection within the gene encoding Spike protein based on the ratio of synonymous to nonsynonymous substitution. Such tests cannot, however, detect changes in the function of RNA molecules.

Methods: Here we apply a test for branch-specific oversubstitution of mutations within narrow windows of the genome without reference to the genetic code.

Results: We recapitulate the finding that the gene encoding Spike protein has been a target of both purifying and positive selection. In addition, we find other likely targets of positive selection within the genome of SARS-CoV-2, specifically within the genes encoding Nsp4 and Nsp16. Homology-directed modeling indicates no change in either Nsp4 or Nsp16 protein structure relative to the most recent common ancestor. These SARS-CoV-2-specific mutations may affect molecular processes mediated by the positive or negative RNA molecules, including transcription, translation, RNA stability, and evasion of the host innate immune system. Our results highlight the importance of considering mutations in viral genomes not only from the perspective of their impact on protein structure, but also how they may impact other molecular processes critical to the viral life cycle.

Subjects Bioinformatics, Evolutionary Studies, Genomics, Virology
Keywords SARS-CoV-2, Evolution, Positive selection, COVID-19, Genome

INTRODUCTION

An important challenge in understanding zoonotic events is identifying the genetic changes that allow a pathogen to infect a new host. Such information can highlight molecular processes in both the pathogen and host that have practical value. The recent outbreak of SARS-CoV-2, a novel coronavirus, provides both a challenge and an opportunity to learn more about the specific adaptations that enable the virus to thrive in

Submitted 15 September 2020
Accepted 4 October 2020
Published 16 October 2020

Corresponding author
Alejandro Berrio,
alejo.berrio@duke.edu

Academic editor
Yuriy Orlov

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.10234**

How to cite this article Berrio A, Gartner V, Wray GA. 2020. Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. **PeerJ 8:e10234 DOI 10.7717/peerj.10234**

human hosts and that endow it with traits distinct from previously described coronaviruses (*Andersen et al., 2020*; *Morens et al., 2020*). Formal tests for natural selection are a powerful tool in this endeavor because they can be applied in an unbiased manner throughout the viral genome: evidence of negative selection can reveal regions of the genome that are broadly constrained functionally and thus unlikely to contribute to species-specific traits, while evidence of branch-specific positive selection can identify candidate regions of the genome where molecular processes may have diverged from that of other species.

Several recent studies have tested for natural selection in the SARS-CoV-2 genome based on the ratio of synonymous to non-synonymous (dN/dS) substitutions relative to other coronaviruses (*Tang et al., 2020*; *Chaw et al., 2020*; *Li et al., 2020a*). The most prominent signal to emerge from these studies is a mix of positive and purifying selection within the gene encoding the Spike glycoprotein, which mediates invasion of host cells by binding to the angiotensin-converting enzyme 2 (ACE2) receptor in host cells (*Gallagher & Buchmeier, 2001*; *Tortorici & Veesler, 2019*). This finding makes good biological sense, because structural changes in the spike protein are common and are known to influence the ability of the virus to infect new hosts and jump between species (*Hulswit, De Haan & Bosch, 2016*). A single nucleotide polymorphism (SNP) that results in an amino acid substitution in Spike protein (A > G at 23,403 bp; D614G) has increased in frequency during the global pandemic more rapidly than other SNPs (*Korber et al., 2020*), leading to speculation that it is an adaptation that alters the interaction between Spike and ACE2, FURIN and TMPRSS2 (*Eaaswarkhanth, Al Madhoun & Al-Mulla, 2020*).

Beyond mutations that alter Spike protein, however, there exists little understanding of positive selection within the SARS-CoV-2 genome and how this may have shaped viral traits. Few convincing signals of positive selection exist for any of the other viral proteins (*Cagliani et al., 2020*; *Velazquez-Salinas et al., 2020*; *Chaw et al., 2020*). For RNA viruses, however, critical aspects of the life cycle rely on molecular processes that are not reflected in protein sequence. In particular, in positive-strand RNA viruses such as coronaviruses, the single RNA molecule that constitutes the genome is first transcribed and translated to produce the replicase polyprotein 1a and 1ab that is cleaved into multiple non-structural proteins, some of which participate in the assembly of a cellular structure known as the replicase-transcriptase complex (RTC), where the proper environment for viral replication and transcription is created. Then, the RNA-dependent-RNA-polymerase (RdRp or Nsp12) produces negative sense genomic and subgenomic RNAs that are used as template strands that are then transcribed in the opposite direction to make more positive-sense viral genomes and a variety of RNA molecules that are translated into structural proteins for packaging (*Fehr & Perlman, 2015*; *Kim et al., 2020*). Although the viral proteins that help mediate these processes are visible to tests for selection that rely on dN/dS ratios, the RNA molecules with which they interact are not. This leaves the operation of natural selection on important molecular functions within the viral life cycle largely unexamined.

In order to test for positive selection on RNA function independent of its role in coding for amino acids, we utilized a test for positive selection, *adaptiPhy* (*Berrio, Haygood & Wray, 2020*), that identifies an excess of nucleotide substitutions within a defined window in the genome relative to neutral expectation using a likelihood ratio framework (*Wong & Nielsen, 2004*; *Haygood et al., 2007*). This test infers regions of the genome that were likely targets of branch-specific positive selection in several *Sarbecovirus* species from bat, pangolin, and human hosts. Our results recapitulate results from dN/dS-based tests that highlight *S*, the gene encoding Spike protein, as a prominent target of natural selection within the SARS-CoV-2 genome (*Cagliani et al., 2020*; *Chaw et al., 2020*; *Li et al., 2020a*). Importantly, we also identify genomic regions not previously reported to be targets of positive selection. Based on structural modeling of RNA and protein, we argue that these newly identified regions of positive selection may affect species-specific RNA, rather than protein, function. These genomic regions are candidates for understanding the molecular mechanisms that endow SARS-CoV-2 with some of its unique biological properties.

MATERIALS AND METHODS

Sequence alignment

To identify branch specific positive selection, it is necessary to obtain a query and a reference alignment. We downloaded six high quality reference genomes from the subgenus *Sarbecovirus* and an outgroup species (Table 1). Next, we used MAFFT (*Katoh & Standley, 2013*) plugin in Geneious Prime v.2.1 (*Kearse et al., 2012*) with default settings to build a sequence alignment. Next, we refined the alignment using a gene by gene procedure. More specifically, each coding sequence annotation (i.e., ORF1a, ORF1b, ORF3a, S, M, N, etc) is selected and realigned using the *Realign Region* tool implemented in Geneious Prime v.2.1 (*Kearse et al., 2012*) using the MAFFT (*Katoh & Standley, 2013*) option.

Testing for positive selection

Although *adaptiPhy* was originally designed to investigate regions of complex genomes under positive selection, it can be used to identify regions of a viral sequence alignment where the foreground branch is evolving at faster rates than the expectation from the background species. We performed a selection analysis on sliding windows of 300 bp with a step of 150 bp along a sequence alignment of five reference genome sequences of coronaviruses of the subgenus *Sarbecovirus* and two sequences of Pangolin Coronavirus recently published (*Liu, Chen & Chen, 2019*; *Lam et al., 2020*). This procedure generates partitions where a tree topology can be fitted. To investigate the extent of positive selection or branches with long substitution rates along the SARS-CoV-2 genome, we used a branch-specific method known as *adaptiPhy* that was initially developed in 2007 (*Haygood et al., 2007*) and recently improved (*Berrio, Haygood & Wray, 2020*). This computational methodology makes use of a likelihood ratio test based on the maximum likelihood estimates obtained from HyPhy v2.5 (*Pond, Frost & Muse, 2005*; *Pond et al., 2020*). The branch of interest (e.g., SARS-CoV-2 branch) is used as the foreground and the rest of the alignment is used as the background. To obtain data from nucleotide substitutions alone, we used *msa_split* from PHAST (*Hubisz, Pollard & Siepel, 2011*) to remove insertions and any sequence gaps that were present in the genomes of the background virus species relative to the SARS-CoV-2 genome. The assumption for the

Table 1 **Coronavirus accessions.**

Coronavirus species	Name used	NCBI Reference sequence
Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome	SARS-CoV-2	NC_045512.2
Bat coronavirus RaTG13, complete genome	Bat-CoV-RaTG13	MN996532.1
Pangolin coronavirus isolate MP789, complete genome	Pan-CoV-GD	MT121216.1
Pangolin coronavirus isolate PCoV_GX-P4L, complete genome	Pan-CoV-GX	MT040333.1
Rhinolophus affinis coronavirus isolate LYRa11, complete genome	Bat-CoV-LYRa11	KF569996.1
SARS coronavirus Tor2, complete genome NCBI Reference	SARS-CoV	NC_004718.3
Bat coronavirus BM48-31/BGR/2008, complete genome NCBI Reference	Bat-CoV-BM48	NC_014470.1

Note:
NCBI accessions of the Coronavirus sequences used in this study

background species is the same for both the null and alternative models; specifically, only neutral evolution and negative (purifying) selection are permitted. While in the foreground, the assumptions are the same as for the background in the null model. In the alternative model, all three types of evolution are permitted (neutral evolution, negative selection, and positive selection) in the foreground of the following topology: (((((SARS_CoV_2, Bat_CoV_RaTG13), Pa_CoV_Guangdong), Pa_CoV_Guangxi_P4L), (Bat_CoV_LYRa11, SARS_CoV)), Bat_CoV_BM48). This method is highly sensitive and specific and can differentiate between positive selection and relaxation of constraint (*Berrio, Haygood & Wray, 2020*). *AdaptiPhy* requires at least three kb reference alignment for each species that is used as a putatively neutral proxy for computing substitution rates. Viruses' genomes lack non-functional regions, therefore, the most reasonable proxy for neutral evolution has to be found in the regions outside the query window. To do this, we concatenated 20 regions of 300 bp of the viral genome alignment that were drawn randomly with replacement from the entire genome alignment. Then, for each query alignment, we built a reference alignment of six kb as it produces a stable evolutionary standard of substitution rates. To control for the stochasticity of the evolutionary process, we run each query against 10 bootstrapped samples of reference alignments. Finally, we used a custom R script to compute the likelihood ratio, which was used as a test statistic for a chi-squared test with one degree of freedom to calculate a *P*-value for each query. Then, we corrected the distribution of all *P*-values per query region using the *p.adjust()* R function with the fdr method. Next, we classified a query window to be under positive selection if the *P*-adjusted value was <0.05. We were unable to successfully run *adaptiPhy* on two windows because the outgroup species (Bat_CoV_BM48) contained a deletion of 406 bp relative to SARS-CoV-2, which spans the entire ORF8.

To visualize the strength of selection comprehensively, we computed the statistic ζ (zeta), representing the evolutionary rate. To calculate this rate, we compared the substitution rate in the query with their respective reference alignments. The distribution of substitution rates for each branch and nodes in each query and reference sequence was calculated using *phyloFit* (*Hubisz, Pollard & Siepel, 2011*). Then, the ratio of substitution rate in the query is divided by the substitution rate in the reference.

This parameter, “ζ”, is analogous to ω (omega), the ratio of dN/dS, where a value of ω < 1 indicates constraint or negative selection; a value of ω = 1 indicates neutrality; and a value of ω > 1 indicates positive selection (*Wong & Nielsen, 2004*).

Testing for conservation

To test for conservation, we used the *PhastCons* computational method from PHAST (*Siepel et al., 2005*; *Hubisz, Pollard & Siepel, 2011*). To run this tool, we used the models obtained with *phyloFit* for the reference alignments and then, we generated an average estimate of the conserved and non-conserved states of the models with *phyloBoot* (*Hubisz, Pollard & Siepel, 2011*). Finally, we run the final analysis using *PhastCons* on the query alignments using the previous models to generate *PhastCons* values for each base-pair along the sequence. To plot these we took the average from each alignment and plot it using the library Gviz and Bioconductor (*Hahne & Ivanek, 2016*) in R.

Testing for recombination

Inference of branch specific selection can be confounded by recombination given that a single phylogenetic tree may not explain the evolution of viruses. Recombination is common in coronaviruses (*Hon et al., 2008*; *Graham & Baric, 2010*; *Lau et al., 2015*; *Hu et al., 2017*; *Li et al., 2020b*; *Lam et al., 2020*) and it should be accounted for as an alternative explanation of selection at the nucleotide level. Here, we screened for evidence of recombination by estimating phylogenetic trees in sliding windows of 500 bp and a step of 150 along coronavirus alignment using RaXML-NG v0.9 (*Kozlov et al., 2019*).

Evaluating polymorphic diversity in the pandemics of 2020

We downloaded complete sequences of SARS-CoV-2 genomes from the NCBI Virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide). As of June 26, 2020, we obtained and aligned 5,597 SARS-CoV-2 genomes sequenced worldwide. To align these sequences, we used MAFFT (*Katoh & Standley, 2013*) plugin from Geneious Prime 2.1 (*Kearse et al., 2012*), eliminating 597 sequences with the highest number of differences and ambiguities relative to the reference sequence (RefSeq: NC_045512.2), for a total of 5,000 sequences. Next, we estimated the frequency of SNP variants using the Find Variations/SNPs tool with a minimum coverage of 4,900 sequences and a minimum frequency of 0.01, to identify nucleotide variants among a subset of high quality sequenced genomes in order to evaluate ongoing evolution in the regions under positive selection.

Analysis of RNA and protein structures

To investigate potential structural changes in Nsp4 and Nsp16 at both the RNA and protein level, we performed minimum free energy (MFE) prediction analysis using the RNAfold WebServer (*Gruber et al., 2008*; *Lorenz et al., 2011*) and consensus homology modeling using PHYRE2’s intensive mode (*Kelley et al., 2016*). These analyses were performed for both Nsp4 and Nsp16 sequences for SARS-CoV-2, Bat-CoV_RaTG13, Pan-CoV-Guangdong, and SARS-CoV.

RNAfold uses a loop-based energy model and a dynamic programing algorithm to predict the structure of the sequence such that the free energy of the structure is minimized. The RNAfold WebServer generates graphical outputs for both the MFE and Centroid structures, which display the base pairing probabilities by color (blue = 0, red =1). These two MFE structures correspond to the MFE and the Centroid traces in the mountain plot, which is a positional representation of the secondary structure. In our figures, we show the MFE structure prediction.

PHYRE2 aligns input protein sequences using Position-Specific Iterated BLAST against sequences of experimentally resolved protein structures. A 3D model of the input sequence is then constructed based on homology-matched templates, optimizing for greatest sequence coverage and highest confidence. Regions of the input sequence without a matching template sequence are modeled ab initio and with Poing, a multi-template modeling tool. Pairwise comparisons of predicted protein structures were visualized using PyMOL software (*DeLano, 2002*). Alignment and structural comparisons performed by FATCAT (*Ye & Godzik, 2004*).

RESULTS

Positive and negative selection are highly localized within coronavirus genomes

We tested for branch-specific selection on nucleotide sequences in coronavirus genomes, focusing on six species from the *Sarbecovirus* Subgenus (Coronaviridae family) and Bat-CoV-BM48-31/BGR/2008 as an outgroup. Using 300 bp windows with a step size of 150 bp, we scanned the genome alignment for concentrations of fixed mutations that exceed the neutral expectation based on the genome as whole relative to that particular window's evolutionary history among the seven species. This test identifies regions of the genome showing the most extreme divergence in nucleotide sequence on a particular branch relative to its specific background rate of evolution across the entire phylogeny and without reference to the genetic code. Figure 1A shows windows of inferred positive selection (red dots) on the branch leading to each species. The results reveal several signals of positive selection that are unique to a single species and others that are recapitulated in multiple species. The latter finding suggests that some segments of the viral genome have repeatedly experienced adaptive modification. In general, the distribution of positive selection is more similar in closely related species than in divergent ones (Figs. 1A and 1B), suggesting that some molecular functions have been altered over an interval that extends beyond the origin of a single species but not across the entire *Sarbecovirus* radiation.

Next, we identified regions of the genome that are highly conserved across the *Sarbecovirus* genomes examined in this study using *PhastCons* (*Siepel et al., 2005*) (Fig. 2A). As with positive selection, conservation is highly localized (Figs. 2A and 2B). Based on a criterion of *PhastCons* >0.9, we found high levels of conservation in regions encoding seven proteins: 3CL-Pro, Nsp6, Nsp8, Nsp9, Nsp10, Nsp11, RdRp, ORF3a (Protein 3a), Nucleocapsid phosphoprotein (NC), and Envelope (E) (Figs. 2A–2D).

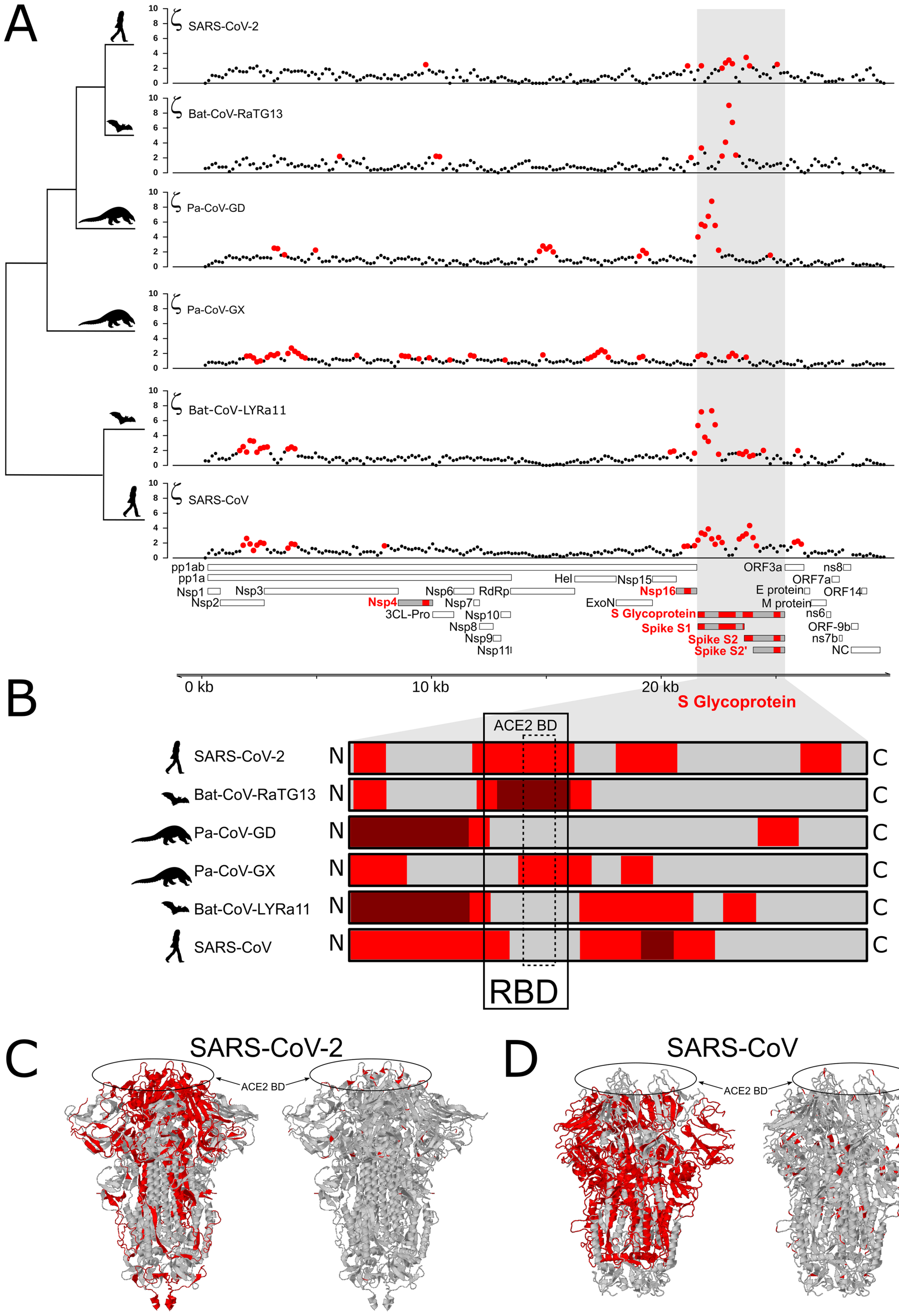

Figure 1 Distribution of evolutionary rate and positive selection across multiple species of coronaviruses of the *Sarbecovirus* subgenus. (A) Distribution of the evolutionary ratio, ζ, along multiple viral genome alignments. Red dots imply significant values of zeta from the *adaptiPhy* test, black dots represent neutral evolution or purifying selection in the foreground branch. (B) Visualization of selection within Spike protein among species. Dark red symbolizes windows where ζ is higher than 4, red is a significant ζ and and gray indicates neutral or purifying selection. RBD, receptor binding domain. Tertiary structure of Spike protein depicting the location positive selection and amino acid substitutions in (C) SARS-CoV-2 and (D) SARS-CoV. Full-size DOI: 10.7717/peerj.10234/fig-1

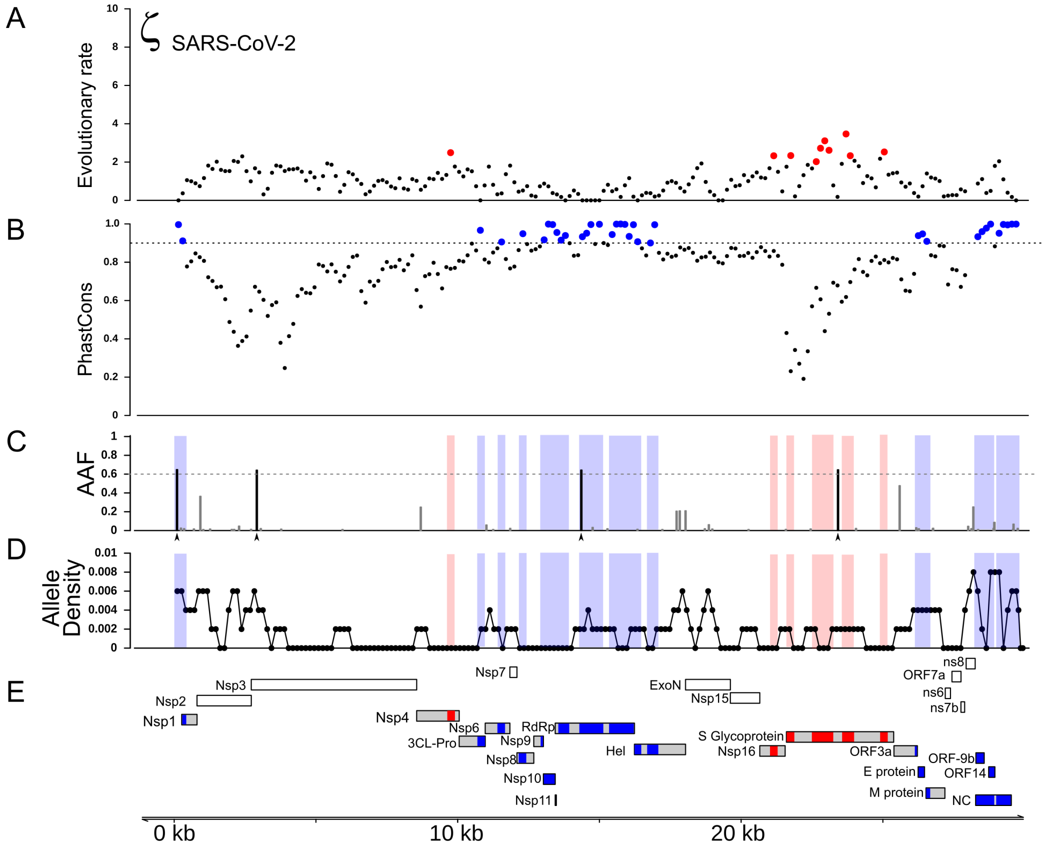

Figure 2 Distribution of positive selection, *PhastCons* conservation, and polymorphic variation across the SARS-CoV-2 genome. (A) Evolutionary rate (ζ) with sites under significant branch specific selection as red dots. (B) Panel depicting conservation values (*PhastCons*) with highly conserved windows (*PhastCons* >0.9) as blue dots over the dashed line along the SARS-CoV-2 genome. (C) Alternative allele frequency for 5,000 high quality genomes available in NCBI. Dotted line represents an arbitrary threshold of 0.6 and SNPs in strong linkage disequilibrium are highlighted with arrowheads under black bars. (D) Allele density in windows of 500 bp with a step of 150 bp. Red boxes in B and C symbolize regions under positive selection, while blue boxes represent high conservation. (E) Annotations for all the mature proteins known to be expressed in SARS-CoV-2. Full-size DOI: 10.7717/peerj.10234/fig-2

These loci of exceptional sequence conservation highlight critical molecular features: NC and E are essential structural proteins of the coronavirus capsid, while the other proteins regulate a variety of molecular process during viral replication (*Tan et al., 2005*; *Lu et al., 2006*; *Minakshi et al., 2009*; *Freundt et al., 2010*; *Fuchs, 2012*; *Yue et al., 2018*).

Because new mutations emerge and new strains replace older ones, we next investigated how much the specific strain used to represent SARS-CoV-2 influences test results. We re-ran the tests for positive selection using a strain of SARS-CoV-2 that contains four derived SNPs that commonly co-occur in currently circulating strains. Using this strain did not change the distribution of inferred regions of positive selection during the origin of SARS-CoV-2 (Fig. S1). We also generated two artificial genomes where we added four and nine mutations in the vicinity of site 14,408 to test the sensitivity of the test. We found that as zeta increased within the window, the test turned significant when more than five mutations are added (Fig. S1). It is important to note that the exact number of mutations that produce a significant test result may differ in other regions of the genome, depending on the degree of sequence conservation among species.

The gene encoding Spike protein is under persistent positive selection

In all ingroup species examined we detected signals of positive selection within the S gene, which encodes the Spike protein. With the exception of Pa-CoV-GX, this was the most prominent signal in the entire genome (Fig. 1A). This finding confirms previous studies that used the dN/dS ratio to test for selection on protein function (*Tang et al., 2020*; *Chaw et al., 2020*; *Li et al., 2020a*). Interestingly, we observed that the specific regions showing signatures of positive selection differed between species (Figs. 1B and 1C). In SARS-CoV-2, we detected signals of positive selection in four segments of the S gene. First, the region encoding the entire receptor binding domain (RBD) shows an extended signal (Figs. 1B, 1C and 3A); as others have noted, structural changes in this region may improve binding to human ACE2 (*Wang et al., 2020*; *Wrapp et al., 2020*; *Wang, Liu & Gao, 2020*). The second segment encodes another externally facing region, the S1 subunit N-terminal (NTD) domain, which includes the first disulfide bond (amino acids 13–113) and several glycosylation sites. The third signal of positive selection within S is located around the derived furin cleavage site (amino acids 664–812) that has been found to be essential for infection of lung cells (*Hoffmann, Kleine-Weber & Pöhlmann, 2020*). The fourth signal is located in a segment encoding the S2 and S2' subunits that includes the Heptad repeat 2 (amino acids 1114–1213). These heptad repeats were previously associated with episodes of selection for amino acids that increase the stability of the six-helix bundle formed by both heptad repeats in MERS and other coronaviruses (*Forni et al., 2015*); they are also thought to determine host expansions and therefore, facilitate virus cross-species transmission (*Graham & Baric, 2010*).

The distribution of inferred positive selection in the S gene of SARS-CoV differed from that of SARS-CoV-2 described above. Notably, there was no signal in the ACE2 binding domain (Figs. 1B and 1C). Moreover, a signal was present throughout the N-terminal domain and in the boundary region between the S1 and the S2 subunits (Fig. 1), a region that includes the proteolytic cleavage (*De Haan et al., 2004*). Interestingly, this region evolved a novel furin cleavage site in SARS-CoV-2 that may increase the cleavage efficiency and cell-cell fusion activity and changes in the virulence of the virion as seen in mutant studies of SARS-CoV and SARS-CoV-2 (*Follis, York & Nunberg, 2006*; *Hoffmann, Kleine-Weber & Pöhlmann, 2020*).

Genes encoding Nsp4 and Nsp16 contain branch-specific signals of positive selection

We also detected two shorter signals of positive selection within the SARS-CoV-2 genome that are located outside of the S gene, in pp1ab and pp1a (Fig. 1A). Interestingly, both encode small proteins that contribute to viral replication. The first is Nsp4, which encodes a membrane-bound protein with a cytoplasmic C-terminal domain; it is thought to anchor the Viral-Replication-Transcription Complex (RTC) to the modified endoplasmic

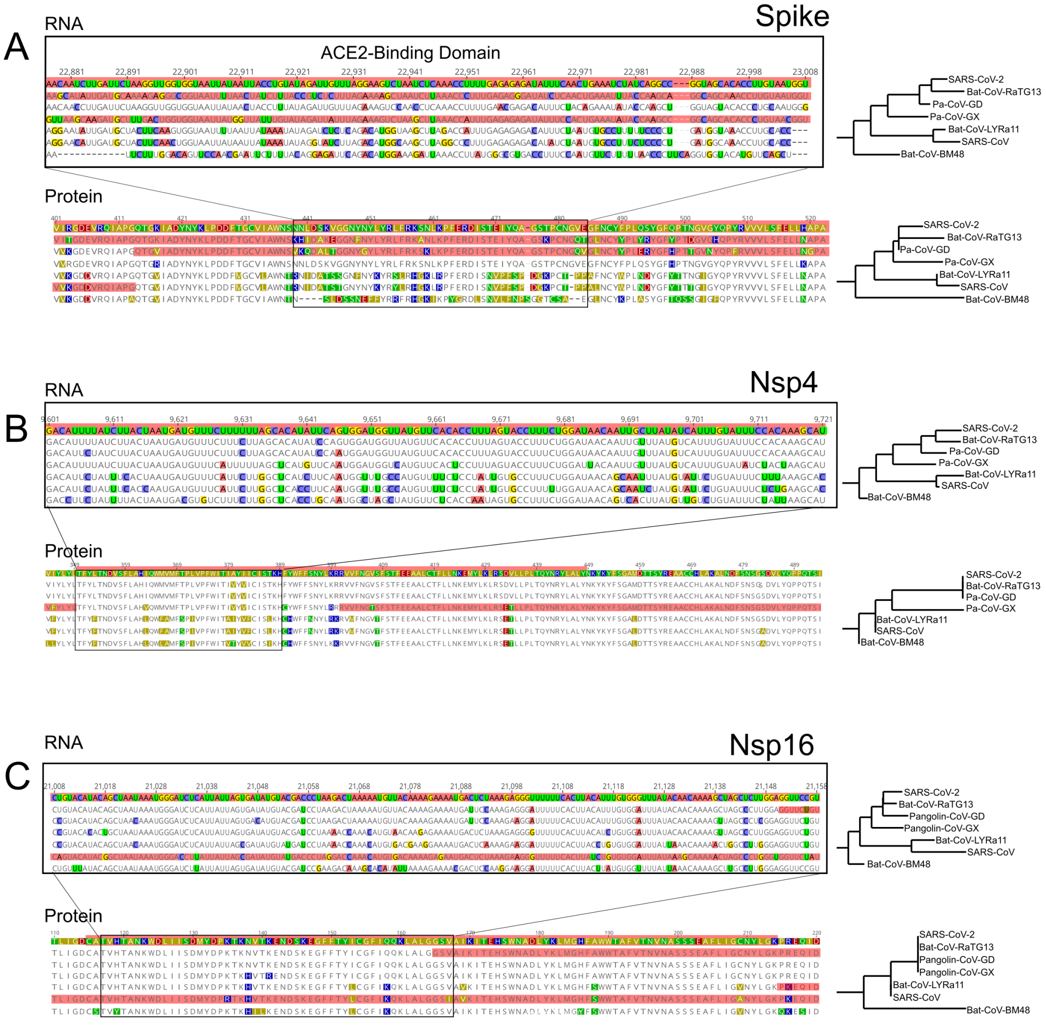

Figure 3 RNA and Protein sequences of Spike, Nsp4 and Nsp16. Each panel (A–C) shows selected RNA and protein sequences scoring high for positive selection in the SARS-CoV-2 branch and other branches (highlighted in red). Changes with respect to SARS-CoV-2 are highlighted in different colors.

Full-size DOI: 10.7717/peerj.10234/fig-3

reticulum membranes in the host cell (*Oostra et al., 2008*; *Hagemeijer et al., 2011*, *2014*; *Snijder, Decroly & Ziebuhr, 2016*). The SARS-CoV-2 Nsp4 protein differs from that of closely related sarbecoviruses by two nearly adjacent amino acids: V380A and V382I. Although this region of the genome as a whole is not highly conserved (Fig. 2), both of these positions are V residues in all of the in-group species we examined except SARS-CoV-2 (Fig. 3B; Fig. S2A). This signal is too weak to be scored as positive selection using dN/dS-based tests (*Tang et al., 2020*; *Chaw et al., 2020*; *Li et al., 2020a*) and indeed may not affect protein function given the biochemically similar side-chains of the amino acids involved.

The second signal of positive selection outside of the S gene lies within Nsp16. This gene encodes a 2′-O-methyltransferase that modifies the 5′-cap of viral mRNAs (*Decroly et al., 2008*; *Bouvet et al., 2010*) and assists in evasion of the innate immune system of host cells (*Züst et al., 2011*; *Menachery, Debbink & Baric, 2014*; *Nelemans & Kikkert, 2019*). Of note, this is the only signal of positive selection within the SARS-CoV-2 genome that lacks any nonsynonymous substitutions (Fig. 3C), and thus could not have been detected by any test that relies on the dN/dS ratio. All of the nucleotide substitutions in Nsp16 during the origin of SARS-CoV-2 are synonymous, while the Nsp16 genes of SARS-CoV-2, Bat-Cov-RaTG13, and Pan-CoV-GD (Guangdong) all encode identical proteins (Fig. 3C; Fig. S3A). This suggests a complex mechanism of selection in the form of purifying selection at the protein level and branch-specific positive selection at the nucleotide level. Ancestral state reconstruction of Nsp16 indicates that 20 synonymous substitutions likely occurred in the lineage leading to SARS-CoV-2 after the split from the common ancestor with BatCoV-RaTG13, while 19 substitutions are synonymous substitutions that occurred in the lineage leading to Bat-CoV-RaTG13 (Supplemental Data). Eleven of these twenty substitutions are concentrated within the region scoring high for positive selection in SARS-CoV-2 and twelve within the positively selected region in Bat-CoV-RaTG13.

As a consequence, we hypothesized that the Nsp16 RNA secondary structure may differ among species in ways that affect molecular functions mediated directly (although not solely) by RNA, such as replication, transcription, translation, or evasion of the host immune system. To investigate this possibility, we first compared the secondary structure and minimum free energy (MFE) of RNA in the vicinity of Nsp4 and Nsp16 among the genomes of SARS-CoV-2, Bat-CoV-RaTG13, Pan-CoV-GD, and SARS-CoV using RNAfold (*Gruber et al., 2008*). Both the predicted secondary structures and mountain plots, which show the free energy predictions along the length of the sequence by position, reveal differences in RNA folding dynamics across the four species (Figs. S2B and S3B). Analysis of the reconstructed sequence of the SARS-CoV-2 + Bat-CoV-RaTG13 ancestor reveal that most of these differences evolved during the origin of SARS-CoV-2 (Fig. S4). These differences among species in predicted secondary structures within Nsp4 and Nsp16 stand in contrast to the 5′ UTR, which is thought to fold into a stable secondary structure that is markedly conserved among *Sarbecovirus* species (Fig. S5). Though the accuracy of MFE predictions is too low to conclusively determine whether there are real between-species differences in the RNA structures of these loci (*Mathews, 2005*), these observations suggest that the signal of positive selection within Nsp16 in the SARS-CoV-2 genome may reflect changes in RNA, rather than protein, function that are unique to this species of coronavirus.

While the focus here is on SARS-CoV-2, it is worth noting that we also detected signals of positive selection outside of the S gene in the other *Sarbecovirus* genomes examined here. The distribution of positive selection in the genome of SARS-CoV, for instance, shows some similarities to, but also notable differences from, that of SARS-CoV-2 (Fig. 1). In both species, S and Nsp16 contain signals of positive selection, although in distinct regions of the two genes (Fig. 1). In addition, the genome of SARS-CoV contains signals of positive selection in Nsp2, Nsp3, and ORF3a, none of which shows elevated rates of substitution in SARS-CoV-2. The first two genes encode proteins with important roles in viral replication: Nsp2 may disrupt intracellular signaling in the host cell (*Cornillez-Ty* et al., 2009) while Nsp3 cleaves itself, Nsp1, and Nsp2 from the replicase polyproteins (Báez-Santos, St. John & Mesecar, 2015), assists in the assembly of the double membrane vesicles of the RTC system (Hagemeijer et al., 2014), and antagonizes the host innate immune response (Tsuchida, Kawai & Akira, 2009; Frieman et al., 2009; Matthews et al., 2014).

Recombination does not account for most signals of positive selection

Recombination from another species can be a confounding factor in the inference of positive selection using the framework employed here, because the inserted genomic segment may be more divergent than the rest of the foreground genome is from nearby background species. Several instances of recombination have been reported in coronaviruses, including SARS-CoV-2 (Hon et al., 2008; Lam et al., 2020; Boni et al., 2020; Li et al., 2020a), making it important to distinguish regions of recombination from positive selection. The two processes produce distinct genetic signatures, with recombination the result of a single event (possibly later further recombined) and positive selection as detected here the result of multiple independent mutations that were fixed over an extended interval and are spatially concentrated. In order to test for regions of the SARS-CoV-2 that contain recombined segments from other species, we estimated the phylogenetic history of 500 bp segments of the genome with a step size of 150 bp among the aligned genomes of the seven species examined in this study. We used RaXML-NG v0.9 (Kozlov et al., 2019) to reconstruct topology for each segment independently and searched for cases where the topology differed from the expected topology based on the entire genome: (Bat-CoV-BM48, (Bat-CoV-LYRa11, SARS-CoV), (Pa-CoV-GX, (Pa-CoV-GD, (SARS-CoV-2, Ba-CoV-RaTG13)))). Recombination from a divergent species should produce an incongruent topology in one or more adjacent windows, revealing a recombined region and its approximate breakpoints. We identified 12 regions where the topology differed from the expected (Fig. 4). Of note, these regions are somewhat more concentrated in the part of the genome that encodes structural proteins. Consistent with a previous report (Li et al., 2020a), we observed overlap between regions scoring high for positive selection and recombination in S, the gene encoding Spike protein (Fig. 4M), specifically the region that encodes for the ACE2 binding domain and a region that includes the furin-cleavage site (Figs. 4F and 4G). Importantly, however, none of the putatively recombined regions overlap with the windows scoring high for positive selection within the genes encoding Nsp4 and Nsp16 proteins in SARS-CoV-2.

Recent changes in allele frequency may result from positive selection and hitch-hiking

To gain insight into the evolutionary mechanisms that have shaped genetic variation more recently within the SARS-CoV-2 genome, we compiled a list of known mutations, based on 5,000 accessions sequenced since the beginning of the current pandemic (see Methods). As expected, the vast majority of variants are singletons, representing either mutations that are not segregating or sequencing errors. The density distribution of polymorphisms

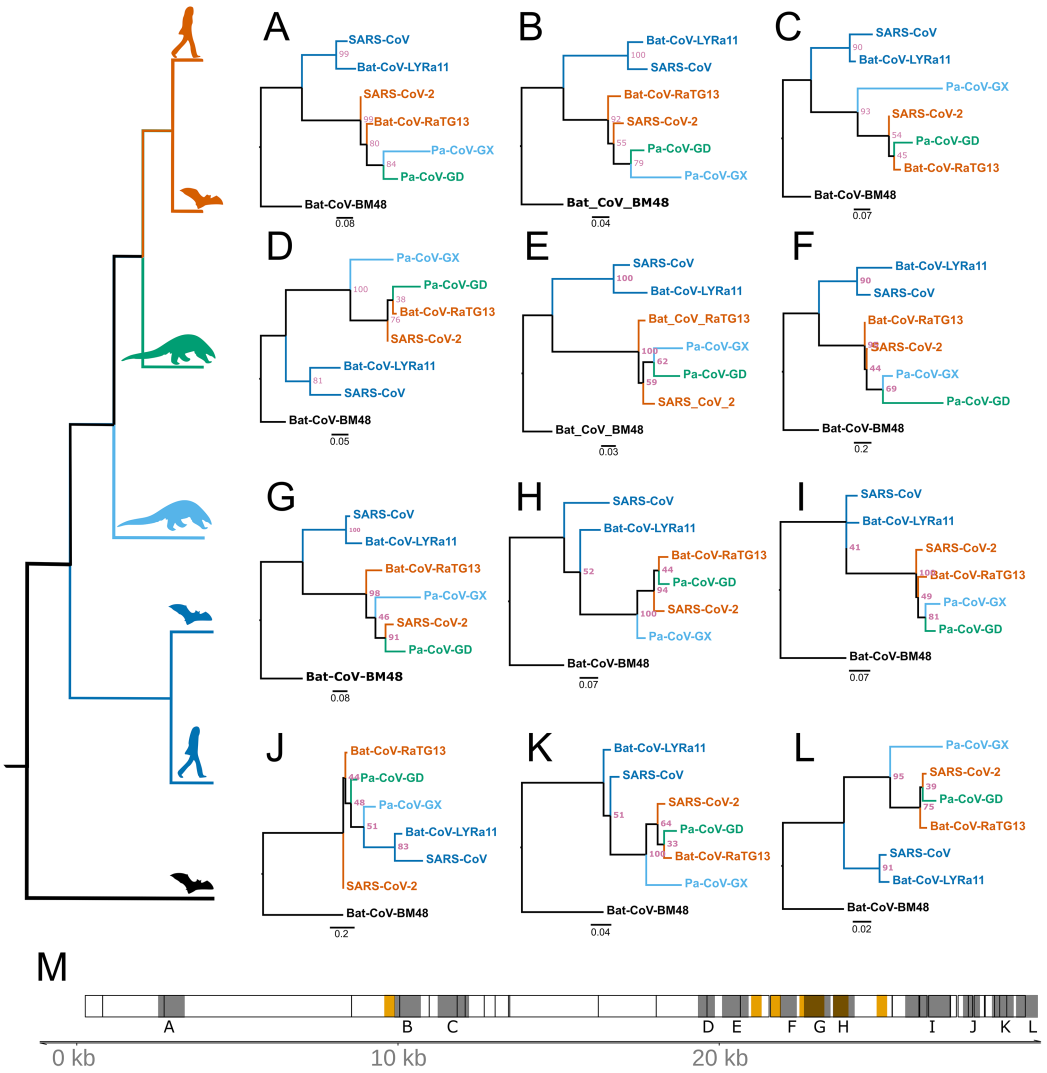

Figure 4 Regions of coronavirus genomes that violate the species tree. The species tree topology is shown on the left. (A–L) Tree topologies that were different from the expected topology. (M) Coronavirus genome track where the regions scoring high for positive selection in SARS-CoV-2 are highlighted in orange, regions with unexpected tree topologies highlighted in dark gray.

Full-size DOI: 10.7717/peerj.10234/fig-4

(regardless of frequency) is elevated within 2–3 kb at both ends the SARS-CoV-2 genome (Fig. 2D) and the site-frequency spectrum is strongly left-skewed (Fig. S6).

We next investigated the likely consequences for altered molecular function due to each of these four high-frequency derived SNPs. Two are located within regions of the genome that are highly conserved among *Sarbecovirus* species (Figs. 2B and 2C). The first is a C > U substitution at position 241 in the 5′UTR, a region of the genome where RNA secondary structure is highly conserved across Coronavirus species (*Madhugiri et al., 2016*; *Rangan et al., 2020*; *Alhatlani, 2020*). Using RNAfold (*Gruber et al., 2008*) we found that this C > U transition had no impact on the stem-loop structure established for SARS-CoV (Fig. S5). The other mutation in a conserved region of the genome is a nonsynonymous substitution in the RdRp gene (14,408; P323L) at the interface domain, which is though to mediate protein-protein interactions (*Pachetti et al., 2020*; *Hillen et al., 2020*). Because proline residues can influence secondary structure, we used PHYRE2 to predict the impact of the P232L mutation on protein structure. Comparison of the two predicted structures using FATCAT shows they are nearly identical (Table S1). The other two high-frequency derived SNPs are located in regions that are neither highly conserved nor highly divergent. One is a synonymous SNP in Nsp3 (3,037) and the other a nonsynonymous SNP in S (23,403; D614G). This last SNP effectively removes a charged side-chain between the receptor binding domain and the furin cleavage site of S, a region of recurrent positive selection among the *Sarbecovirus* species we examined. Thus, of the four high-frequency derived SNPs, the nonsynonymous substitution in S the most plausible candidate for altering molecular function and thus becoming a target of natural selection.

DISCUSSION

A crucial feature contributing to the global spread of COVID-19 is that viral shedding starts before the onset of symptoms (*He et al., 2020*); in contrast, shedding began 2–10 days after the onset of symptoms during the SARS epidemic of 2003 (*Peiris et al., 2003*; *Pitzer, Leung & Lipsitch, 2007*). This striking difference suggests that one or more molecular mechanisms during host cell invasion, virus replication, or immune avoidance may have changed during the origin of SARS-CoV-2. Mutations contributing to viral transmission would likely be favored by natural selection, making tests for positive selection a useful tool for identifying candidate genetic changes responsible for the unique properties of SARS-CoV-2. Here, we searched for regions of possible positive selection within the genomes of six coronavirus species, including SARS-CoV and SARS-CoV-2. The method we used tests for an excess of branch-specific nucleotide substitutions within a defined window relative to a neutral expectation for divergence in that window and without regard to the genetic code (*Wong & Nielsen, 2004*; *Haygood et al., 2007*; *Berrio, Haygood & Wray, 2020*).

Several prior studies have identified *S*, the gene encoding the Spike glycoprotein, as a target of recurrent positive selection in coronavirus genomes, including SARS-CoV-2, based on ω, the ratio of synonymous to nonsynonymous substitutions (*Andersen et al., 2020*; *Cagliani et al., 2020*; *Tang et al., 2020*; *Armijos-Jaramillo et al., 2020*; *Li et al., 2020a*). S thus serves as a positive control for our ability to detect signals of positive selection using a different approach, which considers mutations without respect to the genetic code and uses a likelihood ratio framework to identify regions of elevated, branch-specific nucleotide substitution rates relative to a model that allows only drift (*Wong & Nielsen, 2004*; *Haygood et al., 2007*; *Berrio, Haygood & Wray, 2020*). Consistent with this expectation, we found that portions of the gene encoding Spike showed a striking elevation of sequence divergence relative to the rest of the genome on the branches leading to all six species examined. The specific regions of S containing high divergence differs markedly, however, among species (Fig. 1B). In SARS-CoV and Bat-CoV-LYRa11, these regions include the N-terminal region, which contains glycosylation sites important for viral camouflage (*Watanabe et al., 2019*; *Yang et al., 2020*) and a site of proteolytic cleavage that allows entry into the host cell (*Belouzard, Chu & Whittaker, 2009*) (Figs. 1C and 3A). In contrast, signals of positive selection in SARS-CoV-2 and Bat-CoV-RaTG13 are concentrated in the domain that mediates binding to the host receptor ACE2 (Figs. 1C and 3A). These distinct distributions suggest that modifications in different aspects of Spike function took place as various coronaviruses adapted to novel hosts. In particular, the concentration of derived amino acid substitutions in the receptor binding domain of Spike (Figs. 1B and 1C) in SARS-CoV-2 and Bat-CoV-RaTG13 may reflect selection for amino acid substitutions that result in higher affinity for ACE2 protein in different host species.

Importantly, we also detected signals of positive selection in two additional regions of the SARS-CoV-2 genome, specifically within the genes encoding Nsp4 and Nsp16 (Figs. 1A and 2A). Of note, the Nsp16 region also shows a parallel signal of positive selection on the branch leading to SARS-CoV. To our knowledge, this is the first report of possible adaptive change in molecular function during the evolutionary origin of SARS-CoV-2 outside of the gene encoding Spike protein. Prior scans for positive selection within the SARS-CoV-2 genome used elevated ω as the signal of positive selection, which restricts attention to positive selection based on changes in protein function. For coronaviruses this is a notable limitation, given that many aspects of the lifecycle involve RNA function (*Madhugiri et al., 2016*; *Ziv et al., 2020*; *Alhatlani, 2020*). In addition, the secondary structure of some segments within the RNA genome is well conserved among coronavirus species, which implies a functional role (*Rangan et al., 2020*; *Sanders et al., 2020*; *Huston et al., 2020*). Indeed, the SARS-CoV-2 genome is reported to contain more well-structured regions than any other known virus, including both coding and noncoding regions of the genome (*Huston et al., 2020*). We therefore examined nucleotide substitutions within regions of putative positive selection in Nsp4 and Nsp16 for their likely impact on both protein and RNA structure (Figs. S2 and S3).

In the case of Nsp4 protein, two nearly adjacent nonsynonymous substitutions at residues 380 and 382 occurred on the branch leading to SARS-CoV-2 (Fig. 3B). These both involve changing side chains with similar biochemical properties, respectively valine to alanine and valine to isoleucine. Homology-directed modeling of protein structure suggests that these two amino acid substitutions have very little impact on either secondary or tertiary structure when comparing the SARS-CoV-2 protein orthologue to those of the other species examined (Fig. S2A). In the case of Nsp16 protein, no nonsynonymous substitutions evolved on the branch leading to SARS-CoV-2. Thus, the signal of positive selection within Nsp4 is unlikely to reflect changes in protein structure or function, while the signal within Nsp16 cannot affect either because the encoded polypeptide is identical (Fig. 3C; Fig. S3A).

With highly similar and identical protein structures predicted for Nsp4 and Nsp16, respectively, we considered the possibility that the signals of positive selection instead reflect changes in RNA structure and function. Previous studies found that neither the

Nsp4 nor Nsp16 regions stand out as particularly well folded regions of the genome, although Nsp16 does contain a single well-folded region and Nsp4 two moderately well folded regions (*Rangan et al., 2020*; *Huston et al., 2020*). Further, both genes show significantly decreased sequence divergence among coronavirus species within predicted double-stranded region (*Rangan et al., 2020*; *Sanders et al., 2020*; *Huston et al., 2020*). Indeed, the well-folded region within Nsp16 is the only such region in the SARS-CoV-2 genome that is also well conserved among related coronaviruses (*Sanders et al., 2020*). These published observations suggest possible functional roles for folded structures within Nsp4 and Nsp16. While we have not taken a robust experimental approach to determine between-species differences in RNA secondary structure, our in silico minimum free energy (MFE) predictions suggest that the likely secondary structure of the RNA genome in the region of the Nsp4 and Nsp16 genes may differ among the six coronavirus species we examined (Figs. S2B and S3B, top rows). The MFE predictions also indicate differences among species in entropy across the regions containing the signals of positive selection, indicating possible differences in the stability of the folded molecule (Figs. S2B and S3B, bottom rows). Together, these results indicate that the folded regions of Nsp4 and Nsp16 in the SARS-Cov-2 genome may differ in shape from those of related coronaviruses.

Unfortunately, little is currently known about the molecular functions of secondary structures in coronavirus genomes. Most of the attention has been directed towards the 5′ UTR, 3′ UTR, and frameshift element at the junction between ORF1a and ORF1b, which together contain the most well-folded regions in the SARS-CoV-2 genome (*Andrews et al., 2020*; *Sanders et al., 2020*; *Huston et al., 2020*). Thus, it is not possible at this time to link structural and thermodynamic features within Nsp4 and Nsp16 that are unique to SARS-CoV-2 to specific molecular functions. As discussed above, however, published evidence suggests that RNA secondary structures within these regions of the genome may be functional (*Rangan et al., 2020*; *Sanders et al., 2020*; *Huston et al., 2020*). These functions could, in principle, affect genome or transcript function, or both. Plausible possibilities include secondary structures that recruit specific RNA-binding proteins to mediate transcriptional regulation or transcript processing (*Pirakitikulr et al., 2016*; *Pan et al., 2020*), that mediate looping for other reasons (*Gebhard, Filomatori & Gamarnik, 2011*; *Ziv et al., 2020*), or that simply facilitate or impede processivity of the replication or translation machinery (*MacFadden et al., 2018*).

To investigate what evolutionary mechanisms are shaping the genetic variation at the population level, we examined known mutations among 5,000 accessions from NCBI. Given that the effective population size of SARS-CoV-2 is likely very large, the alternative allele distribution (Fig. S6) suggests that most SNPs are not subject to positive selection and that negative selection prevents most new mutations from rising in frequency due to drift, although this may change as additional whole genomes are examined. However, we did observe four SNPs that are present at high alternative allele frequency (AAF > 0.6) (Fig. 2C), a situation that can reflect positive selection, drift, or hitch-hiking. Interestingly, all four of these SNPs are in tight LD (*Toyoshima et al., 2020*), which suggests that positive selection on one of them may have driven the other three to high frequency due to hitch-hiking. Based on molecular modeling, the high-frequency derived mutation in S is the most plausible to be under positive selection, while the other three may be elevated due to hitch-hiking.

CONCLUSIONS

Scans for positive selection typically focus on changes in protein function and far less often consider the possibility of adaptive change in RNA function. By shining a light on regions of the SARS-CoV-2 genome that appear to be under positive selection yet are unlikely to alter protein function, our results illustrate the value of evaluating the potential for adaptive changes in secondary structures within the genomes of RNA viruses. In particular, we identify Nsp4 and Nsp16 as regions of the SARS-CoV-2 genome that may contain mutations that contribute to the unique biological and epidemiological features of this recently emerged pathogen.

While it is tempting to speculate about the possible adaptive role of changes in RNA structure within these accelerated regions, we suggest that this is best done in the context of relevant experimental results. For example, it might be informative to modify the primary sequence of the genome so as to encode the same protein sequence while altering or disrupting secondary structure within Nsp4 or Nsp16, then assay the consequences for viral replication and for specific molecular functions. We hope that our results inspire these or other experiments aimed at better understand the evolving functions of RNA secondary structure within the SARS-CoV-2 genome.

ACKNOWLEDGEMENTS

The authors would like to thank all the members of the lab groups of David McClay and Greg Wray for comments.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work did not have sources of direct funding but publishing costs were supported by the COPE fund at Duke University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Duke University.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Alejandro Berrio conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

- Valerie Gartner conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Gregory A. Wray conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data are available in the Supplemental Files.

The analytical pipelines and raw data are available in GitHub: https://github.com/wodanaz/adaptiPhy/blob/master/applications/.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10234#supplemental-information.

REFERENCES

Alhatlani BY. 2020. In silico identification of conserved cis-acting RNA elements in the SARS-CoV-2 genome. *Future Virology* **15(7)**:409–417 DOI 10.2217/fvl-2020-0163.

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nature Medicine* **26(4)**:450–452 DOI 10.1038/s41591-020-0820-9.

Andrews RJ, Peterson JM, Haniff HS, Chen J, Williams C, Grefe M, Disney MD, Moss WN. 2020. An in silico map of the SARS-CoV-2 RNA structurome. *BioRxiv* DOI 10.1101/2020.04.17.045161.

Armijos-Jaramillo V, Yeager J, Muslin C, Perez-Castillo Y. 2020. SARS-CoV-2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino acids necessary for complex stability. *Evolutionary Applications* **13(9)**:2168–2178 DOI 10.1111/eva.12980.

Báez-Santos YM, St. John SE, Mesecar AD. 2015. The SARS-coronavirus papain-like protease: structure, function and inhibition by designed antiviral compounds. *Antiviral Research* **115**:21–38 DOI 10.1016/j.antiviral.2014.12.015.

Belouzard S, Chu VC, Whittaker GR. 2009. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proceedings of the National Academy of Sciences of the United States of America* **106(14)**:5871–5876 DOI 10.1073/pnas.0809524106.

Berrio A, Haygood R, Wray GA. 2020. Identifying branch-specific positive selection throughout the regulatory genome using an appropriate proxy neutral. *BMC Genomics* **21(1)**:359 DOI 10.1186/s12864-020-6752-4.

Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, Rambaut A, Robertson DL. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Epub ahead of print 28 July 2014. *Nature Microbiology* DOI 10.1038/s41564-020-0771-4.

Bouvet M, Debarnot C, Imbert I, Selisko B, Snijder EJ. 2010. In vitro reconstitution of SARS-coronavirus mRNA cap methylation. *PLOS Pathogens* **6(4)**:1000863 DOI 10.1371/journal.ppat.1000863.

Cagliani R, Forni D, Clerici M, Sironi M. 2020. Computational inference of selection underlying the evolution of the novel coronavirus, severe acute respiratory syndrome coronavirus 2. *Journal of Virology* **94**:411–431 DOI 10.1128/JVI.00411-20.

Chaw S-M, Tai J-H, Chen S-L, Hsieh C-H, Chang S-Y, Yeh S-H, Yang W-S, Chen P-J, Wang H-Y. 2020. The origin and underlying driving forces of the SARS-CoV-2 outbreak. *Journal of Biomedical Science* **27(1)**:73 DOI 10.1186/s12929-020-00665-8.

Cornillez-Ty CT, Liao L, Yates Iii JR, Kuhn P, Buchmeier MJ. 2009. Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. *Journal of Virology* **83(19)**:10314–10318 DOI 10.1128/JVI.00842-09.

Decroly E, Imbert I, Coutard B, Bouvet M, Selisko B, Alvarez K, Gorbalenya AE, Snijder EJ, Canard B. 2008. Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (Nucleoside-2′O)-methyltransferase activity. *Journal of Virology* **82(16)**:8071–8084 DOI 10.1128/JVI.00407-08.

De Haan CAM, Stadler K, Godeke G-J, Jan Bosch B, Rottier PJM. 2004. Cleavage inhibition of the murine coronavirus spike protein by a furin-like enzyme affects cell-cell but not virus-cell fusion. *Journal of Virology* **78(11)**:6048–6054 DOI 10.1128/JVI.78.11.6048-6054.2004.

DeLano WL. 2002. Pymol: an open-source molecular graphics tool. *Protein Crystallography* **40**:82–92.

Eaaswarkhanth M, Al Madhoun A, Al-Mulla F. 2020. Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? *International Journal of Infectious Diseases* **96**:459–460 DOI 10.1016/j.ijid.2020.05.071.

Fehr AR, Perlman S. 2015. Coronaviruses: an overview of their replication and pathogenesis. In: Maier H, Bickerton E, Britton P, eds. *Coronaviruses: Methods and Protocols*. New York: Springer, 1–23.

Follis KE, York J, Nunberg JH. 2006. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* **350(2)**:358–369 DOI 10.1016/j.virol.2006.02.003.

Forni D, Filippi G, Cagliani R, De Gioia L, Pozzoli U, Al-Daghri N, Clerici M, Sironi M. 2015. The heptad repeat region is a major selection target in MERS-CoV and related coronaviruses. *Scientific Reports* **5(1)**:1–10 DOI 10.1038/srep14480.

Freundt EC, Yu L, Goldsmith CS, Welsh S, Cheng A, Yount B, Liu W, Frieman MB, Buchholz UJ, Screaton GR, Lippincott-Schwartz J, Zaki SR, Xu X-N, Baric RS, Subbarao K, Lenardo MJ. 2010. The open reading frame 3a protein of severe acute respiratory syndrome-associated coronavirus promotes membrane rearrangement and cell death. *Journal of Virology* **84(2)**:1097–1109 DOI 10.1128/JVI.01662-09.

Frieman M, Ratia K, Johnston RE, Mesecar AD, Baric RS. 2009. Severe acute respiratory syndrome coronavirus papain-like protease ubiquitin-like domain and catalytic domain regulate antagonism of IRF3 and NF-kappaB signaling. *Journal of Virology* **83(13)**:6689–6705 DOI 10.1128/JVI.02220-08.

Fuchs SY. 2012. Ubiquitination-mediated regulation of interferon responses. *Growth Factors* **30(3)**:141–148 DOI 10.3109/08977194.2012.669382.

Gallagher TM, Buchmeier MJ. 2001. Coronavirus spike proteins in viral entry and pathogenesis. *Virology* **279(2)**:371–374 DOI 10.1006/viro.2000.0757.

Gebhard LG, Filomatori CV, Gamarnik AV. 2011. Functional RNA elements in the dengue virus genome. *Viruses* **3(9)**:1739–1756 DOI 10.3390/v3091739.

Graham RL, Baric RS. 2010. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *Journal of Virology* **84(7)**:3134–3146 DOI 10.1128/JVI.01394-09.

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Research* **36**:70–74 DOI 10.1093/nar/gkn188.

Hagemeijer MC, Monastyrska I, Griffith J, Van der Sluijs P, Voortman J, Van Bergen en Henegouwen PM, Vonk AM, Rottier PJM, Reggiori F, De Haan CAM. 2014. Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. *Virology* **458–459**:125–135 DOI 10.1016/j.virol.2014.04.027.

Hagemeijer MC, Ulasli M, Vonk AM, Reggiori F, Rottier PJM, De Haan CAM. 2011. Mobility and interactions of coronavirus nonstructural protein 4. *Journal of Virology* **85(9)**:4572–4577 DOI 10.1128/JVI.00042-11.

Hahne F, Ivanek R. 2016. Visualizing genomic data using Gviz and bioconductor. In: Mathé E, Davis S, eds. *Methods in Molecular Biology*. New York: Humana Press Inc, 335–351.

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics* **39(9)**:1140–1144 DOI 10.1038/ng2104.

He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X, Mo X, Chen Y, Liao B, Chen W, Hu F, Zhang Q, Zhong M, Wu Y, Zhao L, Zhang F, Cowling BJ, Li F, Leung GM. 2020. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine* **26(5)**:672–675 DOI 10.1038/s41591-020-0869-5.

Hillen HS, Kokic G, Farnung L, Dienemann C, Tegunov D, Cramer P. 2020. Structure of replicating SARS-CoV-2 polymerase. *Nature* **584(7819)**:154–156 DOI 10.1038/s41586-020-2368-8.

Hoffmann M, Kleine-Weber H, Pöhlmann S. 2020. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular Cell* **78(4)**:779–784.e5 DOI 10.1016/j.molcel.2020.04.022.

Hon C-C, Lam T-Y, Shi Z-L, Drummond AJ, Yip C-W, Zeng F, Lam P-Y, Chi F, Leung C. 2008. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *Journal of Virology* **82(4)**:1819–1826 DOI 10.1128/JVI.01926-07.

Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, Xie J-Z, Shen X-R, Zhang Y-Z, Wang N, Luo D-S, Zheng X-S, Wang M-N, Daszak P, Wang L-F, Cui J, Shi Z-L. 2017. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLOS Pathogens* **13(11)**:e1006698 DOI 10.1371/journal.ppat.1006698.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* **12(1)**:41–51 DOI 10.1093/bib/bbq072.

Hulswit RJG, De Haan CAM, Bosch BJ. 2016. Coronavirus spike protein and tropism changes. *Advances in Virus Research* **96**:29–57 DOI 10.1016/bs.aivir.2016.08.004.

Huston NC, Wan H, De Araujo Tavares RC, Wilen C, Pyle AM. 2020. Comprehensive in-vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *BioRxiv* DOI 10.1101/2020.07.10.197079.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30(4)**:772–780 DOI 10.1093/molbev/mst010.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28(12)**:1647–1649 DOI 10.1093/bioinformatics/bts199.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2016. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10(6)**:845–858 DOI 10.1038/nprot.2015.053.

Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. *Cell* **181(4)**:914–921.e10 DOI 10.1016/j.cell.2020.04.011.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, De Silva TI, Sheffield COVID-19 Genomics Group, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182(4)**:812–827.e19 DOI 10.1016/j.cell.2020.06.043.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A, Wren J. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35(21)**:4453–4455 DOI 10.1093/bioinformatics/btz305.

Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, Tong Y-G, Shi Y-X, Ni X-B, Liao Y-S, Li W-J, Jiang B-G, Wei W, Yuan T-T, Zheng K, Cui X-M, Li J, Pei G-Q, Qiang X, Cheung WY-M, Li L-F, Sun F-F, Qin S, Huang J-C, Leung GM, Holmes EC, Hu Y-L, Guan Y, Cao W-C. 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583(7815)**:282–285 DOI 10.1038/s41586-020-2169-0.

Lau SKP, Feng Y, Chen H, Luk HKH, Yang W-H, Li KSM, Zhang Y-Z, Huang Y, Song Z-Z, Chow W-N, Fan RYY, Ahmed SS, Yeung HC, Lam CSF, Cai J-P, Wong SSY, Chan JFW, Yuen K-Y, Zhang H-L, Woo PCY. 2015. Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. *Journal of Virology* **89(20)**:10532–10547 DOI 10.1128/JVI.01048-15.

Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, Chen Y, Gnanakaran S, Korber B, Gao F. 2020a. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Science Advances* **6(27)**:eabb9153 DOI 10.1126/sciadv.abb9153.

Li Y, Yang X, Wang N, Wang H, Yin B, Yang X, Jiang W. 2020b. The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification. Epub ahead of print 24 March 2020. *Future Virology* DOI 10.2217/fvl-2020-0066.

Liu P, Chen W, Chen J-P. 2019. Viral metagenomics revealed sendai virus and coronavirus infection of Malayan Pangolins (Manis javanica). *Viruses* **11(11)**:979 DOI 10.3390/v11110979.

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA package 2.0. *Algorithms for Molecular Biology* **6(1)**:26 DOI 10.1186/1748-7188-6-26.

Lu W, Zheng BJ, Xu K, Schwarz W, Du L, Wong CKL, Chen J, Duan S, Deubel V, Sun B. 2006. Severe acute respiratory syndrome-associated coronavirus 3a protein forms an ion channel and modulates virus release. *Proceedings of the National Academy of Sciences of the United States of America* **103(33)**:12540–12545 DOI 10.1073/pnas.0605402103.

MacFadden A, Òdonoghue Z, Silva PAGC, Chapman EG, Olsthoorn RC, Sterken MG, Pijlman GP, Bredenbeek PJ, Kieft JS. 2018. Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs. *Nature Communications* **9(1)**:1–11 DOI 10.1038/s41467-017-02604-y.

Madhugiri R, Fricke M, Marz M, Ziebuhr J. 2016. Coronavirus cis-acting RNA elements. *Advances in Virus Research* **96**:127–163 DOI 10.1016/bs.aivir.2016.08.007.

Mathews DH. 2005. Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics* **21(10)**:2246–2253 DOI 10.1093/bioinformatics/bti349.

Matthews K, Schäfer A, Pham A, Frieman M. 2014. The SARS coronavirus papain like protease can inhibit IRF3 at a post activation step that requires deubiquitination activity. *Virology Journal* **11(1)**:209 DOI 10.1186/s12985-014-0209-9.

Menachery VD, Debbink K, Baric RS. 2014. Coronavirus non-structural protein 16: evasion, attenuation, and possible treatments. *Virus Research* **194**:191–199 DOI 10.1016/j.virusres.2014.09.009.

Minakshi R, Padhan K, Rani M, Khan N, Ahmad F, Jameel S. 2009. The SARS coronavirus 3a protein causes endoplasmic reticulum stress and induces ligand-independent downregulation of the type 1 interferon receptor. *PLOS ONE* **4(12)**:e8342 DOI 10.1371/journal.pone.0008342.

Morens DM, Breman JG, Calisher CH, Doherty PC, Hahn BH, Keusch GT, Kramer LD, LeDuc JW, Monath TP, Taubenberger JK. 2020. The origin of COVID-19 and why it matters. *American Journal of Tropical Medicine and Hygiene* **103(3)**:955–959 DOI 10.4269/ajtmh.20-0849.

Nelemans T, Kikkert M. 2019. Viral innate immune evasion and the pathogenesis of emerging RNA virus infections. *Viruses* **11(10)**:961 DOI 10.3390/v11100961.

Oostra M, Hagemeijer MC, Van Gent M, Bekker CPJ, Te Lintelo EG, Rottier PJM, De Haan CAM. 2008. Topology and membrane anchoring of the coronavirus replication complex: not all hydrophobic domains of nsp3 and nsp6 are membrane spanning. *Journal of Virology* **82(24)**:12392–12405 DOI 10.1128/JVI.01219-08.

Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, Zella D, Ippodrino R. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine* **18(1)**:1–9 DOI 10.1186/s12967-019-02189-8.

Pan J, Qian X, Lattmann S, El Sahili A, Yeo TH, Jia H, Cressey T, Ludeke B, Noton S, Kalocsay M, Fearns R, Lescar J. 2020. Structure of the human metapneumovirus polymerase phosphoprotein complex. *Nature* **577(7789)**:275–279 DOI 10.1038/s41586-019-1759-1.

Peiris JSM, Chu CM, Cheng VCC, Chan KS, Hung IFN, Poon LLM, Law KI, Tang BSF, Hon TYW, Chan CS, Chan KH, Ng JSC, Zheng BJ, Ng WL, Lai RWM, Guan Y, Yuen KY. 2003. Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: a prospective study. *Lancet* **361(9371)**:1767–1772 DOI 10.1016/S0140-6736(03)13412-5.

Pirakitikulr N, Kohlway A, Lindenbach BD, Pyle AM. 2016. The coding region of the HCV genome contains a network of regulatory RNA structures. *Molecular Cell* **62(1)**:111–120 DOI 10.1016/j.molcel.2016.01.024.

Pitzer VE, Leung GM, Lipsitch M. 2007. Estimating variability in the transmission of severe acute respiratory syndrome to household contacts in Hong Kong, China. *American Journal of Epidemiology* **166(3)**:355–363 DOI 10.1093/aje/kwm082.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21(5)**:676–679 DOI 10.1093/bioinformatics/bti079.

Pond SLK, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, Wisotsky S, Spielman SJ, Frost SDW, Muse SV. 2020. HyPhy 2.5: a customizable platform for evolutionary hypothesis testing using phylogenies. *Molecular Biology and Evolution* **37(1)**:295–299 DOI 10.1093/molbev/msz197.

Rangan R, Zheludev IN, Hagey RJ, Pham EA, Wayment-Steele HK, Glenn JS, Das R. 2020. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA* **26(8)**:937–959 DOI 10.1261/rna.076141.120.

Sanders W, Fritch EJ, Madden EA, Graham RL, Vincent HA, Heise MT, Baric RS, Moorman NJ. 2020. Comparative analysis of coronavirus genomic RNA structure reveals conservation in SARS-like coronaviruses. *BioRxiv* DOI 10.1101/2020.06.15.153197.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LDW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15(8)**:1034–1050 DOI 10.1101/gr.3715005.

Snijder EJ, Decroly E, Ziebuhr J. 2016. The nonstructural proteins directing coronavirus RNA synthesis and processing. *Advances in Virus Research* **96**:59–126 DOI 10.1016/bs.aivir.2016.08.008.

Tan Y-J, Tham P-Y, Chan DZL, Chou C-F, Shen S, Fielding BC, Tan THP, Lim SG, Hong W. 2005. The severe acute respiratory syndrome coronavirus 3a protein up-regulates expression of fibrinogen in lung epithelial cells. *Journal of Virology* **79(15)**:10083–10087 DOI 10.1128/JVI.79.15.10083-10087.2005.

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. *National Science Review* **7(6)**:1012–1023 DOI 10.1093/nsr/nwaa036.

Tortorici MA, Veesler D. 2019. Structural insights into coronavirus entry. *Advances in Virus Research* **105**:93–116 DOI 10.1016/bs.aivir.2019.08.002.

Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. 2020. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. Epub ahead of print 22 July 2020. *Journal of Human Genetics* 1–8 DOI 10.1038/s10038-020-0808-9.

Tsuchida T, Kawai T, Akira S. 2009. Inhibition of IRF3-dependent antiviral responses by cellular and viral proteins. *Cell Research* **19(1)**:3–4 DOI 10.1038/cr.2009.1.

Velazquez-Salinas L, Zarate S, Eberl S, Gladue DP, Novella I, Borca MV. 2020. Positive selection of ORF3a and ORF8 genes drives the evolution of SARS-CoV-2 during the 2020 COVID-19 pandemic. *BioRxiv* DOI 10.1101/2020.04.10.035964.

Wang Y, Liu M, Gao J. 2020. Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proceedings of the National Academy of Sciences of the United States of America* **117(25)**:13967–13974 DOI 10.1073/pnas.2008209117.

Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, Lu G, Qiao C, Hu Y, Yuen KY, Wang Q, Zhou H, Yan J, Qi J. 2020. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* **181(4)**:894–904.e9 DOI 10.1016/j.cell.2020.03.045.

Watanabe Y, Bowden TA, Wilson IA, Crispin M. 2019. Exploitation of glycosylation in enveloped virus pathobiology. *Biochimica et Biophysica Acta: General Subjects* **1863(10)**:1480–1497 DOI 10.1016/j.bbagen.2019.05.012.

Wong WSW, Nielsen R. 2004. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* **167(2)**:949–958 DOI 10.1534/genetics.102.010959.

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, Graham BS, McLellan JS. 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367(6483)**:1260–1263 DOI 10.1126/science.abb2507.

Yang T-J, Chang Y-C, Ko T-P, Draczkowski P, Chien Y-C, Chang Y-C, Wu K-P, Khoo K-H, Chang H-W, Hsu S-TD. 2020. Cryo-EM analysis of a feline coronavirus spike protein reveals a unique structure and camouflaging glycans. *Proceedings of the National Academy of Sciences of the United States of America* **117(3)**:1438–1446 DOI 10.1073/pnas.1908898117.

Ye Y, Godzik A. 2004. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research* **32**:W582–W585 DOI 10.1093/nar/gkh430.

Yue Y, Nabar NR, Shi CS, Kamenyeva O, Xiao X, Hwang IY, Wang M, Kehrl JH. 2018. SARS-coronavirus open reading frame-3a drives multimodal necrotic cell death. *Cell Death and Disease* **9(1)**:1–15 DOI 10.1038/s41419-017-0012-9.

Ziv O, Price J, Shalamova L, Kamenova T, Goodfellow I, Weber F, Miska EA. 2020. The short-and long-range RNA-RNA interactome of SARS-CoV-2 co-first authors. *BioRxiv* DOI 10.1101/2020.07.19.211110.

Züst R, Cervantes-Barragan L, Habjan M, Maier R, Neuman BW, Ziebuhr J, Szretter KJ, Baker SC, Barchet W, Diamond MS, Siddell SG, Ludewig B, Thiel V. 2011. Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. *Nature Immunology* **12(2)**:137–143 DOI 10.1038/ni.1979.