




# Variable routes to genomic and host adaptation among coronaviruses

Vincent Montoya<sup>1</sup>  | Angela McLaughlin<sup>1,2</sup>  | Gideon J. Mordecai<sup>3</sup>  |  
Rachel L. Miller<sup>1,2</sup>  | Jeffrey B. Joy<sup>1,2,3</sup> 

<sup>1</sup>British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

<sup>2</sup>Bioinformatics Programme, University of British Columbia, Vancouver, BC, Canada

<sup>3</sup>Department of Medicine, University of British Columbia, Vancouver, BC, Canada

## Correspondence

Vincent Montoya and Jeffrey B. Joy, British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada.

Emails: vmontoya@cfenet.ubc.ca; jjoy@cfenet.ubc.ca

## Funding information

Canadian Institutes of Health Research, Grant/Award Number: 440371

## Abstract

Natural selection operating on the genomes of viral pathogens in different host species strongly contributes to adaptation facilitating host colonization. Here, we analyse, quantify and compare viral adaptation in genomic sequence data derived from seven zoonotic events in the *Coronaviridae* family among primary, intermediate and human hosts. Rates of nonsynonymous ( $d_N$ ) and synonymous ( $d_S$ ) changes on specific amino acid positions were quantified for each open reading frame (ORF). Purifying selection accounted for 77% of all sites under selection. Diversifying selection was most frequently observed in viruses infecting the primary hosts of each virus and predominantly occurred in the *orf1ab* genomic region. Within all four intermediate hosts, diversifying selection on the *spike* gene was observed either solitarily or in combination with *orf1ab* and other genes. Consistent with previous evidence, pervasive diversifying selection on coronavirus *spike* genes corroborates the role this protein plays in host cellular entry, adaptation to new hosts and evasion of host cellular immune responses. Structural modelling of spike proteins identified a significantly higher proportion of sites for SARS-CoV-2 under positive selection in close proximity to sites of glycosylation relative to the other coronaviruses. Among human coronaviruses, there was a significant inverse correlation between the number of sites under positive selection and the estimated years since the virus was introduced into the human population. Abundant diversifying selection observed in SARS-CoV-2 suggests the virus remains in the adaptive phase of the host switch, typical of recent host switches. A mechanistic understanding of where, when and how genomic adaptation occurs in coronaviruses following a host shift is crucial for vaccine design, public health responses and predicting future pandemics.

## KEYWORDS

coronaviruses, genomics, molecular evolution, role of selection in host switches, spike protein, viral adaptation, zoonoses

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Journal of Evolutionary Biology* published by John Wiley & Sons Ltd on behalf of European Society for Evolutionary Biology.

## 1 | INTRODUCTION

A mechanistic understanding of how adaptation occurs following a shift from one host to another is crucial for understanding macroevolution, speciation, adaptation and the emergence of infectious diseases (Joy & Crespi, 2012; Longdon et al., 2014; Nosil et al., 2002; Sanchez-Flores et al., 2016; Streicker et al., 2012). Host shifts and subsequent adaptation to the new host are particularly important in viral evolution (Forni et al., 2016; Mollentze et al., 2014) and in the emergence of viral infectious diseases in humans (Longdon et al., 2014). The generally small genome sizes and high mutation rates exhibited by RNA viruses make viral host-switching events excellent models for investigating the genomics of adaptation (Duffy et al., 2007; Forni et al., 2016; Zhao et al., 2019). These properties of RNA viruses also quickly lead to the accretion of significant genetic variation (Drummond et al., 2003). The large pool of genetic variation combined with the generally large population sizes of RNA viruses provide ideal conditions for selection to optimize viral phenotypes (Krakauer & Komarova, 2003). For a viral host switch to take place, there are numerous barriers to be overcome, including adequate ecological contact between the two host species for viral exchange to occur, structural barriers, sufficient host genetic similarity to enable initial infection and overcoming host immune defences (Olival et al., 2017; Parrish et al., 2008). Following a host-switching event, there may either be repeated contact with the new host and viral transfer but no subsequent ongoing transmission (e.g. West Nile Virus infection in humans), limited transmission in novel host species before viral extinction (e.g. Ebolavirus in humans) or sustained transmission among the new host leading to endemic or epidemic disease in the novel host (e.g. HIV, hepatitis C virus, SARS-CoV-2) (Andersen et al., 2020; Boni et al., 2020; Worobey et al., 2004). The amount of viral adaptation necessary for a host switch to result in an endemic disease in the novel host is thought to be dependent upon the extent of genetic overlap between the hosts as well as the frequency of contact between them (Olival et al., 2017). Viral transfer between phylogenetically proximate hosts (i.e. between mammals) is thus expected to be associated with a shallower valley between fitness peaks in the fitness landscape (Gavrilets, 2004), with fewer mutations required for colonization and adaptation to the novel host (Zhao et al., 2019). In contrast, viral transfer between phylogenetically distant hosts is expected to be associated with a deep fitness valley requiring more extensive host-specific adaptation to facilitate colonization of the new host (Olival et al., 2017).

Due to habitat destruction, there is a higher propensity for ecological overlap between displaced animals and humans leading to increased opportunities for viral host switches and emergence of novel viruses (Leao et al., 2020; Wertheim et al., 2013). Prior host jumps of coronaviruses to humans have involved an intermediate host (Chan & Chan, 2013). For example, emergence of Middle East respiratory syndrome coronavirus (MERS-CoV) into humans involved viral exchange between bats and camels before transmission to humans was possible. The role of selection in the intermediate host in facilitating adaptation to humans has been poorly documented; however, MacLean et al. (2020) found evidence that the majority of selection of SARS-CoV-2 occurred in bats not humans.

In this study, we test hypotheses concerning adaptation following a host shift across all known coronaviruses that have switched into humans and have available genomic sequence data. These include endemic human coronaviruses (HCoV-HKU1, HCoV-OC43, HCoV-229E, HCoV-NL63), which commonly infect people globally, and are usually associated with mild respiratory disease. Additionally, we include the three more recently introduced coronaviruses (SARS-CoV, MERS-CoV and SARS-CoV-2).

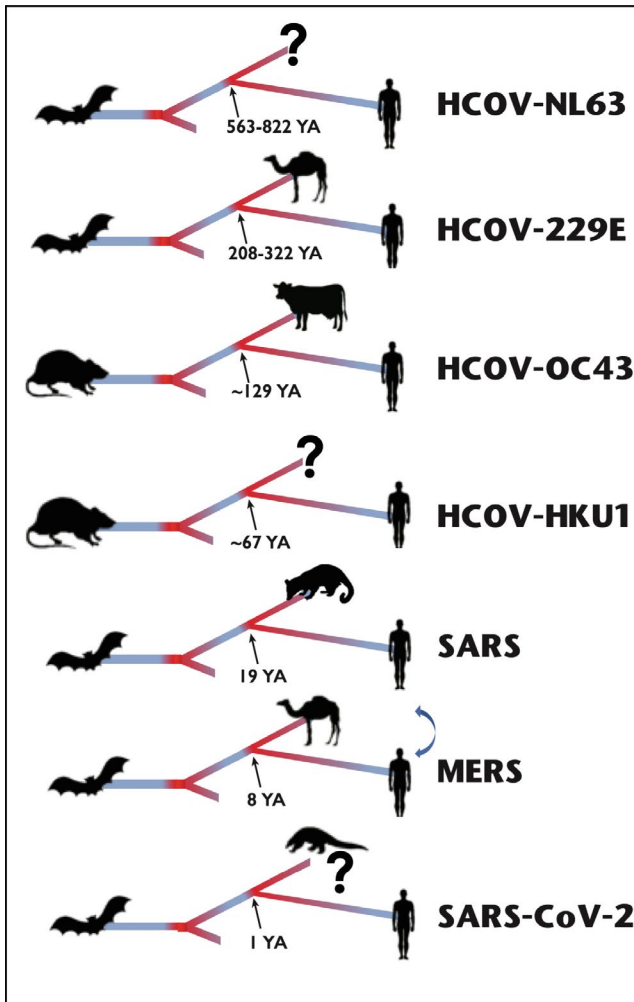
It is thought that the dominant selective regime on viral genes is purifying selection, which maintains gene and protein functionality (Spielman et al., 2019). This is supported by the observation that in the majority of viruses, the number of synonymous substitutions ( $d_s$ ) exceeds the number of nonsynonymous substitutions ( $d_n$ ). However, this observation is expected to be reversed on regions of a gene that are under positive selection ( $d_n > d_s$ ), such as an epitope in a viral protein (Hughes & Hughes, 2007). For example, previous research on coronaviruses has typically identified extensive adaptive evolution in the spike protein (Berrio et al., 2020; Tagliamonte et al., 2020; Xu et al., 2020). However, most investigations of coronavirus evolution and adaptation focus only on one or a limited number of genes and in the context of only a single virus–host system. Previous studies of selection in MERS-CoV have identified a large number of sites undergoing positive selection in genes involved in viral replication (Forni et al., 2016). Nevertheless, a pan-genomic evolutionary analysis is required in order to assess which sites are undergoing adaptation across coronaviruses in order to help identify possible trends that may exist in previous host-switching events. These findings are valuable for vaccine design, managing public health responses and predicting future zoonotic events.

Specifically, using molecular evolutionary analyses of historical selection we test the hypothesis that particular features of coronavirus genomes are imperative for transmission to novel hosts. We hypothesize that selection on coronavirus genes will vary with time depending upon the gene function. Specifically, we predict positive selection on the genomic regions associated with adaptation to the novel host immediately following the host shift followed by a transition to purifying selection (Figure 1). We also predict that genes involved in host adaptation and immune evasion will reveal a correlation between positively selected sites and time since host switching. We further expect these relationships to be consistent across human coronaviruses. Secondly, we evaluate the role that viral selection in intermediate hosts plays in facilitating persistence in humans. Finally, we evaluate associations between the number of positively selected sites since viral establishment and time since emergence in each host type.

## 2 | METHODS

### 2.1 | Data sets

Genome sequences for all SARS-CoV-2 isolates were downloaded from GISAID on 13 June 2020. All other SARS-CoV, MERS-CoV, HCoV-HKU1, HCoV-NL63, HCoV-229E and HCoV-OC43



**FIGURE 1** Mock-phylogenetic tree depicting host-switching events for the coronaviruses in this study. Each virus evolved from within its presumed primary, natural host (bats or rats), subsequently spilled over into an intermediate host, and finally jumped into humans. The estimated time since introduction into the human population is shown at the node preceding each human viral lineage (years ago: YA). However, for HCOV-NL63 and HCOV-229E, these timeframes were estimated from the emergence from bats. At each of these instances of host switching, presumably there was a large increase in adaptive evolution (shown in red on the phylogenetic pathway from primary to human host) followed by purifying selection (shown in blue). Blue arrow signifies documented ongoing cross-species transmission events as exemplified in MERS-CoV

genomes were obtained from NCBI Genbank on 28 May 2020. Viral isolates from bats immediately preceding the SARS-CoV-2 lineage were obtained from Boni et al. (2020). Candidate genes encoding the spike protein from each respective coronavirus species were used to identify additional relatives using BLASTn (Chen et al., 2015) alignments against the nonredundant database with relaxed stringencies. Coordinates for each gene were obtained from each respective Genbank reference sequence, and these identified regions were subsequently used in alignments to retrieve genes for those sequences

without precise gene coordinates. To avoid artefacts associated with in vitro evolution, sequences were removed if they were sequenced from a cultured isolate as described in the respective Genbank record. We also removed sequences if they were derived from a recombinant virus, had greater than 5% unresolved nucleotides ("N") within a gene, or coverage for a particular gene was <50%.

## 2.2 | Phylogenetic methods

Multiple sequence alignments were performed for each species/gene using MAFFT version 7.464, (Kato et al., 2009) and the resulting alignments were visually inspected using AliView version 1.26 (Larsson, 2014). Subsequently, distributions of 10 phylogenetic trees were inferred for each alignment under an approximate maximum likelihood modelling framework employing a general time reversible model of molecular evolution from the consensus sequences as implemented in FastTree version 2.1.10 (Price et al., 2010).

## 2.3 | Molecular evolutionary analyses

All tests for selection were performed using Hyphy version 2.5.8 (Pond, Poon, et al., 2020). Prior to each analysis, recombination events among the sequences in all alignments were detected using GARD (Kosakovsky Pond et al., 2006). Due to computational issues with the GARD analysis for the *orf1ab* and *spike* genes for SARS-CoV-2, the data sets for these genes were down-sampled into 20 groups by randomly selecting unique sequences and performing GARD on each of these groups. This was performed 10 times and the consensus of these analyses were used for all subsequent results. For all other GARD analyses, if any breakpoints were identified, all subsequent analyses were performed on each respective partitioned alignment. To identify positions in the genome under selection for each of the coronavirus species, two site-specific models were used [fixed effects likelihood, FEL (Pond & Frost, 2005) and mixed effects model of episodic selection, MEME (Murrell et al., 2012)]. Differences in selective regimes among human and nonhuman hosts were assessed with an additional site-level model (CONTRAST-FEL) (Pond, Wisotsky, et al., 2020). To account for uncertainty in phylogenetic tree topologies, (Parker et al., 2008) these tests were performed across distributions of ten phylogenetic trees reconstructed for each gene. The default thresholds for each of the resulting *p*-values (www.hyphy.org) were used to determine significance, and the consensus of these results was used to identify the type of selection. In all comparisons of selection we controlled for the difference in number of sequences analyzed by dividing the number of positively selected sites by the  $\log_{10}$  of the sequence counts for each respective virus/gene group. Positions under selection were required to be identified within internal branches for both FEL and MEME. In order to compare the results from CONTRAST-FEL (multi-host derived viral sequences) and FEL/MEME (single host derived viral sequences), the precise codon for each analysis was identified

**TABLE 1** Number of sequences for each coronavirus species

	Human	Bat	Bovine	Camel	Civet	Pangolin
HCOV-229E	191	15		33		
HCOV-HKU1	339					
HCOV-NL63	263	6				
HCOV-OC43	745		40			
MERS-CoV	279	47		264		
SARS-CoV	101	59			46	
SARS-CoV-2	45,721	10				9

by aligning each codon-aware alignment to each other. The estimated time of zoonoses or host-switching events were calculated as the difference between the estimated years since emergence into humans and the most recent collection date for each coronavirus sequence obtained from a clinical isolate. For these results, the total FEL/MEME sites were used, respectively, in each analysis. The estimates were derived from the following studies: HCOV-OC43 (Vijgen et al., 2005), HCOV-HKU1 (Al-Khannaq et al., 2016), HCOV-NL63 (Huynh et al., 2012), HCOV-229E (Pfefferle et al., 2009), SARS-CoV (Lau et al., 2015), MERS-CoV (Zhang et al., 2016) and SARS-CoV-2 (Boni et al., 2020). [Correction added on 27 March 2021, after first online publication: This section has been modified.]

### 2.4 | Protein structure modelling

Reference sequences from Genbank for each respective virus (SARS-CoV-2: NC\_045512.2, HCOV-OC43: NC\_006213.1, HCOV-NL63: NC\_005831.2 SARS-CoV: NC\_004718.3, MERS-CoV: NC\_019843.3, HCOV-229E: NC\_002645.1) were used to model each respective Spike protein using SWISS-MODEL (Waterhouse et al., 2018). The amino acid mutations were mapped onto the respective 3-dimensional structures representative of each species from the RCSB Protein Data Bank (Berman et al., 2000), accession numbers: SARS-CoV-2: 6ZGH, HCOV-OC43: 6NZK, HCOV-NL63: 5SZS, SARS-CoV-1: 6ACD, MERS: 5W9H, HCOV-229E: 6U7H. The Spike protein structures were analysed, annotated and visualized using the PyMOL molecular graphic system, version 2.4.0 (Schrodinger, LLC, 2015). As there was no suitable template for HCOV-HKU1, HCOV-OC43 was used as a surrogate for visual purposes and excluded from further structural analyses. Glycosylation sites were modelled using GlyProt, (Bohne-Lang & Lieth, 2005) and sites in close proximity (10 angstroms) were identified using PyMOL (Schrodinger, LLC, 2015). Structure-based alignments were generated using the TM-Align algorithm (Zhang & Skolnick, 2005). Aligned amino acid positions for each pairwise protein structure alignment found to be in close proximity (5 angstroms, as indicated in the output of the TM-alignment) were then compared with positions that were positively selected. Neutralizing epitopes were obtained from the Immune Epitope Database (Vita et al., 2019). Monte Carlo permutation tests (999 unrestricted permutations,  $p \leq 0.05$ ) were used to test the significance of glycosylation proximity to sites of selection using the Coin package in R (Zeileis et al., 2008).

## 3 | RESULTS

### 3.1 | Data set

A total of 50,354 sequences were obtained from global databases before quality filtering (Table S3, <https://github.com/vmon5813/CoronavirusHostAdaptations>). After removing poor quality sequences, sequences with inadequate coverage, cultured isolates and recombinants a total of 48,168 sequences remained (Table 1, Methods). Due to the extraordinary efforts of researchers around the world, 95% of all genomic sequences in this study belong to SARS-CoV-2. Unsurprisingly, except for SARS-CoV and MERS-CoV, the majority of sequences in each data set were dominated by human sequences (median = 87.4%), whereas the number of sequences from primary and intermediate hosts were typically less well-represented (medians 6.3% and 18.1%, respectively).

### 3.2 | Molecular evolutionary analyses

Rates of nonsynonymous ( $d_N$ ) and synonymous ( $d_S$ ) mutations on specific amino acid positions were quantified for each ORF from seven coronaviruses known to have a history of host-switching events: SARS-CoV-2, SARS-CoV, MERS-CoV, HCOV-229E, HCOV-OC43 and HCOV-NL63 as well as at least one of their nonhuman hosts. HCOV-HKU1 was also investigated; however, all related viruses derived from nonhuman hosts were too divergent. HCOV-OC43-like viruses isolated from rats were also not investigated, as only four suitable genomes were available to the best of our knowledge and most of these ORFs collapsed into identical sequences. Viruses from primary (bats) and intermediate (camels, civets and bovine) hosts were each contrasted with their human-derived counterparts. Selection was measured using sites identified by both a fixed effects likelihood (FEL) analysis that assumes constant selective pressure across the phylogeny and a mixed effects model of episodic selection (MEME) that permits varying positive selective pressure.

Several patterns were observed upon a per-site evolutionary investigation across the genomes of all coronaviruses (Figure 2). Primarily, negative or purifying selection was dominant across the majority of genomes analysed as observed in several other similar analyses (Forni et al., 2016; Pond, 2020; Tang et al., 2009). The majority of selection was revealed to be acting on the primary bat hosts

of each virus, which is not unexpected considering the extraordinary genetic diversity of these viruses circulating in bats (Figure 3).

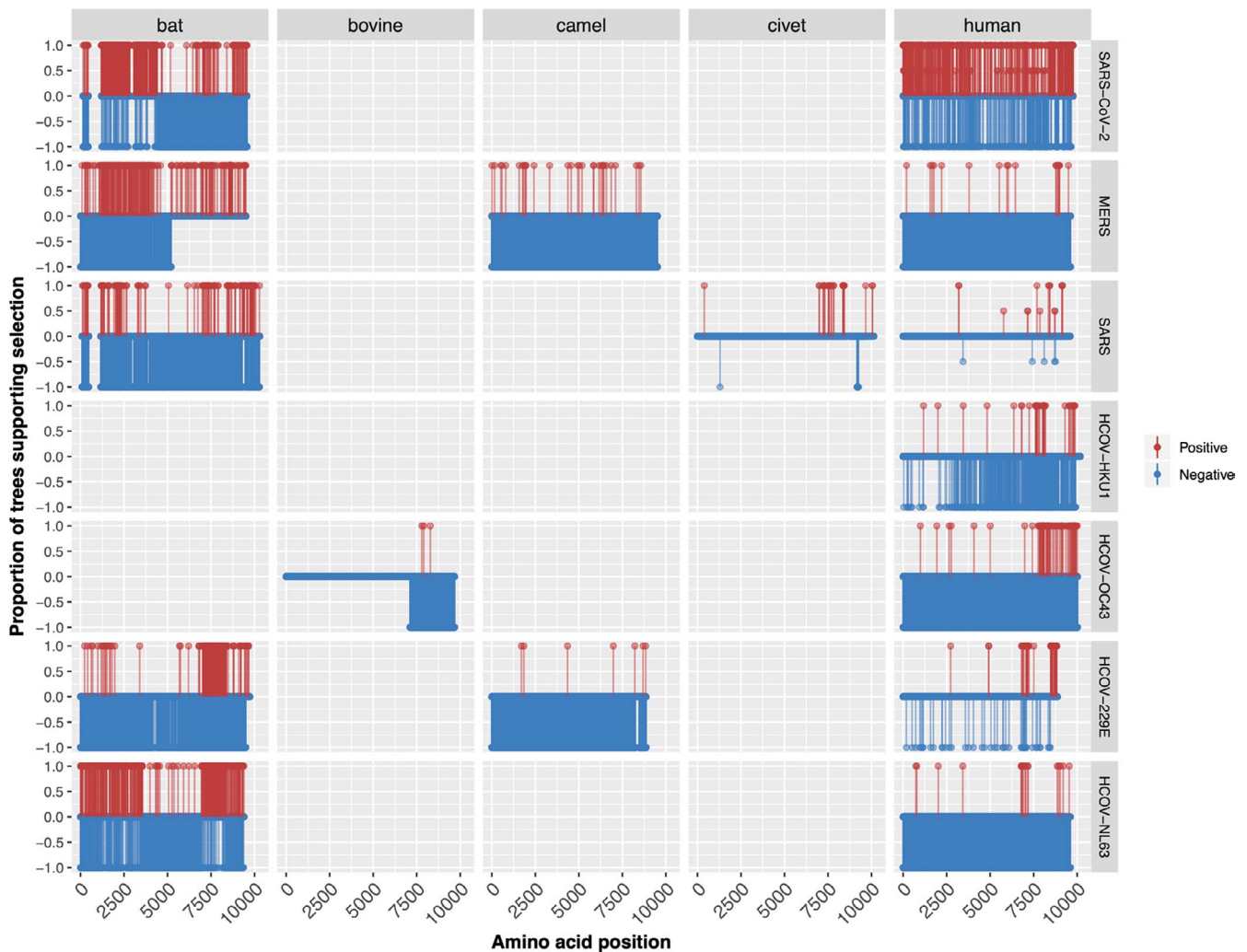
Although selection in bats was not significantly elevated throughout the entire genome relative to intermediate and human hosts, there were significantly more sites under positive selection in the *orf1ab* and spike gene regions ( $p < 0.05$ , Figure 4). Despite the abundance of selection in these gene regions, there were no significant differences identified in the degree of evolution between replication-associated, accessory or structural genes when all hosts were considered (Figure S5).

Within intermediate hosts, we found only limited selection relative to that observed in bats (Figure 4). In all four intermediate hosts, the *spike* gene was under positive selection either solitary (HCOV-OC43) or in combination with *orf1ab* and other genes (MERS-CoV, SARS-CoV and HCOV-229E). This prevalence of selection acting upon the *spike* gene is particularly interesting in that it suggests genes involved in receptor binding and/or frequent contact

with the immune response are undergoing adaptation whereas other genes obscured from such selective pressures are stabilized.

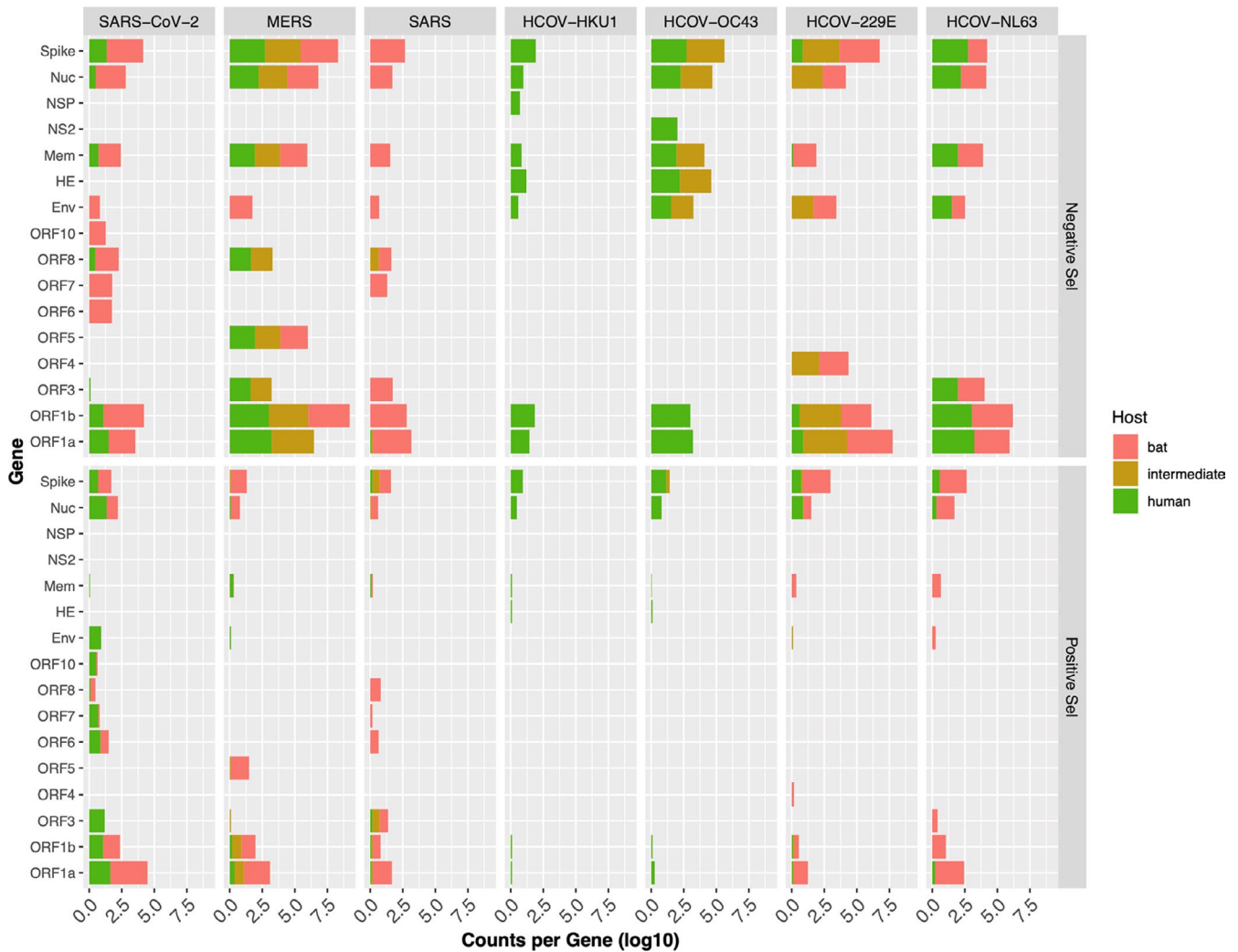
In the human hosts, there was relatively high variability in regions under positive selection; however, in general the 3' end of each genome was undergoing higher selective regimes (Figure 2, particularly evident in SARS-CoV, HCOV-OC43, HCOV-NL63 and HCOV-229E). The gene region with the highest frequency of positive selection was the *nucleocapsid* ( $n = 144$  unique sites). In the *spike* region, the two viruses with a relatively high number of sites undergoing positive selection were HCOV-OC43 ( $n = 39$ ) and SARS-CoV-2 ( $n = 21$ ).

Upon additional evolutionary tests that directly compare the  $d_N/d_S$  ratios of the viral isolates from primary/intermediate hosts relative to the respective viruses isolated from humans, there were a significant number of sites differing in their levels of adaptation (Figures S1 and S3). Furthermore, there were large discrepancies between primary and intermediate  $d_N/d_S$  ratios in comparison with human-derived viruses in genomic location and quantity (Figures S2



**FIGURE 2** Per-site analysis of selection across genomes of all coronaviruses in this study. Vertical lines represent each amino acid site under selection (supported by both FEL and MEME). The length of each line represents the proportion of phylogenetic trees ( $n = 10$ ) supporting each site of selection. Blue lines denote negative (purifying) selection and red lines portray positive (diversifying) selection





**FIGURE 3** Number of sites under positive (“Positive Sel”) or negative selection (“Negative Sel”) for each gene/ORF. Sites under selection are those determined to be significant for both FEL and MEME models. Counts of selection were normalized by the  $\log_{10}$  number of sequences used in each analysis. Results for civets, camels and bovine derived viruses were combined into the “intermediate” host category

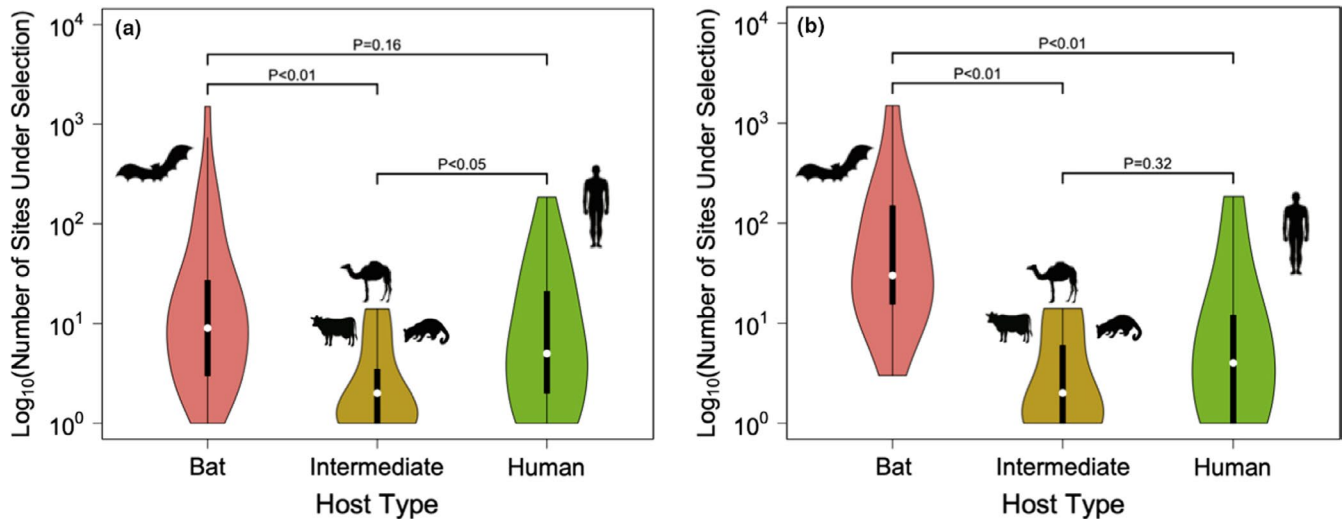
and S3). Specifically, the spike protein was more commonly under positive selection in the intermediate hosts.

Due to no concordant sites under positive selection from linear alignments for each of the coronavirus spike proteins, protein structures were aligned in a pairwise fashion in order to identify alternative positional patterns within their native three dimensional context (Zhang & Skolnick, 2005). All structural alignments resulted in TM-scores greater than 0.5 indicating the proteins are of the same fold (Xu & Zhang, 2010). Two sites identified from the alignments were also undergoing positive selection, one in SARS-CoV-2 and HCOV-OC43 (codons 1,185 and 1,186, respectively) and another for HCOV-NL63 and HCOV-229E (codons 307 and 321, respectively) (Figure 5).

Recent reports have illustrated that in each protomer of the SARS-CoV-2 trimeric spike protein there are 22 glycosylation sites that collectively shield ~40% of the protein surface (Watanabe et al., 2020). Although glycosylation can assist the virus in its ability to evade antibody neutralization, it has been well documented in HIV

and influenza viruses that these complex sugars also affect other aspects of the viral life cycle including receptor binding, expression and assembly (Li et al., 2020). The impact of glycosylation on each of the positively selected sites of each spike protein was investigated by identifying those which were in close proximity (<10 angstroms; Kobayashi & Suzuki, 2012) to the glycan at each respective glycosylation site (Figure S5). Interestingly, SARS-CoV-2 displayed elevated sites of selection in close proximity to glycans (Figure S5), and in conjunction with HCOV-OC43, SARS-CoV-2 displayed a consistently higher fraction of sites in close proximity after accounting for both the total spike positions in close proximity to sites of glycosylation and the number of spike amino acids (Figures S6 and S7). Monte Carlo permutation tests were then used to test whether glycosylation patterns were correlated with sites of positive selection. Only glycosylation sites for SARS-COV-2 and HCOV-229E were shown to be predictive of sites of positive selection ( $p < 0.01$ , Table S2).

The large amount of selection in SARS-CoV-2 suggests that the virus is still in the adaptive phase of the host switch, typical of recent



**FIGURE 4** Comparison of the prevalence of positively selected sites for (a) all genes for each host (bat, intermediate, and human), and (b) total positively selected sites for *orf1ab* and *spike* for each respective host group. Since HCOV-HKU1 and HCOV-OC43 have no bat host, they were removed from this analysis. Positively selected sites for each gene were summed and divided by the  $\log_{10}$  sequence counts. A Kruskal–Wallis statistical test was performed to examine differences between each group, and  $p$ -values are shown above violin plots. [Correction added on 27 March 2021, after first online publication: The caption has been modified.]

host switches (Parrish et al., 2008). To further explore this observation, we compared the estimated time since the introduction into the human population (Table S1, Figure 1) for all coronaviruses with the number of sites undergoing positive selection (Figure 6). Intriguingly, whereas there were varying degrees of correlation for each gene (Figure S2), in the *orf1a* and *orf1b* regions there was a moderate but significant inverse correlation observed in the number of sites under positive selection and time since infecting humans ( $p < 0.01$ ,  $R^2 > 0.93$ ).

## 4 | DISCUSSION

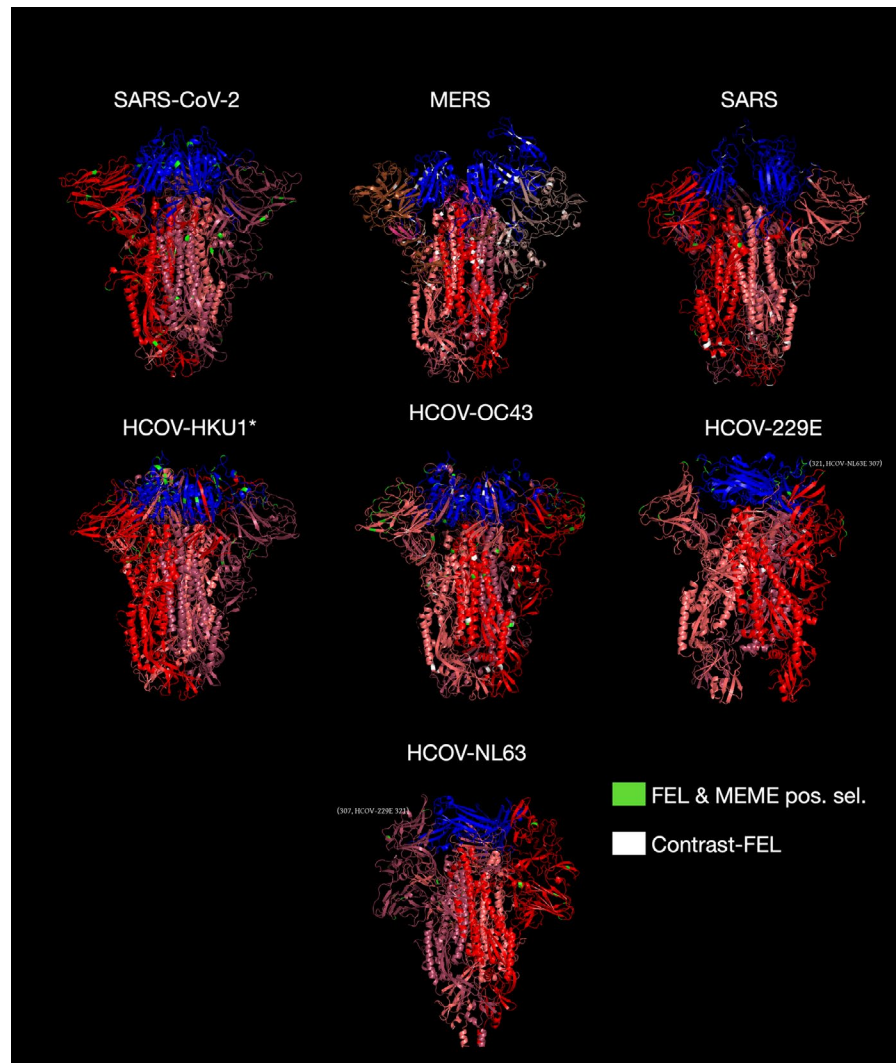
Viruses that frequently change hosts provide a unique opportunity to investigate evolutionary pathways critical for adaptation to novel host environments (Moncla et al., 2016; Morley et al., 2015). In this study, positive selection was observed sporadically across the genomes of each coronavirus, affecting both exogenous and endogenous proteins. Although the pattern of selection in each host was primarily unique to each virus, we provided evidence in line with our hypotheses that the timing of adaptation to each new host is correlated with a decreasing number of sites undergoing positive selection.

It is intuitive, and in accordance with both predictions derived from evolutionary theory and empirical evidence in other contexts (Braga et al., 2018; Ricklefs & Fallon, 2002), that following host switching, viral populations undergo a period of rapid diversification to facilitate adaptation to the novel environment (Abbott et al., 2013; Joy & Crespi, 2007; Turner & Elena, 2000). These adaptations may increase host receptor binding affinity, evade immune detection and promote viral establishment; subsequently, as viruses

ascend a peak of optimality in the fitness landscape (Gavrilets, 2004) (i.e. high affinity binding of the spike protein to the ACE2 receptor in humans for SARS-CoV-2), evolutionary forces favour stabilization and purifying selection to maintain viral phenotypes at new optima in the novel hosts (Morley et al., 2015; Pashley, 1988). This is also consistent with previous genomic investigations of epidemics where substitution rate estimates are inversely correlated with time of sampling (Holmes et al., 2016).

Parallel evolution, well documented throughout the tree of life (Boughman et al., 2005; Nosil et al., 2002; Schluter et al., 2004), has been repeatedly observed in viruses during host shifts both in vitro and in vivo (Bedhomme et al., 2012; Remold et al., 2008; Wichman et al., 1999). For example in the human immunodeficiency virus, identical mutations in codon 30 of the *gag* gene are common to all lineages that are originally derived from chimpanzees (Sharp & Hahn, 2010). Interestingly, this mutation was shown to increase the viral replication rate in human cell cultures and reverts back to the variant observed in simian immunodeficiency viruses upon culturing with chimpanzee T-lymphocytes (Longdon et al., 2014). Several SARS-CoV-2 variants have been observed with identical mutations in the receptor binding domain of the *spike* gene. For example, the variants N501Y.V2 (Ali et al., 2021; Leung et al., 2021) from South Africa, B.1.1.28 (Naveca et al., 2021) from Brazil, B.1.1.7 (also known as VOC202012/01) (Rambaut et al., 2020) from the United Kingdom, have all evolved the same mutations in the receptor binding domain of the *spike* gene in parallel in disparate geographic locations and these variants are associated with enhanced transmissibility. The observation of parallel evolution in the receptor binding domain suggests repeated selection of mutations enhancing transmissibility in humans (Leung et al., 2021; Santos & Passos, 2021).

**FIGURE 5** Spike protein structures. Each Spike protomer is coloured a shade of red, whereas the receptor binding site is shaded blue. Positively selected sites supported by either FEL and MEME are shaded in green and those sites with elevated  $d_N/d_S$  ratios relative to their intermediate host, if available, are shaded in white. \*Sites for HCOV-HKU1 are overlaid onto their closest coronavirus relative with an available structure, HCOV-OC43. White text labels illustrate sites identified in the TM-alignment

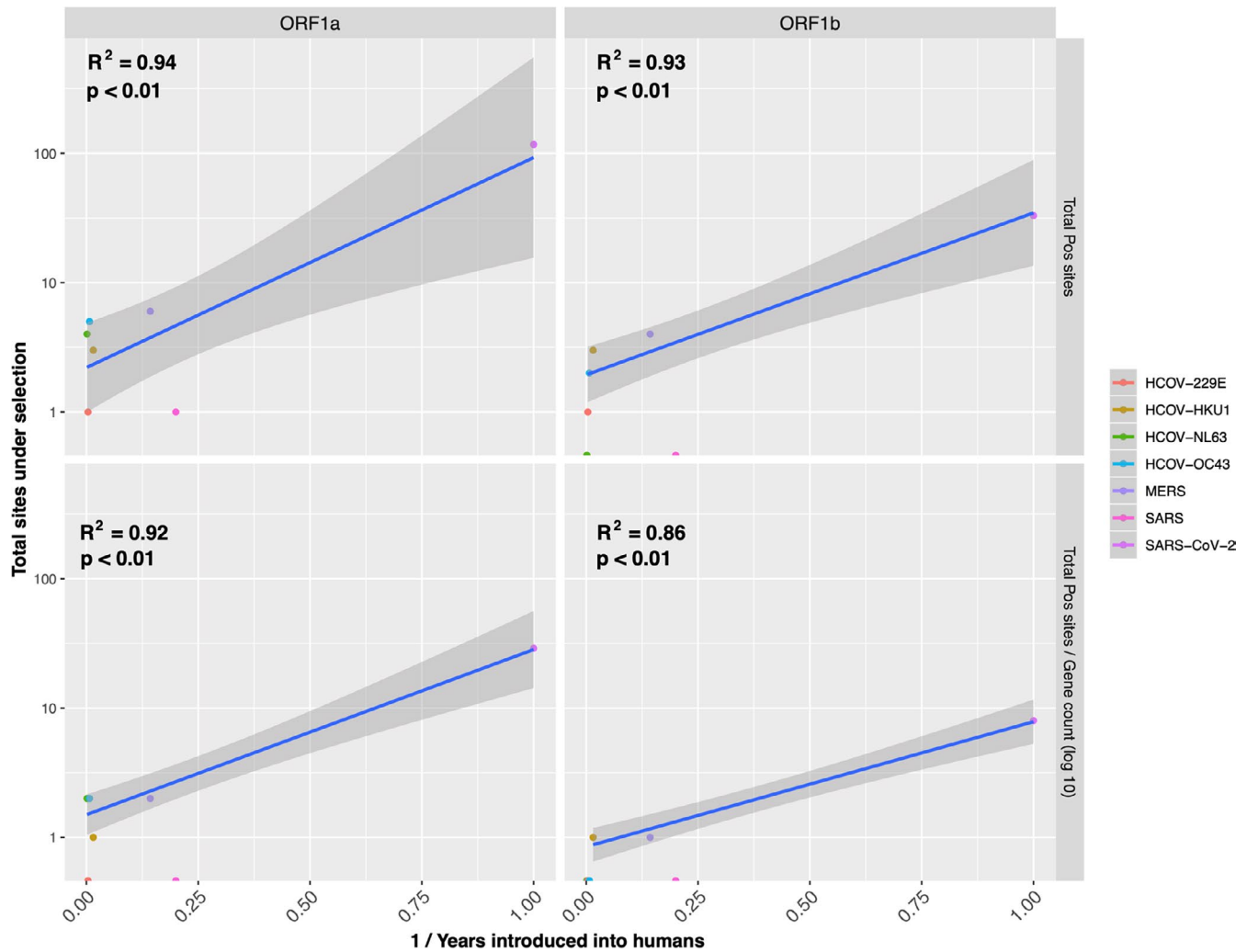


Previous research on selection within coronaviruses has similarly found varying evolutionary signatures. For example, Tang et al. (2009) analysed selection in ORFs across the SARS-CoV genome for humans, civets and bats. Positive selection was identified in the early and middle phases of the epidemic, whereas no selection was detected in the later phases. Although nine sites were identified in the *spike* gene, 53 sites were positively selected for human isolates in the remaining ORFs (Tang et al., 2009). Furthermore, although positive selection predominantly acted upon human and civet clades, there was no positive selection identified in bats. In contrast, our study revealed extensive positive selection in SARS viruses isolated from bats and limited selection on SARS. The observed differences between the two studies are likely due to the 47 additional bat sequences now available and used in our study, whereas the differences in selection in humans could be explained by differences in methodology in that Tang *et al.* used a branch-site model of selection as implemented in PAML whereas in this analysis we used FEL and MEME models of selection in HyPhy. An additional explanation is that in our study sequences that had been cultured prior to sequencing were excluded. Similar to SARS, analyses of MERS-CoV have found a higher degree of positive selection

for genes involved in replication relative to the Spike gene (Forni et al., 2016), largely consistent with the results obtained in our study (e.g. Figures 2 and 3). The implications of a larger degree of positive selection in replicase genes for both SARS-CoV and MERS-CoV are intriguing as it suggests genes other than spike are under strong diversifying selection immediately following host-switching events consistent with patterns observed in influenza (Bhatt et al., 2013). An alternative and more simplified explanation, however, could simply be due to the significantly longer lengths for genes involved in replication. Studies investigating the evolution of SARS-CoV-2 have largely found that purifying selection has dominated its evolution (MacLean et al., 2020). It was also demonstrated that extensive positive selection likely occurred prior to its introduction into the human population (MacLean et al., 2020). Analyses using more recent genomes (December 2020) have revealed evidence for positive selection on 1,860 sites, again predominantly affecting replication-associated genes ( $n = 1,222$ ) compared with those sites selected in the spike gene ( $n = 237$ ) (Pond, 2020).

The proximity between glycans and sites of positive selection in the SARS-CoV-2 spike protein are intriguing specifically within the context of other coronaviruses. Previous research has found that





**FIGURE 6** The relationships between the number of positively selected sites and estimated years since introduction into humans. Using estimated years from previous research, the inverse estimated years since introduction is plotted against the total number of unique sites under positive selection (FEL and MEME) and the same sites divided by the  $\log_{10}$  normalized counts of sequences used in each selection analysis for both the ORF1a and ORF1b genes. Adjusted  $R^2$  values and  $p$ -values derived from a linear regression (blue line, shaded regions represent 95% confidence intervals) are also shown for each plot

the SARS-CoV-2 spike protein has a relatively large glycan shield (Watanabe et al., 2020), and when compared with other coronaviruses, there were few patterns concerning the precise sites of glycosylation (Xu et al., 2020). In this study the proximity of glycans to sites of selection was investigated. Proximity to glycans was significantly elevated for sites of selection in both HCoV-OC43 and SARS-CoV-2. These elevated levels for SARS-CoV-2 suggests that selection in this virus is multifarious and not predominantly driven by antibody mediated selection. Furthermore, an additional comparison of the 32 unique neutralizing epitopes composed of 125 unique amino acid positions for SARS-CoV-2 revealed that only two of these sites were identified as sites of selection. Due to a lack of studies investigating the neutralization responses for most other coronaviruses, there were no epitopes for any of the other human coronaviruses available.

Our results are subjected to a number of caveats. First, we were limited to the sequencing data available on public repositories.

Therefore, widely sequenced viruses such as SARS-CoV-2 were compared to less well-represented viruses such as HCoV-229E. Thus, our findings may have been partially confounded by the disproportionate number of sequences available for viruses that demonstrated the most divergence. However, we attempted to account for this discrepancy when comparing each of the coronaviruses in this study by normalizing counts of selection by the number of sequences for each virus (Figures 3 and 6). Nevertheless, this discrepancy may still influence the results of this study, and once further data arise for other coronaviruses, future work may profitably investigate the relationships outlined in this study. Concerning the moderate relationship identified between the number of sites under selection and estimated divergence times, it is interesting that this relationship was only identified in *orf1a* and *orf1b* (Figure S2). In fact, the opposite trend was observed for the spike protein where the number of sequences was inversely correlated with the number of positively selected sites. This may imply that while endogenous proteins

stabilize following host-switching events, exogenous proteins are continuously adapting in an evolutionary arms race with host immune responses. However, when SARS-CoV-2 is removed from pan-viral comparisons of selection relative to divergence times, the relationship then lacks significance ( $p > .1$ ) indicating that with the available data, the abundant selection observed in SARS-CoV-2 is driving the significance of this relationship. Further limitations exist within virus–host groups where the proportional representation of a given host level (primary, intermediate, human) could also affect our ability to detect selection. Finally, the limited number of groups being compared reduced the power of further statistical analyses.

Our findings add strength to the existing body of literature on diversification associated with colonization of novel environments and contribute to understanding the genomic and phenotypic complexity of adaptation required during the colonization of novel environments with some granularity. Our results also complement those observed in other viruses that have documented host switches, such as within the *Rhaboviridae* family (Streicker et al., 2012) where each virus has followed a relatively unique evolutionary pathway between hosts. The selective regimes placed upon the genomes of each virus seems to reflect the delicate balance between viral genetics and the unique ecologies of both the viruses and their respective hosts. However, for several viruses in this study, the number of amino acid sites under positive selection decreases with time since emergence into the human population. Although not solely predictive of future pathogen emergence, additional cases of documented host-switching events may permit more statistically robust analyses in order to further examine the possibility that hierarchical patterns will emerge. Emerging signatures reflecting each viral evolutionary pathway could then be stratified with comparisons of viral genomes supplemented with additional variables including the biology, ecology and evolution of their respective donor and recipient hosts. Our results in documenting genomic adaptations of SARS-CoV-2 and the observation of repeated evolution of particular mutations favouring enhanced transmissibility, highlight the importance of genomic surveillance of emerging viruses to identify and track variants of interest for public health purposes.

## ACKNOWLEDGMENTS

This project was supported by an operating grant from the Canadian Institutes of Health Research Coronavirus Rapid Response Programme grant number 440371 a Genome Canada Bioinformatics and Computational Biology Programme grant 287PHY, and the BC Centre for Excellence in HIV/AIDS. We thank the editors and anonymous reviewers of our manuscript for their time and for providing insightful comments which greatly improved our manuscript.

## CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

## AUTHOR CONTRIBUTIONS

VM conceived of and coordinated the study, performed data analyses, generated figures and tables and wrote the first draft of the manuscript. JBJ conceived of and coordinated the study, critically

examined analyses and results, generated figures and contributed to writing the first draft of the manuscript. GJM, RLM and AM critically examined the analyses and results and contributed to draft versions of the manuscript. All authors approved the final version.

## DATA AVAILABILITY STATEMENT

All data used in this analysis are freely available to access on GenBank, GISAID and the RSCB Protein Databank. The accession numbers for all sequences used in the analysis are available in Table S3 and all positively selected sites are in Table S4, on GitHub. <https://github.com/vmon5813/CoronavirusHostAdaptations>

All alignment files are available on request.

## ORCID

Vincent Montoya  <https://orcid.org/0000-0002-0615-4591>  
 Angela McLaughlin  <https://orcid.org/0000-0001-5606-9080>  
 Gideon J. Mordecai  <https://orcid.org/0000-0001-8397-9194>  
 Rachel L. Miller  <https://orcid.org/0000-0002-3152-905X>  
 Jeffrey B. Joy  <https://orcid.org/0000-0002-7013-1482>

## REFERENCES

- Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., ... Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26, 229–246. <https://doi.org/10.1111/j.1420-9101.2012.02599.x>
- Ali, F., Kasry, A., & Amin, M. (2021). The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutation. *ArXiv*.
- Al-Khannaq, M. N., Ng, K. T., Oong, X. Y., Pang, Y. K., Takebe, Y., Chook, J. B., Hanafi, N. S., Kamarulzaman, A., & Tee, K. K. (2016). Molecular epidemiology and evolutionary histories of human coronavirus OC43 and HKU1 among patients with upper respiratory tract infections in Kuala Lumpur, Malaysia. *Virology Journal*, 13, 33. <https://doi.org/10.1186/s12985-016-0488-4>
- Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, 26, 450–452. <https://doi.org/10.1038/s41591-020-0820-9>
- Bedhomme, S., Lafforgue, G., & Elena, S. F. (2012). Multihost experimental evolution of a plant RNA virus reveals local adaptation and host-specific mutations. *Molecular Biology and Evolution*, 29, 1481–1492. <https://doi.org/10.1093/molbev/msr314>
- Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., & Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology*, 7(Suppl), 957–959. <https://doi.org/10.1038/80734>
- Berrio, A., Gartner, V., & Wray, G. A. (2020). Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. *PeerJ*, 8, e10234. <https://doi.org/10.7717/peerj.10234>
- Bhatt, S., Lam, T. T., Lycett, S. J., Leigh Brown, A. J., Bowden, T. A., Holmes, E. C., Guan, Y., Wood, J. L. N., Brown, I. H., Kellam, P., Pybus, O. G., Brown, I., Brookes, S., Germundsson, A., Cook, A., Williamson, S., Essen, S., Garcon, F., Gunn, G., ... Enstone, J. (2013). The evolutionary dynamics of influenza A virus adaptation to mammalian hosts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368, 20120382. <https://doi.org/10.1098/rstb.2012.0382>
- Bohne-Lang, A., & von der Lieth, C. W. (2005). GlyProt: In silico glycosylation of proteins. *Nucleic Acids Research*, 33, W214–W219. <https://doi.org/10.1093/nar/gki385>

- Boni, M. F., Lemey, P., Jiang, X., Lam, T.-T.-Y., Perry, B. W., Castoe, T. A., Rambaut, A., & Robertson, D. L. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology*, 5, 1408–1417. <https://doi.org/10.1038/s41564-020-0771-4>
- Boughman, J. W., Rundle, H. D., & Schluter, D. (2005). Parallel evolution of sexual isolation in sticklebacks. *Evolution*, 59, 361–373. <https://doi.org/10.1111/j.0014-3820.2005.tb00995.x>
- Braga, M. P., Guimaraes, P. R. Jr, Wheat, C. W., Nylin, S., & Janz, N. (2018). Unifying host-associated diversification processes using butterfly-plant networks. *Nature Communications*, 9, 5155. <https://doi.org/10.1038/s41467-018-07677-x>
- Chan, P. K., & Chan, M. C. (2013). Tracing the SARS-coronavirus. *Journal of Thoracic Disease*, 5(Suppl 2), S118–S121. <https://doi.org/10.3978/j.issn.2072-1439.2013.06.19>
- Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High speed BLASTN: An accelerated MegaBLAST search tool. *Nucleic Acids Research*, 43, 7762–7768. <https://doi.org/10.1093/nar/gkv784>
- Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., & Rodrigo, A. G. (2003). Measurably evolving populations. *Trends in Ecology and Evolution*, 18, 481–488. [https://doi.org/10.1016/S0169-5347\(03\)00216-7](https://doi.org/10.1016/S0169-5347(03)00216-7)
- Duffy, S., Burch, C. L., & Turner, P. E. (2007). Evolution of host specificity drives reproductive isolation among RNA viruses. *Evolution*, 61, 2614–2622. <https://doi.org/10.1111/j.1558-5646.2007.00226.x>
- Forni, D., Cagliani, R., Mozzi, A., Pozzoli, U., Al-Daghri, N., Clerici, M., & Sironi, M. (2016). Extensive positive selection drives the evolution of nonstructural proteins in lineage C betacoronaviruses. *Journal of Virology*, 90, 3627–3639. <https://doi.org/10.1128/JVI.02988-15>
- Gavrilets, S. (2004). *Fitness landscapes and the origin of species*. Princeton University Press.
- Holmes, E. C., Dudas, G., Rambaut, A., & Andersen, K. G. (2016). The evolution of Ebola virus: Insights from the 2013–2016 epidemic. *Nature*, 538, 193–200. <https://doi.org/10.1038/nature19790>
- Hughes, A. L., & Hughes, M. A. (2007). More effective purifying selection on RNA viruses than in DNA viruses. *Gene*, 404, 117–125. <https://doi.org/10.1016/j.gene.2007.09.013>
- Huynh, J., Li, S., Yount, B., Smith, A., Sturges, L., Olsen, J. C., Nagel, J., Johnson, J. B., Agnihotram, S., Gates, J. E., Frieman, M. B., Baric, R. S., & Donaldson, E. F. (2012). Evidence supporting a zoonotic origin of human coronavirus strain NL63. *Journal of Virology*, 86, 12816–12825. <https://doi.org/10.1128/JVI.00906-12>
- Joy, J. B., & Crespi, B. J. (2007). Adaptive radiation of gall-inducing insects within a single host-plant species. *Evolution*, 61, 784–795. <https://doi.org/10.1111/j.1558-5646.2007.00069.x>
- Joy, J. B., & Crespi, B. (2012). Island phytophagy: Explaining the remarkable diversity of phytophagous insects. *Proceedings of the Royal Society B: Biological Sciences*, 279, 3250–3255.
- Katoh, K., Asimenos, G., & Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. *Methods in Molecular Biology*, 537, 39–64.
- Kobayashi, Y., & Suzuki, Y. (2012). Evidence for N-glycan shielding of antigenic sites during evolution of human influenza A virus hemagglutinin. *Journal of Virology*, 86, 3446–3451. <https://doi.org/10.1128/JVI.06147-11>
- Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. (2006). GARD: A genetic algorithm for recombination detection. *Bioinformatics*, 22, 3096–3098. <https://doi.org/10.1093/bioinformatics/btl474>
- Krakauer, D. C., & Komarova, N. L. (2003). Levels of selection in positive-strand virus dynamics. *Journal of Evolutionary Biology*, 16, 64–73. <https://doi.org/10.1046/j.1420-9101.2003.00481.x>
- Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30, 3276–3278.
- Lau, S. K., Feng, Y., Chen, H., Luk, H. K. H., Yang, W.-H., Li, K. S. M., Zhang, Y.-Z., Huang, Y., Song, Z.-Z., Chow, W.-N., Fan, R. Y. Y., Ahmed, S. S., Yeung, H. C., Lam, C. S. F., Cai, J.-P., Wong, S. S. Y., Chan, J. F. W., Yuen, K.-Y., Zhang, H.-L., & Woo, P. C. Y. (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, 10532–10547. <https://doi.org/10.1128/JVI.01048-15>
- Leao, J. C., Gusmao, T. P. L., Zarzar, A. M., Leao Filho, J. C., Barkokebas Santos de Faria, A., Morais Silva, I. H., Gueiros, L. A. M., Robinson, N. A., Porter, S., & Carvalho, A. A. T. (2020). Coronaviridae-Old friends, new enemy!. *Oral Diseases*, 1–6. <https://doi.org/10.1111/odi.13447>
- Leung, K., Shum, M. H. H., Leung, G. M., Lam, T. T. Y., & Wu, J. T. (2021). Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Eurosurveillance*, 26(1). <https://doi.org/10.2807/1560-7917.ES.2020.26.1.2002106>
- Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., Qin, H., Wang, M., Lu, Q., Li, X., Sun, Q., Liu, J., Zhang, L., Li, X., Huang, W., & Wang, Y. (2020). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*, 182(5), 1284–1294.e9. <https://doi.org/10.1016/j.cell.2020.07.012>
- Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J., & Jiggins, F. M. (2014). The evolution and genetics of viral host shifts. *PLoS Pathogens*, 10, e1004395. <https://doi.org/10.1371/journal.ppat.1004395>
- MacLean, O. A., Lytras, S., Weaver, S., Singer, J. B., Boni, M. F., Lemey, P., Kosakovsky Pond, S. L., & Robertson, D. L. (2020). Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. *bioRxiv*. <https://doi.org/10.1101/2020.05.28.122366>. [Epub ahead of print].
- Mollentze, N., Biek, R., & Streicker, D. G. (2014). The role of viral evolution in rabies host shifts and emergence. *Current Opinion in Virology*, 8, 68–72. <https://doi.org/10.1016/j.coviro.2014.07.004>
- Moncla, L. H., Zhong, G., Nelson, C. W., Dinis, J. M., Mutschler, J., Hughes, A. L., Watanabe, T., Kawaoka, Y., & Friedrich, T. C. (2016). Selective bottlenecks shape evolutionary pathways taken during mammalian adaptation of a 1918-like avian influenza virus. *Cell Host and Microbe*, 19, 169–180. <https://doi.org/10.1016/j.chom.2016.01.011>
- Morley, V. J., Mendiola, S. Y., & Turner, P. E. (2015). Rate of novel host invasion affects adaptability of evolving RNA virus lineages. *Proceedings of the Royal Society B: Biological Sciences*, 282, 20150801. <https://doi.org/10.1098/rspb.2015.0801>
- Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, 8, e1002764. <https://doi.org/10.1371/journal.pgen.1002764>
- Naveca, F., Nascimento, V., Souza, V., Corado, A., Nascimento, F., Silva, F., Costa, A., Duarte, D., Pessoa, K., Gonçalves, L., Brandão, M. J., Jesus, M., Fernandes, C., Pinto, R., Silva, M., Mattos, T., Wallau, G. L., Mendonça Siqueira, M., Cristina Resende, P., ... Bello, G. (2021). Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein. *Virological.org*. <https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585>
- Nosil, P., Crespi, B. J., & Sandoval, C. P. (2002). Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, 417, 440–443. <https://doi.org/10.1038/417440a>
- Olival, K. J., Hosseini, P. R., Zambrana-Torrel, C., Ross, N., Bogich, T. L., & Daszak, P. (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature*, 546, 646–650. <https://doi.org/10.1038/nature22975>
- Parker, J., Rambaut, A., & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution*, 8, 239–246. <https://doi.org/10.1016/j.meegid.2007.08.001>

- Parrish, C. R., Holmes, E. C., Morens, D. M., Park, E.-C., Burke, D. S., Calisher, C. H., Laughlin, C. A., Saif, L. J., & Daszak, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiology and Molecular Biology Reviews*, 72, 457–470. <https://doi.org/10.1128/MMBR.00004-08>
- Pashley, D. P. (1988). Quantitative genetics, development, and physiological adaptation in host strains of fall armyworm. *Evolution*, 42, 93–102. <https://doi.org/10.1111/j.1558-5646.1988.tb04110.x>
- Pfefferle, S., Oppong, S., Drexler, J. F., Gloza-Rausch, F., Ipsen, A., Seebens, A., Müller, M. A., Annan, A., Vallo, P., Adu-Sarkodie, Y., Kruppa, T. F., & Drosten, C. (2009). Distant relatives of severe acute respiratory syndrome coronavirus and close relatives of human coronavirus 229E in bats, Ghana. *Emerging Infectious Diseases*, 15, 1377–1384. <https://doi.org/10.3201/eid1509.090224>
- Pond, S. L. K. (2020). *Natural selection analysis of SARS-CoV-2/COVID-19*. <https://observablehq.com/@spond/revised-sars-cov-2-analytics-page?collection=@spond/sars-cov-2>
- Pond, S. L. K., & Frost, S. D. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, 22, 1208–1222. <https://doi.org/10.1093/molbev/msi105>
- Pond, S. L. K., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., Shank, S. D., Magalis, B. R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S. J., Frost, S. D. W., & Muse, S. V. (2020). HyPhy 2.5—A customizable platform for evolutionary hypothesis testing using phylogenies. *Molecular Biology and Evolution*, 37, 295–299. <https://doi.org/10.1093/molbev/msz197>
- Pond, S. L. K., Wisotsky, S., Escalante, A. A., Magalis, B. R., & Weaver, S. (2020). Contrast-FEL: A test for differences in selective pressures at individual sites among clades and sets of branches. *Molecular Biology and Evolution*. <https://doi.org/10.1093/molbev/msaa263>
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One*, 5, e9490. <https://doi.org/10.1371/journal.pone.0009490>
- Schrodinger, LLC (2015). *The PyMOL molecular graphics system, Version 2.4*. Schrodinger, LLC.
- Rambaut, A. et al. (2020). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Genomic Epidemiology*.
- Remold, S. K., Rambaut, A., & Turner, P. E. (2008). Evolutionary genomics of host adaptation in vesicular stomatitis virus. *Molecular Biology and Evolution*, 25, 1138–1147. <https://doi.org/10.1093/molbev/msn059>
- Ricklefs, R. E., & Fallon, S. M. (2002). Diversification and host switching in avian malaria parasites. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269, 885–892. <https://doi.org/10.1098/rspb.2001.1940>
- Sanchez-Flores, A., Peñaloza, F., Carpinteyro-Ponce, J., Nazario-Yepiz, N., Abreu-Goodger, C., Machado, C. A., & Markow, T. A. (2016). Genome evolution in three species of cactophilic *Drosophila*. *G3 (Bethesda)*, 6, 3097–3105. <https://doi.org/10.1534/g3.116.033779>
- Santos, J. C., & Passos, G. A. (2021). The high infectivity of SARS-CoV-2 B. 1.1. 7 is associated with increased interaction force between Spike-ACE2 caused by the viral N501Y mutation. *bioRxiv*. <https://doi.org/10.1101/2020.12.29.424708>
- Schluter, D., Clifford, E. A., Nemethy, M., & McKinnon, J. S. (2004). Parallel evolution and inheritance of quantitative traits. *The American Naturalist*, 163, 809–822. <https://doi.org/10.1086/383621>
- Sharp, P. M., & Hahn, B. H. (2010). The evolution of HIV-1 and the origin of AIDS. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365, 2487–2494. <https://doi.org/10.1098/rstb.2010.0031>
- Spielman, S. J., Weaver, S., Shank, S. D., Magalis, B. R., Li, M., & Kosakovsky Pond, S. L. (2019). Evolution of viral genomes: Interplay between selection, recombination, and other forces. *Methods in Molecular Biology*, 1910, 427–468. [https://doi.org/10.1007/978-1-4939-9074-0\\_14](https://doi.org/10.1007/978-1-4939-9074-0_14)
- Streicker, D. G., Altizer, S. M., Velasco-Villa, A., & Rupprecht, C. E. (2012). Variable evolutionary routes to host establishment across repeated rabies virus host shifts among bats. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 19715–19720. <https://doi.org/10.1073/pnas.1203456109>
- Tagliamonte, M. S., Abid, N., Borocci, S., Sangiovanni, E., Ostrov, D. A., Kosakovsky, S. L., Pond, M. S., Chillemi, G., & Mavian, C. (2020). Multiple recombination events and strong purifying selection at the origin of SARS-CoV-2 spike glycoprotein increased correlated dynamic movements. *International Journal of Molecular Sciences*, 22, 80. <https://doi.org/10.3390/ijms22010080>
- Tang, X., Li, G., Vasilakis, N., Zhang, Y., Shi, Z., Zhong, Y., Wang, L.-F., & Zhang, S. (2009). Differential stepwise evolution of SARS coronavirus functional proteins in different host species. *BMC Evolutionary Biology*, 9, 52. <https://doi.org/10.1186/1471-2148-9-52>
- Turner, P. E., & Elena, S. F. (2000). Cost of host radiation in an RNA virus. *Genetics*, 156, 1465–1470.
- Vijgen, L., Keyaerts, E., Moës, E., Thoelen, I., Wollants, E., Lemey, P., Vandamme, A.-M., & Van Ranst, M. (2005). Complete genomic sequence of human coronavirus OC43: Molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *Journal of Virology*, 79, 1595–1604. <https://doi.org/10.1128/JVI.79.3.1595-1604.2005>
- Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47, D339–D343. <https://doi.org/10.1093/nar/gky1006>
- Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S., & Crispin, M. (2020). Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*, 369, 330–333. <https://doi.org/10.1126/science.abb9983>
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46, W296–W303. <https://doi.org/10.1093/nar/gky427>
- Wertheim, J. O., Chu, D. K., Peiris, J. S., Kosakovsky Pond, S. L., & Poon, L. L. (2013). A case for the ancient origin of coronaviruses. *Journal of Virology*, 87, 7039–7045. <https://doi.org/10.1128/JVI.03273-12>
- Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M., & Bull, J. J. (1999). Different trajectories of parallel evolution during viral adaptation. *Science*, 285, 422–424. <https://doi.org/10.1126/science.285.5426.422>
- Worobey, M., Santiago, M. L., Keele, B. F., Ndjango, J.-B.-N., Joy, J. B., Labama, B. L., Dheda, B. D., Rambaut, A., Sharp, P. M., Shaw, G. M., & Hahn, B. H. (2004). Origin of AIDS: Contaminated polio vaccine theory refuted. *Nature*, 428, 820. <https://doi.org/10.1038/428820a>
- Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26, 889–895. <https://doi.org/10.1093/bioinformatics/btq066>
- Xu, W., Wang, M., Yu, D., & Zhang, X. (2020). Variations in SARS-CoV-2 spike protein cell epitopes and glycosylation profiles during global transmission course of COVID-19. *Frontiers in Immunology*, 11, 565278. <https://doi.org/10.3389/fimmu.2020.565278>
- Zeileis, A., Wiel, M., Hornik, K., & Hothorn, T. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28, 1–23.
- Zhang, Y., & Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33, 2302–2309. <https://doi.org/10.1093/nar/gki524>
- Zhang, Z., Shen, L., & Gu, X. (2016). Evolutionary dynamics of MERS-CoV: Potential recombination, positive selection and transmission. *Scientific Reports*, 6, 25049. <https://doi.org/10.1038/srep25049>
- Zhao, L., Seth-Pasricha, M., Stemate, D., Crespo-Bellido, A., Gagnon, J., Draghi, J., & Duffy, S. (2019). Existing host range mutations constrain

further emergence of RNA viruses. *Journal of Virology*, 93, e01385-18.  
<https://doi.org/10.1128/JVI.01385-18>

#### SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Montoya V, McLaughlin A, Mordecai GJ, Miller RL, Joy JB. Variable routes to genomic and host adaptation among coronaviruses. *J Evol Biol.* 2021;34:924–936. <https://doi.org/10.1111/jeb.13771>