Submission is intended as: Article in the Discoveries section of MBE

Proposed title:

Identification of site-specific evolutionary trajectories shared across human betacoronaviruses

Authors listed according to affiliation
Marina Escalera-Zamudio [1*]
Sergei L. Kosakovsky Pond [2]
Natalia Martínez de la Viña [1]
Bernardo Gutiérrez [1]
Julien Thézé [1,3]
Thomas A. Bowden [4]
Oliver G. Pybus [1]
Ruben J.G. Hulswit [4*]

Affiliations:

1. Department of Zoology, University of Oxford, Parks Rd Oxford, OX1 3PS, UK
2. Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, USA
3. Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, Saint-Genès-Champanelle, France
4. Division of Structural Biology, Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK

Email addresses (*corresponding authors):
marina.escalerazamudio@zoo.ox.ac.uk *
spond@temple.edu
natalia.martinezdelavina@zoo.ox.ac.uk
bernardo.gutierrez@zoo.ox.ac.uk
julien.theze@inrae.fr
thomas.bowden@strubi.ox.ac.uk
oliver.pybus@zoo.ox.ac.uk
ruben.hulswit@strubi.ox.ac.uk *

ABSTRACT (228)

Comparison of evolution among related viruses can provide insights into shared adaptive processes, for example following host switching to a mutual host species. Whilst phylogenetic methods can help identify mutations that may be important for evolutionary processes such as adaptation to a new host, these can be enhanced by positioning candidate mutations to known functional sites on protein structures. Over the past two decades, three zoonotic betacoronaviruses have significantly impacted human public health: SARS-CoV-1, MERS-CoV and SARS-CoV-2, whilst two other betacoronaviruses, HKU1 and OC43, have circulated endemically in the human population for over 100 years. In this study, we use a comparative approach to prospectively search for potentially evolutionarily-relevant mutations within the Orf1ab and S genes across betacoronavirus species that have demonstrated sustained human-to-human transmission (HKU1, OC43, SARS-CoV-1 and SARS-CoV-2). We used a combination of molecular evolution methods to identify 30 sites that display evidence of homoplasy and/or stepwise evolution, that may be suggestive of adaptation across emerging and endemic betacoronaviruses. Of these, seven sites also display evidence of being selectively relevant. Drawing upon known protein structure data, we find that four of the identified mutations [18121 (exonuclease/27), 21623 (spike/21), 21635 (spike/25) and 23948 (spike/796), in SARS-CoV-2 genome coordinates] are proximal to regions of known functionality. Our results provide a molecular-level context for common evolutionary pathways that betacoronaviruses may undergo during adaptation to the human host.

INTRODUCTION

Mutation is a fundamental process for virus evolution, generating genetic variability and enabling evolutionary change (Loewe and Hill 2010). The vast majority of mutations are expected to be either detrimental to virus fitness and eliminated through purifying selection, or selectively neutral and subjected to random genetic drift. Only a small proportion of mutations are expected to be adaptive and subsequently maintained through positive selection (Pond, et al. 2012). Although RNA viruses evolve rapidly due to their relatively small genomes and high mutation rates, mutational pathways leading to adaptation are limited by the functional constraints of the interacting genes they encode (Dolan, et al. 2018). If admissible genetic variability is indeed limited, then adaptive evolutionary trajectories may exhibit constrained, and sometimes recurrent patterns (Gutierrez, et al. 2019).

In the context of virus evolution, *homoplasy* (or parallel evolution) is defined as the appearance of the same mutations in lineages that do not share direct common ancestry, a phenomenon that is potentially informative on adaptation (Gutierrez, et al. 2019). For example, the parallel loss of the hemagglutinin-esterase (HE) protein lectin function in endemic betacoronaviruses HKU1 and OC43 likely reflects their convergent adaptation to human hosts (Bakkers, et al. 2017). Recurring mutation patterns have also been linked to the emergence of highly-pathogenic genotypes/phenotypes in avian influenza A and polio viruses (Stern, et al. 2017; Escalera-Zamudio, et al. 2020). Another evolutionary pattern that may reflect adaptation is *stepwise evolution*, in which genome sites are subjected to mutational change between at least two states (A→B) occurring in directed steps towards a local fitness optimum (*e.g.* mutation A→B, but without immediate reversion B→A) (Figure 1) (Delport, et al. 2008) (Farris 1977). For example, stepwise evolution may be reflective of selective pressure exerted by host immune responses (Boni, et al. 2006; Starr, et al. 2021)

Coronaviruses are well known for their propensity to switch host species, as evidenced by the zoonotic introduction of three such viruses over the past two decades. Whilst the emergence of SARS-CoV-2 variants in recent months indicates that adaptation of SARS-CoV-2 to the human host environment is ongoing (O'Toole, et al. 2021), the relatively homogeneous genetic composition of the SARS-CoV-2 population (Rausch, et al. 2020) results in a limited power to detect emerging adaptive mutations using standard analytical methods (van Dorp, Richard, et al. 2020). Under these circumstances, the contextualization of comparative molecular evolution within a protein structure framework may provide a complementary approach for identifying evolutionary change related to adaptation in different virus populations (Ellegren 2008; Hulswit, et al. 2016; Avanzato, et al. 2019; Escalera-Zamudio, et al. 2020).

The four betacoronaviruses capable of sustained human-to-human transmission (OC43, HKU1, SARS-CoV-1, SARS-CoV-2) were introduced into human populations through independent zoonotic events, and fall within two virus lineages (International Committee on Taxonomy of Viruses 2012, Zhou, et al. 2020). Lineage A (LinA, *Embecovirus* subgenus) includes the OC43 viruses, first described in 1967 (McIntosh, et al. 1967), and the HKU1 viruses identified in 2005 (Woo, et al. 2005), both associated with a mild respiratory disease (Su, et al. 2016). HKU1 and OC43 became endemic to humans following their introduction, which is estimated to have occurred >100 years ago (Vijgen, et al. 2005). Viruses of lineage B (LinB, *Sarbecovirus* subgenus, SARS-CoV-1 and SARS-CoV-2) were introduced through more recent independent zoonotic events (Li, Shi, et al. 2005; Vijaykrishna, et al. 2007; Andersen, et al. 2020; Boni, et al. 2020; Banerjee, et al. 2021). After its emergence in 2002 (Peiris, et al. 2003), SARS-CoV-1 spread to >20 countries in six months, resulting in a short-lived but severe outbreak characterized by sustained human-to-human virus transmission (Cheng, et al. 2007). Even though the circulation of SARS-CoV-1 in humans was terminated, there is evidence that the virus adapted to transmission among humans (Chinese SARS Molecular Consortium 2004). Since its emergence in 2019 (Zhou, et al. 2020), SARS-CoV-2 has displayed highly efficient human-to-human transmission resulting in global spread, with an apparently low rate of adaptive change in humans during the early stage of the pandemic (MacLean, et al. 2020). Although MERS-CoV (*Merbecovirus* subgenus; Fehr, et al. 2017) can also infect humans, MERS-CoV outbreaks so far have been the result of independent zoonotic events characterized by limited transmission chains (Fehr, et al. 2017; WHO 2021), and

MERS-CoV has not yet shown signs of ongoing adaptation to the human host (Kim, et al. 2016; Fehr, et al. 2017).

Continuous circulation of OC43 and HKU1 in humans has been accompanied by ongoing host-specific adaptation, a process that is at an early stage for SARS-CoV-2. If SARS-CoV-2 becomes endemic in humans (Shaman and Galanti 2020), it will similarly need to overcome the selective pressures exerted by collective immune responses of the human host population (Kissler, et al. 2020). In an attempt to shed light on the possible existence of convergent adaptive evolution across human betacoronaviruses, we undertook a comparative evolutionary analysis of four human-infecting virus species: HKU1, OC43, SARS-CoV-1 and SARS-CoV-2. As MERS-CoV infections only result in short-lived human-to-human transmission chains, it was excluded from our analysis. We identify 30 mutations at sites displaying evidence for homoplasy and/or stepwise evolution across the different virus species studied. Using a comparative *in silico* approach, we find that four mutations (18121 [exonuclease/27], 21623 [spike/21], 21635 [spike/25] and 23948 [spike/796], in SARS-CoV-2 genome coordinates) additionally display evidence of being selectively relevant and localize near known functional surfaces of non-structural proteins in Orf1ab (nsp14) (Ma, et al. 2015), the spike protein S1 subunit, and the virus fusion machinery in S2, respectively.

RESULTS

Conserved and variable sites across HKU1, OC43, SARS-CoV-1, and SARS-CoV-2

The LinA (HKU1 and OC43) and LinB viruses (SARS-related viruses, SARS-CoV-1 and SARS-CoV-2) (International Committee on Taxonomy of Viruses 2012) were consistently identified in all phylogenetic trees, in agreement with previously published phylogenies (Figure 2) (Woo, et al. 2006; Woo, et al. 2010; Lau, et al. 2011; Oong, et al. 2017; Zhu, et al. 2018; Bedford 2020). We also identified the previously described genotypes for HKU1 (A-C) and OC43 (A-H) (Woo, et al. 2006; Oong, et al. 2017) (Supplementary Data 1). We found that 2.2% of all homologous sites in the Orf1ab+S alignment (205/8962 codons) corresponded to non-synonymous amino acid changes shared between virus species, *i.e.,* appearing in any of the SARS-related viruses and in HKU1 and/or OC43. For Orf1a, 2.7% of sites (129/4774 codons within the Orf1a alignment) fell within this category, whilst for Orf1b 0.9% of sites (25/2623 within the Orf1b codon alignment) were identified as shared. The highest proportion of shared mutations was identified within Orf S (3.2% of all sites, 48/1457 codons within the alignment).

The S protein sequence alignment showed a greater proportion of variable sites relative to conserved sites across virus species, indicating a low degree of sequence conservation for the S gene, characteristic of coronaviruses (Li F, 2012). Only 16% of homologous sites in the alignment were conserved (243/1457 within S codon alignment), whilst the remainder (84%) were variable (Supplementary Data 2). The highest proportion of conserved sites was found within the S2 domain, presumably reflecting functional constraints and conservation of the viral membrane fusion machinery across virus species (Bosch, et al. 2003). A greater number of variable sites was observed within S1. Conserved sites were mostly observed within the $S1^A$ domain (also known as N-terminal domain, NTD) compared to the $S1^B$ domain, showing no conserved sites across virus species. This is likely attributable to the differences in receptor engagement mediated by the S1 subunit between the LinA and Lin B viruses, as the LinA viruses use domain $S1^A$ to interact with sialoglycan-based receptors, whilst LinB viruses use the S1 ^{B} domain to interact with the human 'angiotensin-converting enzyme 2' receptor (ACE2) (Hulswit, et al. 2019; Lan, et al. 2020). Additionally, a conserved residue 'R' at site 685 within the S1/S2 cleavage site (numbering according to the SARS-CoV-2 protein, codon sites 23615-23617) was found to be shared across all virus species (Supplementary Data 2), reflective of a conserved proteolytic maturation for the spike protein in the lifecycle of coronaviruses (Millet, et al. 2015).

In contrast to inter-species analysis, a high degree of sequence conservation within single species was observed (Figure 3). In general, the vast majority of sites were conserved and distributed predominantly within the membrane proximal S2, whilst variable sites were fewer and tend to be predominantly distributed within the membrane distal S1 subunit of the protein structures. The predominance of variable sites within S1 and conserved sites within S2 is most evident for the LinA viruses (HKU1 and OC43) and less so for LinB viruses (SARS-related viruses), for which the distribution of sites across the Orf S seems to be more homogeneous (Figure 3).

Identifying putative homoplasy and/or stepwise evolution

Not all variation observed for homologous sites may be reflective of common evolutionary trajectories across virus species. Thus, among the variable sites identified in the section above, we searched for those sites displaying putative homoplasy and/or stepwise evolution across virus species (Figure 1, see Methods section 3). We identified 30 mutations (representing 0.3% of all sites within the Orf1ab+S codon alignment) at sites displaying patterns indicative of homoplasy and/or stepwise evolution. Two of these sites were found within Orf1a, nine within Or1b, and nineteen within S (Table 1). Ancestral reconstructions for amino acid evolution patterns are shown for three illustrative examples in Figure 4.

Codon sites 18121-18123 (designated here as 18121 to indicate the start of the codon) in Orf1b/nsp14 correspond to the amino acid state ‘S’, which is homoplasic between HKU1 genotype B, SARS-CoV-1 and SARS-CoV-2 (Table 1, Figure 4, Supplementary Data 3). Codon sites 21623-21626 (designated here as site 21623) in Orf S correspond to the amino acid state ‘R’, which is homoplasic between SARS-CoV-2 and OC43 genotypes D, F, G and H. This site also exhibits the amino acid state ‘I’ present in a small SARS-CoV-2 cluster (represented by isolate SARS_CoV_2|PHWC-252C3|Human|Wales|2020, belonging to Pango lineage B.1.1.237) (O’Toole, et al. 2021). This state is homoplasic between this small SARS-CoV-2 cluster and OC43 genotypes E and H. Thus, site 21623 shows both evidence for homoplasy across different virus species (*i.e.* within OC34 and SARS-CoV-2), and stepwise evolution within a single virus species (in OC43, represented by the sequential amino acid changes V→ I→ K→ R, with no reversions observed to date to the immediate ancestral state). Comparably, codon sites 21635-21637 (designated here as site 21635) in Orf S correspond to the amino acid state ‘P’, also homoplasic between SARS-CoV-2 and OC43 genotypes D, F, G and H. This site again shows both evidence for homoplasy across different virus species (i.e. within OC34 and SARS-CoV-2), and stepwise evolution within a single virus species (in OC43, represented by the sequential amino acid changes V→P with no reversions) (Table 1, Figure 4, Supplementary Data 3).

Codon sites 23948-23950 (designated here as site 23948) in Orf S correspond to amino acid state ‘D’, present in all virus species with the exception of the OC43 viruses that have an ‘N’ at this site. Of interest, SARS-CoV-1 at this site shows a change from state ‘D’ (present in the earlier isolates) to ‘Y’ (appearing in the later isolates) (Table 1, Figure 4, Supplementary Data 3). Thus, this site displays evidence for stepwise evolution within SARS-CoV-1. Finally, codon sites 24614-24616, 24620-24623 and 24632-24635 (designated here as sites 24614, 24620 and 24632) in Orf S correspond to amino acid states ‘I’, ‘A’ and ‘L’. These amino acid states at three separate sites are homoplasic between HKU1 genotype B and the SARS-CoV-1 and SARS-CoV-2 viruses (Supplementary Data 3).

Estimating and comparing positive selection across virus species

Using a genome-wide comparison of dN/dS estimates across different virus genome regions (see Methods section 4), we detected evidence for positive selection in the complete Orf1b and S regions of the SARS-CoV-2 genome compared to the other virus genomes. Specifically, episodic diversifying selection analysis revealed positive selection in 5/14 non-recombinant fragments (three in Orf1b and two in Orf S, for details see https://observablehq.com/@spond/beta-cov-analysis). When comparing the alignment for all virus species and looking for evidence of selection across sites/internal branches of the tree (see Methods section 4), we found that 0.7% of all sites (67 codons in the Orf1ab+S alignment) were inferred to be under episodic diversifying positive selection (scored under MEME p≤0.05 as PSS, positively-selected sites) (Supplementary Table 3), and an additional 5% (461 of the codons in the Orf1ab+S alignment) were inferred to be under pervasive negative selection (scored under FEL p≤0.05 as NSS, negatively-selected site). For the sites categorized under homoplasy/stepwise evolution under our pipeline, 19048, 21623, 21635, 22124 and 23048 were scored as PSS. Sites 21623 and 21635 were inferred as PSSs along the branches ancestral to HKU1, OC43 and SARS-CoV-2, sites 19048 and 22124 were inferred as PSSs along the OC43 branches, and site 23048 was inferred as a PSS along the HKU1 branch ($p < 0.05$) (Table 1, Supplementary Table 3).

Using the Contrast-FEL method to detect differential positive selection across branches separating different virus lineages, we found that 36 sites (0.4%) across the alignment/tree were scored to be under differential selective pressure between some or all the different viral clades. For sites categorized under homoplasy/stepwise evolution in our pipeline, analysis under branch and site models (MEME; see Methods section 4) revealed that site 18121 is under selection for the HKU1 clade/branch compared to LinB (SARS-related viruses), in agreement with our observation for this site being homoplasic between HKU1 genotype B and SARS-CoV-1 and SARS-CoV-2 (Table 1, Figure 4). Site 23948 was also inferred to be under positive selective for the SARS-CoV-1 clade/branch compared to other virus species (Supplementary Table 3, Table 1).

We subsequently mapped the identified PSS and NSSs onto the SARS-CoV-2 S protein structure. For the 22 PSSs identified in S, 18 were located within S1 (11 in $S1^A$, 5 in $S1^B$, 1 in $S1^C$ and 1 in $S1^D$), whilst the 4 remaining PSSs mapped onto S2. For the 82 NSSs identified in S, 46 mapped onto S1 (18 in $S1^A$, 21 in $S1^B$, 3 in $S1^C$ and 4 in $S1^D$), and remaining 36 were found within S2 (Supplementary Figure 2).

Comparing positive selection with ongoing selection in the SARS-CoV-2 population

We compared our abovementioned results with the selection analysis available for currently sampled SARS-CoV-2 genomes as of February 2021 (available at https://observablehq.com/@spond/evolutionary-annotation-of-sars-cov-2-covid-19-genomes-enab) (Kosakovsky Pond 2021). Of the 30 mutations identified here, 16 showed evidence of being under positive or negative selection currently within SARS-CoV-2, with 13 of these sites mapping directly onto potential T cell epitopes derived from HLA class I and HLA-DR binding peptides in SARS-CoV-2 (Campbell, et al. 2020; Nelde, et al. 2021) (Table 1). Among the sites identified here as displaying homoplasy or stepwise evolution, sites 7478, 21614, 23948, 24620 and 25166 were also inferred to be under ongoing positive selection within SARS-CoV-2, whilst genome sites 21635, 24863, and 25037 were identified to be under negative selection within SARS-CoV-2 (Table 1). Some of the amino acid changes observed at these sites within the SARS-CoV-2 population had already occurred in other betacoronavirus species. An example of this is site 21614, with an observed amino change 'L' within the sequences sampled for our study (see Methods section 1), and an observed change L→F later occurring within the SARS-CoV-2 population sampled. Amino acid state 'F' had been already observed in the OC43 and SARS-CoV-1-like viruses (Table 1). Contrastingly, other amino acid changes at these sites currently observed for SARS-CoV-2 are not seen within other human betacoronaviruses in our data set. However, we note that some of these newly observed amino acid changes may represent evolutionary dead-ends within the long-term evolution of the virus population, as exemplified by the early emergence and extinction of different SARS-CoV-2 lineages through time and space (van Dorp, Acman, et al. 2020; van Dorp, Richard, et al. 2020).

Relating the locations of identified mutations to known functional sites on reported protein structures.

We found that 8 of the 30 identified mutations that exhibited homoplasy and/or stepwise evolution are structurally proximal to regions of known protein function (not considering whether these if sites were shown to be selectively relevant). The main results are summarised below and in Table 2:

- **Orf1ab**

The Orf1ab gene encodes 16 non-structural proteins (nsp1-16), all of which have a functional role related to viral RNA synthesis and processing (Wu, et al. 2020). Site 18121 in Orf1ab corresponds to the 'S' to 'A' mutation identified as homoplasic in some HKU1 and the SARS-related viruses (for details, see results section 'Identifying putative homoplasy and/or stepwise evolution'), and inferred to be under positive selection for the HKU1, OC43 and SARS-CoV-2 branches (Figure 4, Supplementary Data 3, Table1). This site corresponds to residue 28 located within the exonuclease ExoN domain of the nsp14 protein (numbering according to the SARS-CoV-1 protein) (Table 2, Figure 5a). Nsp14 protein functions as a methyltransferase and is involved 5′-capping to the viral mRNA (Ma, et al. 2015). The cap core structure is essential for viral mRNA transcription, but is also implicated in protecting the 5′-triphosphate from activating the host innate immune response (Wang, et al. 2015). This 'S' to 'A' amino acid change is expected to result in the loss of an intra-protein hydrogen-bond (formed with the main chain of residue T25 of nsp14, Figure 5a) within nsp14's interaction surface with its activator nsp10 (Ma, et al. 2015) (as assessed by PISAebi; http://www.ebi.ac.uk/pdbe/prot_int/pistart.html), potentially modulating this protein-protein interaction.

Site 20344 also in Orf1ab corresponds to a non-conservative 'H' to 'Y' mutation observed to be homoplasic in some HKU1 and SARS-related viruses (Supplementary Data 3, Table1). This site corresponds to residue 243 in nsp15 (numbering according to the SARS-CoV-2 protein) (Table 2, Figure 5b), and is located within the NendoU catalytic domain of the endoRNAse. This domain specifically targets and degrades viral mRNA polyuridine sequences to prevent host immune sensing (Hackbart, et al. 2020). However, this mutation is distal (~12 Å) to the nucleotide binding pocket of the active site, and the impact of this change upon endoRNAse activity, if any, is unknown.

- **S1**

The S1 subunit of the spike (S) protein mediates attachment of the virus to the host cell (Hulswit, et al. 2016). The LinA viruses (HKU1 and OC43) recognize glycan-based cell receptors carrying 9-*O*-acetylated sialic acids and receptor recognition is accomplished via two hydrophobic pockets separated by a conserved Trp (W) located within the S1^{A} region of the protein (Hulswit, et al. 2019, Tortorici and Veesler 2019). In contrast, the receptor-binding site for LinB viruses (SARS-CoV-1 and SARS-CoV-2) consists of an extended loop located within the S1^{B} domain of the protein (Li, et al. 2005; Lan, et al. 2020; Shang, et al. 2020). Despite a limited level of sequence conservation amongst the contact residues in the RBD between the SARS-CoV-1 and SARS-CoV-2 viruses, both recognize ACE2 for cell entry (Lan, et al. 2020).

Sites 21623 and 21635 in Orf S were identified as homoplasic across lineages of distinct virus species (certain OC34 and SARS-CoV-2 lineages) and exhibit stepwise evolution within a single virus species (OC34). These sites were also inferred to be under positive selection for the HKU1, OC43 and SARS-CoV-2 branches (Figure 4, Supplementary Data 3, Table1). Site 21623 maps to domain $S1^A$, and corresponds to the non-conservative mutation 'R' to 'I' at residue 21 in the SARS-CoV-2 S protein, and to residue 29 of the OC43 S protein. Site 21635 is situated 4 residues downstream of 21623, and corresponds to residue 25 in the SARS-CoV-2 S protein, and to residue 33 of the OC43 S protein (Table 2, Figure 6). For OC43, these residues located within a loop neighbouring the hydrophobic pockets in $S1^A$ instrumental for receptor recognition, and changes to this region have been previously shown to modulate receptor affinity (Hulswit, et al. 2019). Given residue location within the protein and a highly variable evolutionary pattern exhibiting both homoplasy and stepwise evolution (Figure 4), it seems possible that this site may reflect antigenic drift shaped by the selective pressure exerted by the host immune response (Kistler and Bedford, 2021), and might be of particular relevance for LinA viruses. In the case of SARS-CoV-2, two variants of concern (B.1.351 and P.1) have independently accumulated mutations at this region (Faria, et al. 2021; Tegally, et al. 2021). Given that LinB viruses engage their ACE2 receptor via domain $S1^B$, mutations at this particular site in SARS-related viruses may reflect relaxed constraints within the local protein surface, unrelated to receptor functionality.

- **S2**

Binding of the virus to the cell surface is followed by fusion of the viral and host membrane to release the virus genome into the cell. The S protein needs to be primed in order to mediate membrane fusion, and this is achieved through cleavage by host cell proteases (Xia, et al. 2020). Betacoronavirus spike proteins contain a conserved cleavage site at the S1-S2 junction (consensus RRAR|S in SARS-CoV-2), and an additional R|S site within the S2' subunit, termed the S2' cleavage site. The S2 subunit of the S protein harbours the protein's fusion machinery, with the characteristic structural features of class I fusion proteins (Bosch, et al. 2003; Benton, et al. 2020), including a fusion peptide that is inserted into the host membrane during the fusion process that triggers major conformational changes within the central helix to facilitate merger of the virus-host membranes

Site 23948 within S displays evidence for stepwise evolution within the SARS-CoV-1 viruses (Figure 4, Supplementary Data 3, Table1), and corresponds to a non-conservative ‘D’ to ‘Y’ mutation at residue 778 within the S2 subunit of SARS-CoV-1, and an ‘N’ at residue 890 of OC43 S (Table 2). This protein region is located immediately upstream of the S2' cleavage site, key for the release of the fusion peptide (Millet, et al. 2015). The amino acids between site 23948 and the fusion peptide form a loop with some degree of variability across different betacoronavirus species (Figure S1), suggesting relaxed functional constraints and/or that flexibility is important for functionality within the local protein region. In agreement with this observation, the corresponding region remains unresolved in the HKU1 structure, indicating local protein flexibility. Due to the position of this site near the S2' cleavage site, it is possible that changes to this region may affect the maturation and/or activity of the spike protein. Mutations within this region have been detected in the currently circulating SARS-CoV-2 population (Table 1), whilst evidence for positive selection at this site may reflect ongoing adaptation.

Finally, sites 24614, 24620 and 24632 correspond to the conservative mutation ‘V’ to ‘I’ at residue 1018, ‘F’ to ‘A’ at residue 1020, and the non-conservative mutation ‘L’ to ‘R’ at residue 1024 (numbering according to the SARS-CoV-2 protein) in S2, observed to be homoplasic for some HKU1 and SARS-related viruses (Table 1). These three residues are within close proximity of each other and are positioned within the central helix of S2 (Table 2, Figure 6), a region for which conformational rearrangements may facilitate membrane fusion (Kirchdoerfer, et al. 2016; Pallesen, et al. 2017; Kirchdoerfer, et al. 2018). Again, given their proximity, it is possible that changes to these amino acids may alter the fusogenic functionality of the S2, as has been observed for other changes at central helix domain of coronaviruses (Hulswit, et al. 2016).

DISCUSSION

Current genomic studies on betacoronavirus mutational patterns have focused mostly on the intra-species variation of SARS-CoV-2, yet the vast majority of the observed variation in the SARS-CoV-2 population is not expected to be related to adaptive processes (van Dorp, Acman, et al. 2020; van Dorp, Richard, et al. 2020). In addition, emerging mutations in the sampled SARS-CoV-2 virus population may also reflect mutational rate biases inherent of the viral genome (*i.e.* with C→ U transitions being more likely and resulting a high degree of apparent homoplasy in synonymous sites), or even systematic errors related to sequencing and bioinformatic methodologies (De Maio, et al. 2020; Worobey, et al. 2020; Wang, et al. 2021). Thus, the comparison of mutations co-occurring across human-infecting betacoronaviruses (OC43, HKU1, SARS-CoV-1, SARS-CoV-2) has the potential to improve our understanding on the common mutations associated with betacoronavirus adaptation to the human host.

Within individual virus species, the majority of variable sites were observed within Orf S, encoding for the main viral antigenic protein (Yoshimoto 2020). Analysis of the distribution of variable versus conserved sites on the S protein structures of the different virus species showed more variable sites within S1 compared to S2, with this pattern being particularly evident for the endemic LinA viruses (OC34 and HKU1). Despite major differences in the receptors and sites used for engagement between the LinA and LinB viruses (Hulswit, et al. 2019; Lan, et al. 2020), similar structural constraints on the S1 subunit domains and exposure to comparable immune-derived selective pressures may explain the occurrence of potentially homoplasic mutations in S1 across distinct virus species. Thus, we speculate that a key evolutionary force driving fixation of mutations in S1 may arise from the host humoral immune response (Kistler and Bedford, 2021; Li, et al. 2019; Dejnirattisai, et al. 2021). Due to the recent zoonotic introduction of SARS-CoV-2, the effects of immune-derived selective forces may be more pronounced for the endemic viruses, for which antigenic drift (Kistler and Bedford, 2021) may be associated with the emergence of viral genotypes that result in recurrent infections (Dejnirattisai, et al. 2021).

Through the genomic comparison across virus species, we find that only four sites (i) display evidence of homoplasy and/or stepwise evolution, (ii) show evidence to be evolving under positive selection, and (iii) are proximal to regions of established protein function. The emergence of mutations observed in the non-structural genes may be related to adaptation for a more efficient replication in the human host (Menachery, et al. 2017), such as those mutations identified here and observed to occur in the Orf1ab/Exonuclease domain of nsp14, and in the Orf1ab/endonuclease domain of nsp15 (sites 4265, 18121 and 20344, respectively). In contrast to the antibody-mediated immune response expected to be a key driver of evolution for the spike protein (Kistler and Bedford, 2021; Li F, 2016)., immune selection within the Orf1ab non-structural genes can be driven by impairment of interferon and cytokine signalling cascades and antigen presentation suppression, among other cellular pathways that can be affected by viral immune hijacking mechanisms (Wang, et al. 2015; Hackbart, et al. 2020; Taefehshokr, et al. 2020; Yuen, et al. 2020). Therefore, these may also arise through immune-derived selective pressures. However, it is important to note that the different selective pressures are not mutually exclusive, and that a single mutational change can have pleiotropic effects on multiple phenotypes and components of virus fitness (Polster et al, 2016).

Detecting molecular evolution related to adaptation in human-infecting betacoronaviruses, as represented by homoplasy and/or stepwise evolution, can be hampered by the long divergence times between the virus species studied, which can limit alignment confidence. This divergence is also reflected by major differences in the basic biology of the viruses, such as receptor usage, and thus restrict the conclusions that can be drawn from this comparative analysis. Other limitations of our study include (i) the low availability of genomes sampled longitudinally through time (especially for HKU1 and SARS-CoV-1), and (ii) the low genetic variability for SARS-CoV-2 (Rausch, et al. 2020), which restrict the statistical power to detect mutations likely to denote adaptation (van Dorp, Richard, et al. 2020). Further, it is not possible to be certain that the mutations identified by our pipeline are indeed adaptive, as apparent homoplasy and stepwise evolution can also result from nonadaptive evolutionary processes such as genetic drift, mutational hitchhiking, and mutational rate biases (Delport, et al. 2008; Pond, et al. 2012; De Maio, et al. 2020; Simmonds 2020; Wang, et al. 2021). Further genomic surveillance of these viruses, as well as other beta-coronaviruses that may potentially emerge, will be necessary to confirm that the mutational panel presented here may represent common pathways reflecting betacoronavirus adaptation to the human host. The mutations identified here may be informative on ongoing adaptation of betacoronavirus circulating in the human population, but require further experimental evidence to interpret their adaptive effect and biological significance.

MATERIAL AND METHODS

1. Data collation

For HKU1, OC43 and SARS-CoV-1, complete virus genomes from sampled from human across all geographical regions and collection years were downloaded from the Virus Pathogen Resource (ViPR-NCBI 2021) (Supplementary Data 4). Sequences were removed from the datasets if (i) they were >1000nt shorter than full genome length, (ii) they were 100% similar to any other sequence, or (iii) if >10% of site were ambiguities (including N or X). A total of 53 HKU1, 136 OC43 and 40 SARS-CoV-1 sequences were retained for analysis. We aimed to limit genetic diversity of the sampled SARS-CoV-2 virus population to the first wave of the pandemic, in order to better reflect its recent zoonotic introduction into the human population (MacLean, et al. 2020).Thus, for SARS-CoV-2, ~23000 full genomes sampled worldwide before May 2020 and available in the GSAID platform (GSAID 2021) were downloaded and aligned as part of the public dataset provided by the COG-UK consortium (COG-UK Consortium 2021) (Supplementary Data 4). To make analyses computationally feasible, the original SARS-CoV-2 alignment was subsampled to 5% of its original size, removing sequences using the criteria above. A total of SARS-CoV-2 1120 sequences were retained. For all virus species, we focused only on sequences derived from human hosts within our initial sequence sampling scheme, so that identified mutations reflect host-specific adaptation processes.

2. Initial phylogenetic analysis

Only the main viral ORFs (Orf1ab and S) were used for phylogenenetic analysis, as these are shared among the four viral species used in this study (HKU1, OC43, SARS-CoV-1 and SARS-CoV-2). These ORFs code for proteins essential for virus function, such as the genome replication machinery and other essential non-structural proteins (Orf1ab), and the receptor engagement and the virus-host membrane fusion apparatus (S) (Yoshimoto 2020). For each virus species, individual ORFs were extracted, translated to amino-acids, and aligned using MAFFT v7.471(Katoh and Standley 2013). UTRs and short non-coding intergenic regions were excluded. The virus species and accession numbers used for this work are listed in Supplementary Data 1. Aligned ORFs were concatenated to generate an Orf1ab+S alignment for each virus species. Concatenated alignments were combined to generate a global dataset that was re-aligned at amino acid level using a profile-to-profile approach following taxonomic relatedness (Wang and Dunbrack 2004).

Although recombination is known to occur among betacoronaviruses (Woo, et al. 2006; Su, et al. 2016; Oong, et al. 2017), recombinant sequences were not removed for this initial step of the analysis, as it was important first to identify general evolutionary patterns and to detect recombinant isolates that may display relevant mutations. However, recombinant sequences were further removed for detailed phylogenetic analysis (see Methods sections 6 and 7). In total, 1314 sequences were used to generate an alignment with 26883 columns. Maximum likelihood phylogenies for the individual and global alignments were estimated using RAxML v8 (Stamatakis 2015) under a general time reversible nucleotide substitution model with gamma-distributed among-site rate variation (GTR+G) and branch support assessed using 100 bootstrap replicates. All trees were midpoint-rooted, and general patterns of ancestry among virus species were validated by comparing to previously published phylogenies (Woo, et al. 2006; Woo, et al. 2010; Lau, et al. 2011; Oong, et al. 2017; Zhu, et al. 2018; Bedford 2020).

3. Identifying evidence for homoplasy and/or stepwise evolution

Following the pipeline described in Escalera et al. (Escalera-Zamudio, et al. 2020), we identified all variable sites within the global alignment representing non-synonymous amino acid changes occurring in ≥1% of the sampled sequences. Variable sites were identified by comparing homologous sites across sequences to a consensus generated under a 95% threshold using the 'Find Variations/SNPs' function in Geneious Prime v2020.0.4 (Kearse, et al. 2012). Ancestral amino acid states at these sites were inferred for nodes in the RAxML tree (global ML tree) using TreeTime (Sagulenko, et al. 2018) under a ML approach (RAS-ML) and a time-reversible model (GTR) for state transitions. In parallel, conserved amino acid states within the alignment were identified and extracted by using a profile-to profile alignment comparison of global consensus in amino acid sequences generated under a 99% threshold, and re-aligned using MAFFT v 7.471 (Katoh and Standley 2013) (Supplementary Data 2).

The resulting 6681 variable amino acid sites were mapped onto the global ML tree (referred here as Ancestral Reconstruction Trees, ARTs) and analysed visually. We further developed a computational algorithm to sort ARTs according to whether they evidenced patterns of molecular homoplasy and/or stepwise evolution. Homoplasy can occur at an interspecies or intraspecies level, and is defined here as any given amino acid change occurring in at least one internal node of a given virus species, which is also present in at least another internal node of the same or other virus species (Figure 1). Nodes with the same amino acid state must not share direct common ancestry. Stepwise evolution can occur only at an intraspecies level, and is defined here as those sites subjected to directional mutational change involving at least two states (A→B) and occurring sequentially towards a local fitness optimum, but without immediate reversions (B→A) (see Figure 1 and Methods Section 3). A full description for the basic steps in the algorithm, including a schematic representation and validation data is available in the Supplementary Information (Supplementary Text 1, Supplementary Figure 3 and 4).

4. Estimating dN/dS

Using the global alignment and ML tree, we estimated dN/dS (or ω, defined as the ratio of non-synonymous substitution rate per non-synonymous site to the synonymous substitution rate per synonymous site) using both site, branch and branch-site dN/dS models: Mixed Effects Model of Evolution (MEME), Fixed Effects Likelihood (FEL), and the fixed effects site-level model (Contrast-FEL) (Kosakovsky Pond and Frost 2005; Murrell, et al. 2012; Kosakovsky Pond, et al. 2020). The concatenated codon alignment for Orf1ab/S was partitioned into 14 putatively non-recombinant regions using the Genetic Algorithm for Recombination Detection (GARD) (Kosakovsky Pond, et al. 2006) and all subsequent analyses were conducted on partitioned data. The dN/dS models use the GTR component for the nucleotide evolutionary rate, so biased mutation rates are handled. Testing for selection was restricted to internal branches of the phylogeny to mitigate the inflation in dN/dS due to unresolved or maladaptive evolution in individual hosts (Pond, et al. 2006). Importance of biochemical properties at selected sites were assessed under the PRoperty Informed Models of Evolution method (PRIME) (HYPHY 2013). Genome-wide comparison of dN/dS estimates across different viral genome regions was performed using the Branch-Site Unrestricted Statistical Test for Episodic Diversification method (BUSTED) (Murrell, et al. 2015). An interactive notebook with the full selection analysis results is available at https://observablehq.com/@spond/beta-cov-analysis.

5. Mapping mutations onto betacoronavirus protein structures

To relate the positions of amino acid changes to regions of known protein function, the mutations identified in section 3 were mapped using PyMOL v 2.4.0 (https://pymol.org/2/) onto the available protein structures listed in Table 2 and in Data Availability section. N-linked glycosylation sites in S protein sequences were identified by searching for the N-[not P]-[S or T] consensus sequence (Watanabe, et al. 2019). None of the mutations identified in this study resulted in generation or deletion of N-linked glycosylation sequons. In parallel, conserved and variable sites for single virus species, and variable sites evidencing homoplasy and/or stepwise evolution across virus species, were mapped onto published S protein structures for the four different betacoronaviruses (Figure 6, Supplementary Figure 2). To compare dN/dS distributions between specific domains of the S protein within and across virus species, sites inferred to be under positive and negative selection were mapped onto S protein structures (Supplementary Data 2).

6. Resampling datasets

We undertook further analysis to detect if the mutations identified within human-infecting betacoronaviruses were also present in genomes of the most closely related viruses derived from non-human hosts within LinA and LinB. First, we subsampled the SARS-CoV-2 sequences from the global alignment in a phylogenetically-informed way to reduce over-representation. Based on the ML tree, only the most basal SARS-CoV-2 sequences and those displaying the mutations of interest were retained, whilst randomly subsampling the rest to preserve overall tree structure. Using this approach, a total of 70 SARS-CoV-2 sequences were retained. We then added the most closely related viruses from non-human host, using the betacoronavirus dataset at https://github.com/blab/beta-cov to retrieve sequences based on percentage identity and phylogenetic clustering (Bedford 2020). Adding related non-human host genomes will also mitigate the effect of long branches separating virus species (*i.e.* LinA and LinB). Recombinant sequences identified were removed using ClonalFrameML (Didelot and Wilson 2015), whilst non-recombinant fragments were verified using GARD (Kosakovsky Pond, et al. 2006). Minor recombination events (with fewer than 10 sequences) were detected amongst the OC43 viruses (within the C, E, F and G genotypes) (Oong, et al. 2017). None of the recombinant sequences displayed any mutations of interest, and thus were excluded from further analysis (see below). In total, 430 non-human virus sequences were added to the dataset, yielding a total of 686 sequences that were realigned under a progressive profile-to-profile approach based on taxonomic relatedness, resulting in an alignment with a total length of 27392 bases. The resulting alignment was used to estimate a new ML tree using the approach in Methods Section 2 (Figure 2). MERS virus sequences were included only for tree rooting purposes.

7. Reconstruction of amino acid evolution for selected sites

For examplary mutations with cumulative evolutionary and structural evidence of being potentially informative about adaptation processes (genome sites 18121, 21623 and 23948) (Table 1), we used the resampled dataset to infer ancestral states under a Bayesian framework. For this, we first estimated an MCC tree from the resampled codon alignment (Methods Section 6) using a SRD06 substitution model (Shapiro, et al. 2006) and a strict molecular clock fixed to 1. For each site of interest, coded amino acid traits were mapped onto the nodes of the MCC tree by performing reconstructions of ancestral states under an asymmetric discrete trait evolution model (DTA) in BEAST v1.8.4 (Lemey, et al. 2009; Suchard, et al. 2018). The DTA model was run using a Bayesian Skygrid tree prior for 100X10^6 generations and sampled every 10000 states until all DTA-relevant parameters reached an ESS >200.

FIGURE LEGENDS

Figure 1. Patterns of evolution potentially informative of adaptation

Patterns of molecular evolution that may be informative of adaptation include homoplasy (also called parallel evolution) and stepwise evolution. (a) Homoplasy can occur at an interspecies or intraspecies level, and is defined here as any given amino acid change occurring in at least one internal node of a given virus clade, and which is also present in at least another internal node of the same or another virus clade. Nodes with the same amino acid state must not share direct common ancestry. (b) Stepwise evolution can occur only at an intraspecies level, and is defined here as those sites subjected to directional mutational change involving at least two states (A→B) and occurring sequentially towards a local fitness optimum, but without immediate reversions (B→A). Homoplasy and Stepwise evolution are not mutually exclusive events and may co-occur (see Supplementary Text 1). Here we study the amino acid evolution patterns denoting homoplasy and stepwise evolution for mutations shared across the LinB viruses (SARS-CoV-1 and SARS-CoV-2) and any of the LinA virus clades (HKU1 and/or OC43).

Figure 2. Phylogeny of human-infecting betacoronaviruses

Maximum likelihood tree estimated from the Orf1ab+S alignment summarizing the evolution of four human-infecting betacoronaviruses: HKU1, OC43, SARS-CoV-1 and SARS-CoV-2. Both the LinA (*Embecovirus* subgenus, HKU1 and OC43) and LinB (*Sarbecovirus* subgenus, SARS-CoV-1 and SARS-CoV-2) are shown. The most closely related animal virus isolates to each group (where available) have been included. The three genotypes of HKU1 (A, B and C) and the eight different genotypes (A–H) of OC43 are also shown (for details see Supplementary Data 3). MERS-related viruses representing the betacoronavirus Lineage C are included for tree rooting purposes.

Figure 3. Distribution of highly conserved/variable sites within S across different virus species.

(a) Top (upper panel) and side view (bottom panel) of a cartoon representation of the multidomain architecture of the trimeric SARS-CoV-2 S ectodomain (PDB: 6ZGI). The S1 subunit is divided into S1^A (cream), S1^B (teal), S1^C (orange), and S1^D (blue) domains, while and the S2 subunit is indicated in grey. (b) Top-down and side views of sphere-based representations of trimeric betacoronavirus S protein ectodomains for the viruses studied here: SARS-CoV-2 (PDB: 6VXX), SARS-CoV-1 (PDB: 6ACC), OC43 (PDB: 6OHW) and HKU1 (PDB: 5I08). The sphere-based representation shows intra-species conserved (grey; conservation in ≥99% of sequences) and variable residues (blue; changes in ≥1% of sequences. Variable sites at an inter-species denoting homoplasy or stepwise evolution are shown in red (see Methods section 3). Residues that do not all in the abovementioned criteria are not shown. Asparagine residues representing N-linked glycosylation sequons are indicated in purple.

Figure 4. Reconstruction of amino acid evolution at selected sites

Maximum clade credibility (MCC) trees for three representative sites (18121, 21623 and 23948 in SARS-CoV-2 genome coordinates) that are under selective relevance and that co-localize to known functional surfaces of proteins. Illustrative reconstruction of ancestral states for these sites show amino acid evolution patterns that denote homoplasy and/or stepwise evolution. Different amino acid states at nodes are shown with circles whilst amino acid changes indicated with different colours. The posterior probabilities for a given amino acid state occurring at the specified node are indicated. Sites 18121 display evidence of homoplasy across species, site 21623 shows evidence of both homoplasy across species and stepwise evolution within single virus species (*i.e.* OC43), and site 23948 shows evidence of stepwise evolution within single virus species (*i.e.* SARS-CoV-1).

Figure 5. Mutations that co-localize to known functional sites on reported protein betacoronavirus structures.

(a) Cartoon representation of the SARS-CoV-1 nsp14-nsp10 protein complex (PDB: 5C8S) showing residue Ser^{28} (corresponding to site 18121 in SARS-CoV-2 genome coordinates) as a red sphere. This residue is located within the nsp14 ExoN domain (cream) and approximately 9 Å from the interface with nsp10 (light blue, the proximal residue Cys^{41} used to calculate the distance is indicated as a sphere). The distance between nsp14's Ser^{28} and the nsp10's Cys^{41} is annotated and indicated by a dashed black line. Zoomed-in panel: detailed representation of the intra-nsp14 hydrogen-bond between the side chain of Ser^{28} and the main chain of Thr^{25}. The side chain of Ser^{25} is indicated as a red stick and Thr^{25} is indicated in sticks and coloured according to atom (C, cream; O, red; N, blue). The hydrogen-bond is indicated as a dashed black line. (b) Cartoon representation of the SARS-CoV-2 nsp15 protein (PDB: 6WLC, grey), showing residue His^{243} (site 20344 in SARS-CoV-2 genome coordinates) as a red sphere. This residue is located approximately 12 Å from the nucleotide binding pocket of the active site of the endoribonuclease. The nucleotide ligand uridine-5'-monophosphate (UMP) is shown within the active site in a stick representation and coloured according to atom

(C, white; N, blue; O, red; P, orange). The distance between His^{243} and UMP is indicated in black with a dashed line. All proteins are shown with a transparent surface for clarity.

Figure 6. S protein structure of SARS-CoV-2 with mutations that exhibit homoplasy indicated.

Top-down (left) and side view (right) of a cartoon representation of the multidomain architecture of the trimeric SARS-CoV-2 S ectodomain (PDB: 6ZGI). The S2 subunit is highlighted in grey and the S1 ectodomain is divided into S1^{A} (highlighted in cream), S1^{B} (teal), S1^{C} (orange), and S1^{D} (blue) domains, following the colour scheme in Figure 3. Homoplasic mutations co-localizing to known functional surfaces (see Table 2) are indicated in the structure and coloured in groups: Arg^{21} (corresponding to site 21623 in SARS-CoV-2 genome coordinates, in green), Pro^{25} (site 21635, in green), Asp^{796} (site 23948, in yellow), Ile^{1018} (site 24614, in red), Ala^{1020} (site 24620, in red) and Leu^{1024} (site 24632, in red). All representations are shown with a transparent protein surface for clarity.

DATA AVAILABILITY

Taxa IDs and accession numbers for sequences used and GISAID acknowledgements for the COG dataset are provided in the Supplementary Data 4 file. All sequence data supporting the findings of this study are publicly available from GSAID/GenBank. PBD files used are listed as follows: S protein (HKU1 PDB:5I08, OC43 PDB:6OHW, SARS-CoV-1 PDB:6ACC and SARS-CoV-2 PDB:6VXX). Orf1a (SARS-CoV-1 nsp3 PDB:2W2G). Orf1b (SARS-CoV-2 nsp13 PDB:6XEZ, SARS-CoV-1 nsp14 PDB:5C8S and SARS-CoV-2 nsp15 PDB:6WLC) (Tan, et al. 2009; Ma, et al. 2015; Kirchdoerfer, et al. 2016; Song, et al. 2018; Tortorici and Veesler 2019; Walls, et al. 2020; Kim, et al. 2021). Full code for our pipeline is available as open source: https://github.com/nataliamv/SARS-CoV-2-ARTs-Classification. Full selection analysis is available at https://observablehq.com/@spond/beta-cov-analysis.

AUTHOR CONTRIBUTIONS

MEZ and OGP designed research. MEZ and RJGH performed research. MEZ, RJGH, BG, SKP, JT and LDP analysed data. NM developed the code for implementing the computational pipeline. OGP and TAB supervised data analysis. MEZ and RJGH wrote the manuscript, with comments from all authors.

COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENTS

MEZ is supported currently supported by Leverhulme Trust ECR Fellowship (ECF-2019-542). RJGH is supported by the European Molecular Biology Organisation (ALTF 869-2019). RJGH and TAB are supported by the Medical Research Council (MR/S007555/1). The Wellcome Centre for Human Genetics is supported by Wellcome Trust grant 203141/Z/16/Z. SKP is supported in part by the the NIH (grants AI134384, AI140970, GM110749) and the NSF (grant 2027196). BG is supported by the 2017 Universities of Academic Excellence Scholarship Program of the Secretariat for Higher Education, Science, Technology and Innovation of the Republic of Ecuador (ARSEQ-BEC-003163-2017) and by Universidad San Francisco de Quito. JT is supported by European Union's Horizon 2020 project MOOD (874850). We thank Dr. Louis Du Plessis for his help with the phylogenomic analysis. We thank Dr. Nicola De Maio for constructive comments on our analysis approach and on the manuscript.

Table 1. Potentially relevant sites across human-infecting betacoronaviruses

			Amino acid state observed									
SARS-CoV-2 genome coordinates [†]	ORF	Protein/ Residue [†]	Ancestral LinA	OC43	HKU1	Ancestral LinB	SARS-CoV-1	SARS-CoV-2[§]	Homoplasy (H)/ Stepwise Evolution (SWE)	Selection across species, PSS p-value [† #]	Selection in SARS-CoV-2, recent amino acid changes [¶]	Epitopes*
2557	Orf1a	nsp2 585	P	P	S	S	A/T	P/S	H/SWE			0
7478	Orf1a	nsp3 1587	N	S/N	N	N	T	N	H		PSS, N→S/D (OC43-like and new state)	0
16189	Orf1b	nsp12 917	D	D	E/D	E	E	E	H/SWE	Overall negative selection (FEL 0.02)		1
17809	Orf1b	nsp13 525	V	V	V/I	I	I	I	H			0
18121	**Orf1b**	**nsp14 28**	**A**	**A**	**A/S**	**S**	**S**	**S**	**H**	**Different overall positive selection (CF 0.022)**		**1**
18334	Orf1b	nsp14 100	D	D	E/D	E	D	E	H/SWE	Overall negative selection (FEL 0.004)		0
18442	Orf1b	nsp14 136	K	K	K/R	R	R	R	H			0
19048	Orf1b	nsp14 338	A	G/A	G	A	A	A	H/SWE	OC43 branch (MEME 0.035)		0
20344	Orf1b	nsp15 243	Q	Q	H/Y	H	H	H	H/SWE			2
20554	Orf1b	nsp15 313	N	N	S/N	S	S	S	H	Overall negative selection (FEL 0.04)		0
21400	Orf1b	nsp16 249	A	A	T/S	S	S	S	H/SWE			2
21614	Orf S	S1 18	F	F/I/L	I	L	F	L	H/SWE		PSS, L→F (OC43 and SARS-CoV-1-like)	1
21623	**Orf S**	**S1 21**	**V**	**R/V/K/I**	**K/Y/L**	**R**	**V**	**R/I**	**H/SWE**	**HKU1, OC43 and SARS-2 branches (MEME 0.047)**	**NSS, R→ I/K/T (OC43 and HKU1-like and new state)**	**1**
21635	Orf S	S1 25	V	P/V/S/ L/H	V/I	P	N	P	H/SWE	HKU1, OC43 and SARS-2 branches (MEME 0.048)	NSS, P→S and L (OC43-like)	0
21800	Orf S	S1 81	K	K	Q/K	D	G/D	D	SWE		PSS, D→Y/A/G (SARS-CoV-1-like and new states)	0
21863	Orf S	S1 102	Y	F/I/T	Y	I	V	I	H/SWE		PSS, I→V (SARS-CoV-1-like)	0
21920	Orf S	S1 120	V	V	V/I	V	I	V	H/SWE			0
21926	Orf S	S1 122	T	T	N/T	N	N	N	H/SWE	Overall negative selection (FEL 0.002)	NSS	0
22004	Orf S	S1 149	N	N/K	K/I	N	G	N	H		NSS, N→D (new state)	0
22124	Orf S	S1 189	D	T/D/N	D	H	H	N	H/SWE	OC43 branch (MEME 0.008)	NSS	0
22553	Orf S	S1 332	N	D/N	D/N	N	N	N	H/SWE			1
23048	Orf S	S1 497	S	A/G/S	D/S	G	G	G	H/SWE	HKU1 branch (MEME 0.044)		2
23948	**Orf S**	**S2 796**	**D**	**N**	**D**	**D**	**Y/D**	**D**	**SWE**	**Different overall positive selection (CF 0.031)**	**PSS, D → Y/G/H (SARS-CoV-1-like and new states)**	**0**

24614	Orf S	S2 1018	V	V	V/I	I	I	I	H	NSS	1
24620	Orf S	S2 1020	F	F	F/A/L	A	A	A	H	PSS, A→S/V (new states)	2
24632	Orf S	S2 1024	Q	Q	L/R	L	L	L	H/SWE		2
24863	Orf S	S2 1101	T	T	H/S	H	S	H	H/SWE	NSS, H→Y (new state)	1
25037	Orf S	S2 1159	Q	Q	Q/H	H	H	H	H	NSS, H→Y (new state)	0
25166	Orf S	S2 1202	D	D/Y	D/E	E	E	E	H	PSS, E→Q/G (new states)	0
25247	Orf S	S2 1230	V	V	V/M	M	M	M	H	PSS, M→I/T/ L (new states)	1

[†] Positions indicate the start of the codon for reference genome Wuhan-Hu-1 (NC_045512.2). Sites in bold refer to those represented in Figure 4.

[#] Sites/branches scored under MEME/FEL and Contrast-FEL (CF); CF tests for differences is selective pressures between clades

[¶] Available from https://observablehq.com/@spond/sars_cov_2_sites. Results representing virus diversity as of February 2021

[§] Representing virus diversity as of May 2020

* Potential T cell epitopes derived from HLA class I and HLA-DR SARS-CoV-2 binding peptides (Campbell, et al. 2020; Nelde, et al. 2021)

Table 2. Identified sites that are structurally proximal to regions of known protein function

SARS-CoV-2 reference genome coordinates *	ORF/ protein	Protein function	Structural Correspondence † ‡	Structural proximity to known functional sites
18121	Orf1ab/ nsp14	ExoRNAse	**S28** in SARS-CoV-1 (PDB:5C8S)	Residue in the ExoN domain Proximal to the nsp10 interaction site., which cleaves terminal nucleotides during replication
20344	Orf1ab/ nsp15	EndoRNAse	**H243** in SARS-CoV-2 (PDB:6WLC); **H242** in SARS-CoV-1 (PDB:2H85)	Within the NendoU catalytic domain, which cleaves non-terminal uracil nucleotides during replication
21623	Spike (S1^A)	RBD	**R21** in SARS-CoV-2; **V25** in SARS-CoV-1; **K29** in OC43; **K28** in HKU1	Proximal to the S1^A domain involved in receptor recognition for the LinA viruses
21635			**P25** in SARS-CoV-2; **N29** in SARS-CoV-1; **P33** in OC43; **V32** in HKU1	Proximal to the S1^A domain involved in receptor recognition for the LinA viruses
23948	Spike (S2)	Viral fusion	**D796** in SARS-CoV-2; **Y778** in SARS-CoV-1; **N890** in OC43; D**878** in HKU1	Near the trimerization surface, which undergoes conformational rearrangements during viral fusion
24614			**I1018** in SARS-CoV-2; **I1000** in SARS-CoV-1; **V1112** in OC43; **I1099** in HKU1	In central helix domain, which undergoes conformational rearrangements during viral fusion
24620			**A1020** in SARS-CoV-2; **A1002** in SARS-CoV-1; **F1114** in OC43; **A1101** in HKU1	In central helix domain, which undergoes conformational rearrangements during viral fusion
24632			**L1024** in SARS-CoV-2; **L1006** in SARS-CoV-1; **Q1118** in OC43; **R1105** in HKU1	In central helix domain, which undergoes conformational rearrangements during viral fusion

* For SARS_CoV_2|Wuhan-Hu-1|MN908947 reference sequence.

† Structures of the relevant protein (domains) have not been solved for all four betacoronaviruses studied here.

‡ Available structures used in this study: PDB 2W2G, PDB 5C8S, PDB 6WLC, PDB 5I08, PDB 6OHW, PDB 6ACC and PDB 6VXX (see Methods section 5)

REFERENCES

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. Nat Med 26:450-452.

Avanzato VA, Oguntuyo KY, Escalera-Zamudio M, Gutierrez B, Golden M, Kosakovsky Pond SL, Pryce R, Walter TS, Seow J, Doores KJ, et al. 2019. A structural basis for antibody-mediated neutralization of Nipah virus reveals a site of vulnerability at the fusion glycoprotein apex. Proc Natl Acad Sci U S A 116:25057-25067.

Bakkers MJ, Lang Y, Feitsma LJ, Hulswit RJ, de Poot SA, van Vliet AL, Margine I, de Groot-Mijnes JD, van Kuppeveld FJ, Langereis MA, et al. 2017. Betacoronavirus Adaptation to Humans Involved Progressive Loss of Hemagglutinin-Esterase Lectin Activity. Cell Host Microbe 21:356-366.

Banerjee A, Doxey AC, Mossman K, Irving AT. 2021. Unraveling the Zoonotic Origin and Transmission of SARS-CoV-2. Trends Ecol Evol 36:180-184.

Benton DJ, Wrobel AG, Xu P, Roustan C, Martin SR, Rosenthal PB, Skehel JJ, Gamblin SJ. 2020. Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. Nature 588:327-330.

Boni MF, Gog JR, Andreasen V, Feldman MW. 2006. Epidemic dynamics and antigenic evolution in a single season of influenza A. Proc Biol Sci 273:1307-1316.

Boni MF, Lemey P, Jiang X, Lam TT, Perry BW, Castoe TA, Rambaut A, Robertson DL. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol 5:1408-1417.

Bosch BJ, van der Zee R, de Haan CA, Rottier PJ. 2003. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol 77:8801-8811.

Campbell KM, Steiner G, Wells DK, Ribas A, Kalbasi A. 2020. Prioritization of SARS-CoV-2 epitopes using a pan-HLA and global population inference approach. bioRxiv.

Cheng VC, Lau SK, Woo PC, Yuen KY. 2007. Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. Clin Microbiol Rev 20:660-694.

Chinese SMEC Consortium C-GU. 2021. COVID-19 Genomics UK Consortium. 2004. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science 303:1666-1669.

Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 176:104742.

Cui J, Li F, Shi ZL. 2019. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17:181-192.

Dejnirattisai, W., Zhou, D., Ginn, H. M., Duyvesteyn, H., Supasa, P., Case, J. B., Zhao, Y., Walter, T. S., Mentzer, A. J., Liu, C., Wang, B., Paesen, G. C., Slon-Campos, J., López-Camacho, C., Kafai, N. M., Bailey, A. L., Chen, R. E., Ying, B., Thompson, C., Bolton, J., … Screaton, G. R. 2021. The antigenic anatomy of SARS-CoV-2 receptor binding domain. Cell, 184(8), 2183–2200.e22. https://doi.org/10.1016/j.cell.2021.02.032

Delport W, Scheffler K, Seoighe C. 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLoS Pathog 4:e1000242.

Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041.

Dolan PT, Whitfield ZJ, Andino R. 2018. Mapping the Evolutionary Potential of RNA Viruses. Cell Host Microbe 23:435-446.

Egloff MP, Malet H, Putics A, Heinonen M, Dutartre H, Frangeul A, Gruez A, Campanacci V, Cambillau C, Ziebuhr J, et al. 2006. Structural and functional basis for ADP-ribose and poly(ADP-ribose) binding by viral macro domains. J Virol 80:8493-8502.

Ellegren H. 2008. Comparative genomics and the study of evolution by natural selection. Mol Ecol 17:4586-4596.

Escalera-Zamudio M, Golden M, Gutierrez B, Theze J, Keown JR, Carrique L, Bowden TA, Pybus OG. 2020. Parallel evolution in the emergence of highly pathogenic avian influenza A viruses. Nat Commun 11:5511.

Evolutionary annotation of global SARS-CoV-2/COVID-19 genomes enabled by data from GSAID [Internet]. 2021. Available from: https://observablehq.com/@spond/sars_cov_2_sites

Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DDS, Mishra S, Crispim MAE, Sales FCS, Hawryluk I, McCrone JT, Hulswit RJG, Franco LAM, Ramundo MS, de Jesus JG, Andrade PS, Coletti TM, Ferreira GM, Silva CAM, Manuli ER, Pereira RHM, Peixoto PS, Kraemer MUG, Gaburo N Jr, Camilo CDC, Hoeltgebaum H, Souza WM, Rocha EC, de Souza LM, de Pinho MC, Araujo LJT, Malta FSV, de Lima AB, Silva JDP, Zauli DAG, Ferreira ACS, Schnekenberg RP, Laydon DJ, Walker PGT, Schlüter HM, Dos Santos ALP, Vidal MS, Del Caro VS, Filho RMF, Dos Santos HM, Aguiar RS, Proença-Modena JL, Nelson B, Hay JA, Monod M, Miscouridou X, Coupland H, Sonabend R, Vollmer M, Gandy A, Prete CA Jr, Nascimento VH, Suchard MA, Bowden TA, Pond SLK, Wu CH, Ratmann O, Ferguson NM, Dye C, Loman NJ, Lemey P, Rambaut A, Fraiji NA, Carvalho MDPSS, Pybus OG, Flaxman S, Bhatt S, Sabino EC. 2021. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. Apr 14:eabh2644. doi: 10.1126/science.abh2644. Epub ahead of print. PMID: 33853970.

Farris SJ. 1977. Phylogenetic Analysis Under Dollo's Law. Systematic Biology 26:77–88.

Fehr AR, Athmer J, Channappanavar R, Phillips JM, Meyerholz DK, Perlman S. 2015. The nsp3 macrodomain promotes virulence in mice with coronavirus-induced encephalitis. J Virol 89:1523-1536.

Fehr AR, Channappanavar R, Jankevicius G, Fett C, Zhao J, Athmer J, Meyerholz DK, Ahel I, Perlman S. 2016. The Conserved Coronavirus Macrodomain Promotes Virulence and Suppresses the Innate Immune Response during Severe Acute Respiratory Syndrome Coronavirus Infection. mBio 7.

Fehr AR, Channappanavar R, Perlman S. 2017. Middle East Respiratory Syndrome: Emergence of a Pathogenic Human Coronavirus. Annu Rev Med 68:387-399.

Fehr AR, Jankevicius G, Ahel I, Perlman S. 2018. Viral Macrodomains: Unique Mediators of Viral Replication and Pathogenesis. Trends Microbiol 26:598-610.

Genetic diversity of betacoronaviruses including novel coronavirus (nCoV) [Internet]. Available from: https://nextstrain.org/groups/blab/beta-cov

Global Initiative on Sharing Avian Influenza Data [Internet]. 2021. Available from: https://www.gisaid.org/

Gutierrez B, Escalera-Zamudio M, Pybus OG. 2019. Parallel molecular evolution and adaptation in viruses. Curr Opin Virol 34:90-96.

Hackbart M, Deng X, Baker SC. 2020. Coronavirus endoribonuclease targets viral polyuridine sequences to evade activating host sensors. Proc Natl Acad Sci U S A 117:8094-8103.

Han W, Li X, Fu X. 2011. The macro domain protein family: structure, functions, and their potential therapeutic implications. Mutat Res 727:86-103.

Hulswit RJG, Lang Y, Bakkers MJG, Li W, Li Z, Schouten A, Ophorst B, van Kuppeveld FJM, Boons GJ, Bosch BJ, et al. 2019. Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. Proc Natl Acad Sci U S A 116:2681-2690.

Hulswit, R. J., de Haan, C. A., & Bosch, B. J. 2016. Coronavirus Spike Protein and Tropism Changes. Advances in virus research, 96, 29–57. https://doi.org/10.1016/bs.aivir.2016.08.004

Issues with SARS-CoV-2 sequencing data [Internet]. 2020. Available from: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649.

Kim Y, Cheon S, Min CK, Sohn KM, Kang YJ, Cha YJ, Kang JI, Han SK, Ha NY, Kim G, et al. 2016. Spread of Mutant Middle East Respiratory Syndrome Coronavirus with Reduced Affinity to Human CD26 during the South Korean Outbreak. mBio 7:e00019.

Kim Y, Wower J, Maltseva N, Chang C, Jedrzejczak R, Wilamowski M, Kang S, Nicolaescu V, Randall G, Michalska K, et al. 2021. Tipiracil binds to uridine site and inhibits Nsp15 endoribonuclease NendoU from SARS-CoV-2. Commun Biol 4:193.

Kirchdoerfer RN, Cottrell CA, Wang N, Pallesen J, Yassine HM, Turner HL, Corbett KS, Graham BS, McLellan JS, Ward AB. 2016. Pre-fusion structure of a human coronavirus spike protein. Nature 531:118-121.

Kirchdoerfer RN, Wang N, Pallesen J, Wrapp D, Turner HL, Cottrell CA, Corbett KS, Graham BS, McLellan JS, Ward AB. 2018. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. Sci Rep 8:15701.

Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science 368:860-868.

Kistler, K. E., & Bedford, T. 2021. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. eLife, 10, e64509. https://doi.org/10.7554/eLife.64509

Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208-1222.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096-3098.

Kosakovsky Pond SL, Wisotsky SR, Escalante A, Magalis BR, Weaver S. 2020. Contrast-FEL - a test for differences in selective pressures at individual sites among clades and sets of branches. Mol Biol Evol.

Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, et al. 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581:215-220.

Lau SK, Lee P, Tsang AK, Yip CC, Tse H, Lee RA, So LY, Lau YL, Chan KH, Woo PC, et al. 2011. Molecular epidemiology of human coronavirus OC43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination. J Virol 85:11325-11337.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol 5:e1000520.

Li F, Li W, Farzan M, Harrison SC. 2005. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 309:1864-1868.

Li F. 2012. Evidence for a common evolutionary origin of coronavirus spike protein receptor-binding subunits. Journal of virology, 86(5), 2856–2858. https://doi.org/10.1128/JVI.06882-11

Li F. 2016. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annual review of virology, 3(1), 237–261. https://doi.org/10.1146/annurev-virology-110615-042301

Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, et al. 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science 310:676-679.

Li, Z., Tomlinson, A. C., Wong, A. H., Zhou, D., Desforges, M., Talbot, P. J., Benlekbir, S., Rubinstein, J. L., & Rini, J. M. 2019. The human coronavirus HCoV-229E S-protein structure and receptor binding. eLife, 8, e51230. https://doi.org/10.7554/eLife.51230

Loewe L, Hill WG. 2010. The population genetics of mutations: good, bad and indifferent. Philos Trans R Soc Lond B Biol Sci 365:1153-1167.

Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, Lou Z, Yan L, Zhang R, Rao Z. 2015. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. Proc Natl Acad Sci U S A 112:9436-9441.

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, Kosakovsky Pond SL, Robertson DL. 2021. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19:e3001115.

McIntosh K, Becker WB, Chanock RM. 1967. Growth in suckling-mouse brain of "IBV-like" viruses from patients with upper respiratory tract disease. Proc Natl Acad Sci U S A 58:2268-2273.

Menachery VD, Graham RL, Baric RS. Jumping species-a mechanism for coronavirus persistence and survival. Curr Opin Virol. 2017 Apr;23:1-7. doi: 10.1016/j.coviro.2017.01.002. Epub 2017 Mar 31. PMID: 28214731; PMCID: PMC5474123.

Middle East respiratory syndrome [Internet]. 2021. Available from: http://www.emro.who.int/health-topics/mers-cov/mers-outbreaks.html

Millet JK, Whittaker GR. 2015. Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. Virus Res. 2015;202:120-134.

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. Mol Biol Evol 32:1365-1371.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:e1002764.

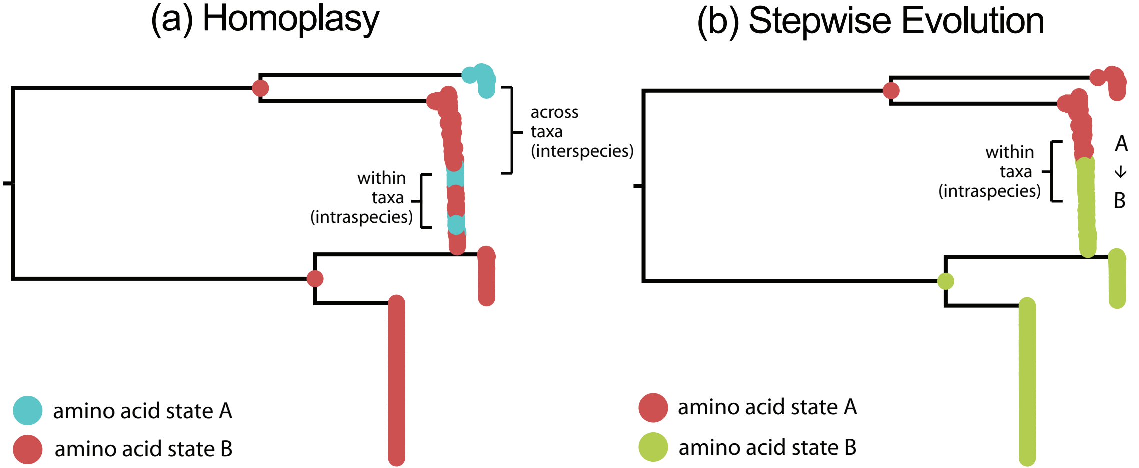
Nelde A, Bilich T, Heitmann JS, Maringer Y, Salih HR, Roerden M, Lubke M, Bauer J, Rieth J, Wacker M, et al. 2021. SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. Nat Immunol 22:74-85.

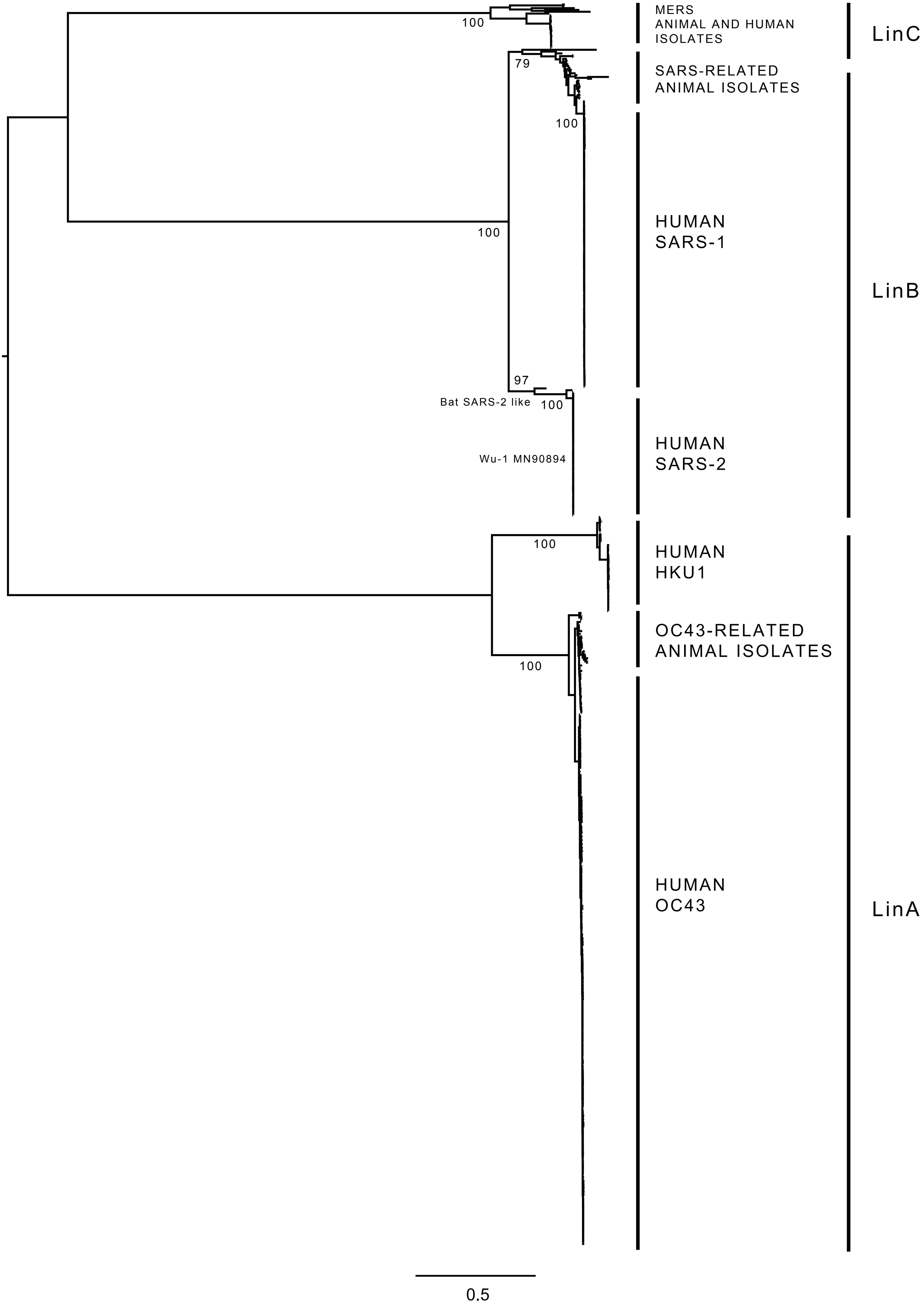
Okba NMA, Muller MA, Li W, Wang C, GeurtsvanKessel CH, Corman VM, Lamers MM, Sikkema RS, de Bruin E, Chandler FD, et al. 2020. Severe Acute Respiratory Syndrome Coronavirus 2-Specific Antibody Responses in Coronavirus Disease Patients. Emerg Infect Dis 26:1478-1488.

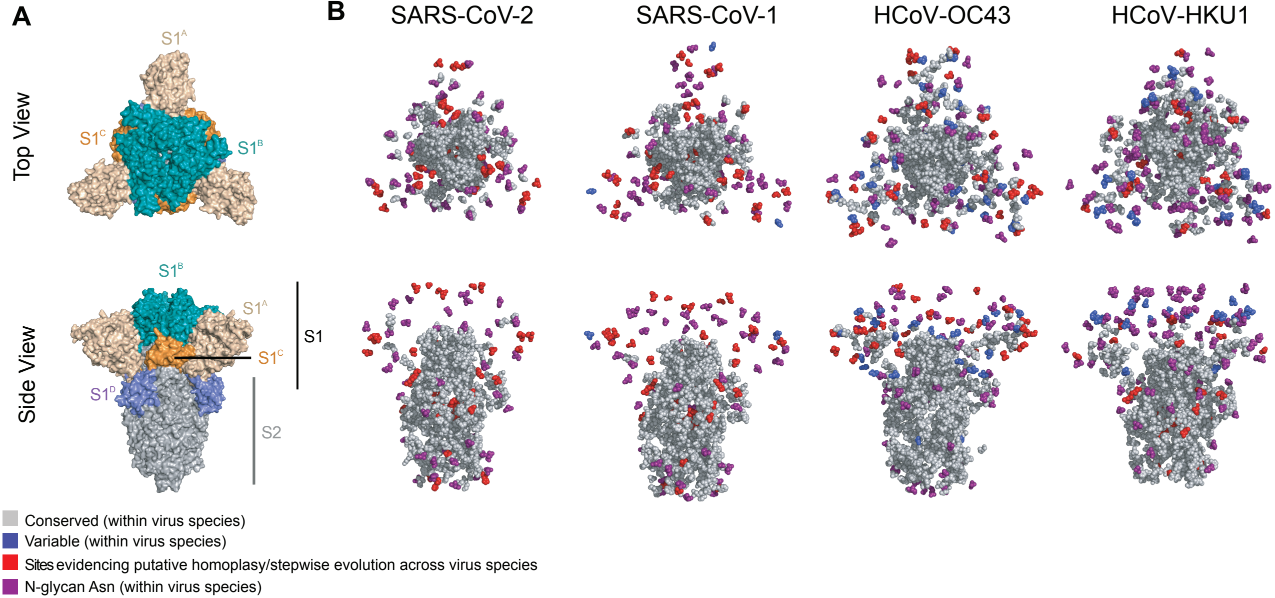
Oong XY, Ng KT, Takebe Y, Ng LJ, Chan KG, Chook JB, Kamarulzaman A, Tee KK. 2017. Identification and evolutionary dynamics of two novel human coronavirus OC43 genotypes

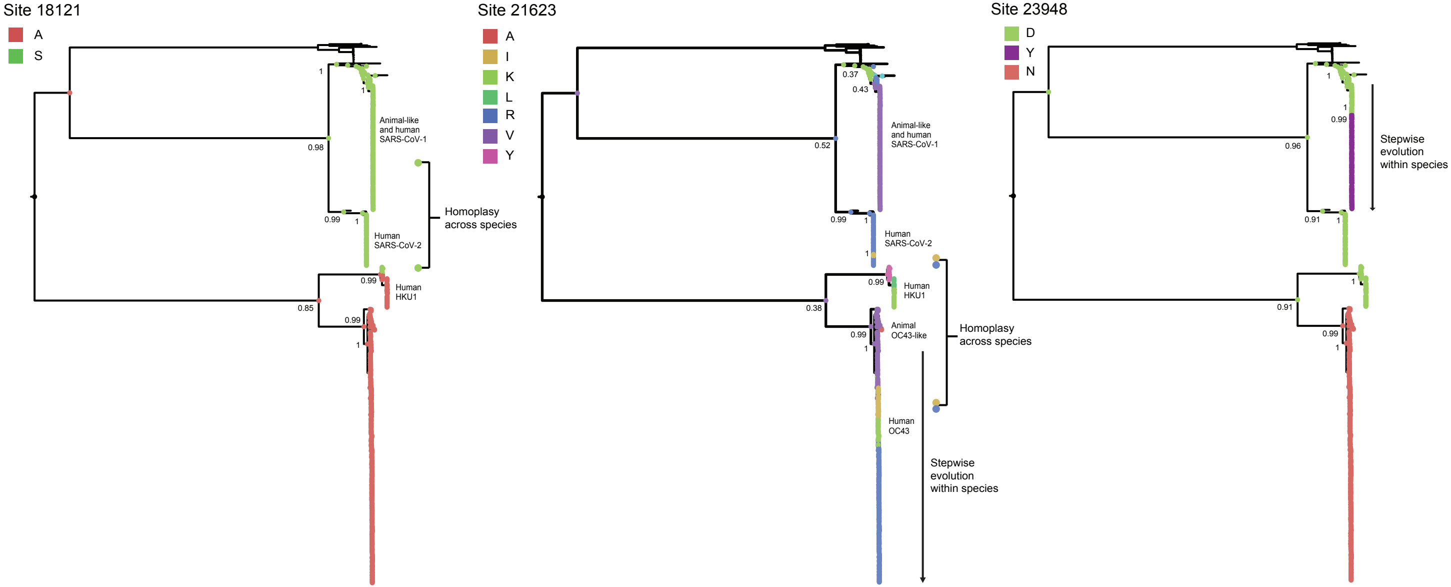
associated with acute respiratory infections: phylogenetic, spatiotemporal and transmission network analyses. Emerg Microbes Infect 6:e3.
Pallesen J, Wang N, Corbett KS, Wrapp D, Kirchdoerfer RN, Turner HL, Cottrell CA, Becker MM, Wang L, Shi W, et al. 2017. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proc Natl Acad Sci U S A 114:E7348-E7357.
Pangolin COVID-19 Lineage Assigner [Internet]. Available from: https://pangolin.cog-uk.io/
Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P. 2008. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev 72:457-470.
Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, Lim W, Nicholls J, Yee WK, Yan WW, Cheung MT, et al. 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet 361:1319-1325.
Polster, R., Petropoulos, C. J., Bonhoeffer, S., & Guillaume, F. 2016. Epistasis and Pleiotropy Affect the Modularity of the Genotype-Phenotype Map of Cross-Resistance in HIV-1. Molecular biology and evolution, 33(12), 3213–3225. https://doi.org/10.1093/molbev/msw206
Pond SL, Frost SD, Grossman Z, Gravenor MB, Richman DD, Brown AJ. 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput Biol 2:e62.
Pond SL, Murrell B, Poon AF. 2012. Evolution of viral genomes: interplay between selection, recombination, and other forces. Methods Mol Biol 856:239-272.
PRIME [Internet]. 2013. Available from: http://hyphy.org/w/index.php/PRIME
Rausch JW, Capoferri AA, Katusiime MG, Patro SC, Kearney MF. 2020. Low genetic diversity may be an Achilles heel of SARS-CoV-2. Proc Natl Acad Sci U S A 117:24614-24616.
Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol 4:vex042.
Shaman J, Galanti M. 2020. Will SARS-CoV-2 become endemic? Science 370:527-529.
Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, Geng Q, Auerbach A, Li F. 2020. Structural basis of receptor recognition by SARS-CoV-2. Nature 581:221-224.
Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23:7-9.
Simmonds P. 2020. Rampant C-->U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere 5.
Song W, Gui M, Wang X, Xiang Y. 2018. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS Pathog 14:e1007236.
Stamatakis A. 2015. Using RAxML to Infer Phylogenies. Curr Protoc Bioinformatics 51:6 14 11-16 14 14.
Starr TN, Greaney AJ, Addetia A, Hannon WW, Choudhary MC, Dingens AS, Li JZ, Bloom JD. 2021. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. Science 371:850-854.
Stern A, Yeh MT, Zinger T, Smith M, Wright C, Ling G, Nielsen R, Macadam A, Andino R. 2017. The Evolutionary Pathway to Virulence of an RNA Virus. Cell 169:35-46 e19.
Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF. 2016. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol 24:490-502.
Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4:vey016.
Taefehshokr N, Taefehshokr S, Hemmat N, Heit B. 2020. Covid-19: Perspectives on Innate Immune Evasion. Front Immunol 11:580641.
Tan J, Vonrhein C, Smart OS, Bricogne G, Bollati M, Kusov Y, Hansen G, Mesters JR, Schmidt CL, Hilgenfeld R. 2009. The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. PLoS Pathog 5:e1000428.
Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, Mlisana K, Bhiman JN, von Gottberg A, Walaza S, Fonseca V, Allam M, Ismail A, Glass AJ, Engelbrecht S, Van Zyl G, Preiser W, Williamson C, Petruccione F, Sigal A, Gazy I, Hardie D, Hsiao NY, Martin D, York D, Goedhals D, San EJ, Giovanetti M, Lourenço J, Alcantara LCJ, de Oliveira T. Sixteen novel lineages of SARS-CoV-2 in South Africa. Nat Med. 2021 Mar;27(3):440-446. doi: 10.1038/s41591-021-01255-3. Epub 2021 Feb 2. PMID: 33531709.
Tortorici MA, Veesler D. 2019. Structural insights into coronavirus entry. Adv Virus Res 105:93-116.

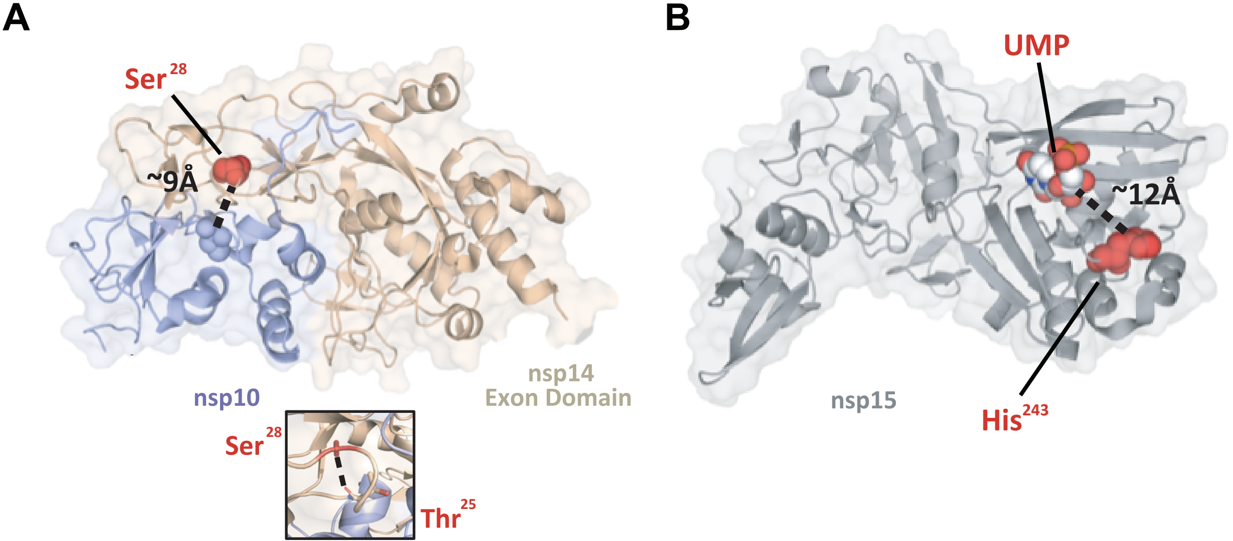
Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 [Internet]. 2021. Available from: https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592
van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, et al. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 83:104351.
van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F. 2020. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat Commun 11:5986.
Vijaykrishna D, Smith GJ, Zhang JX, Peiris JS, Chen H, Guan Y. 2007. Evolutionary insights into the ecology of coronaviruses. J Virol 81:4012-4020.
Vijgen L, Keyaerts E, Moes E, Thoelen I, Wollants E, Lemey P, Vandamme AM, Van Ranst M. 2005. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. J Virol 79:1595-1604.
Virus Pathogen Resource [Internet]. 2021. Available from: https://www.viprbrc.org/brc/home.spg?decorator=vipr
Viruses NRotICoTo. 2012. Family - Coronaviridae. In: Andrew MQ, King MJ, editors. Virus Taxonomy p. 806-828.
Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 183:1735.
Wang G, Dunbrack RL, Jr. 2004. Scoring profile-to-profile sequence alignments. Protein Sci 13:1612-1626.
Wang H, Pipes L, Nielsen R. 2021. Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. Virus Evol 7:veaa098.
Wang Y, Sun Y, Wu A, Xu S, Pan R, Zeng C, Jin X, Ge X, Shi Z, Ahola T, et al. 2015. Coronavirus nsp10/nsp16 Methyltransferase Can Be Targeted by nsp10-Derived Peptide In Vitro and In Vivo To Reduce Replication and Pathogenesis. J Virol 89:8416-8427.
Watanabe Y, Bowden TA, Wilson IA, Crispin M. 2019. Exploitation of glycosylation in enveloped virus pathobiology. Biochim Biophys Acta Gen Subj 1863:1480-1497.
Woo PC, Huang Y, Lau SK, Yuen KY. 2010. Coronavirus genomics and bioinformatics analysis. Viruses 2:1804-1820.
Woo PC, Lau SK, Chu CM, Chan KH, Tsoi HW, Huang Y, Wong BH, Poon RW, Cai JJ, Luk WK, et al. 2005. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J Virol 79:884-895.
Woo PC, Lau SK, Yip CC, Huang Y, Tsoi HW, Chan KH, Yuen KY. 2006. Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1. J Virol 80:7136-7145.
Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, Rambaut A, Suchard MA, Wertheim JO, Lemey P. 2020. The emergence of SARS-CoV-2 in Europe and North America. Science 370:564-570.
Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, et al. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265-269.
Xia S, Lan Q, Su S, Wang X, Xu W, Liu Z, Zhu Y, Wang Q, Lu L, Jiang S. 2020. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. Signal Transduct Target Ther 5:92.
Yoshimoto FK. 2020. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. Protein J 39:198-216.
Yuen CK, Lam JY, Wong WM, Mak LF, Wang X, Chu H, Cai JP, Jin DY, To KK, Chan JF, et al. 2020. SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists. Emerg Microbes Infect 9:1418-1428.
Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270-273.
Zhu Y, Li C, Chen L, Xu B, Zhou Y, Cao L, Shang Y, Fu Z, Chen A, Deng L, et al. 2018. A novel human coronavirus OC43 genotype detected in mainland China. Emerg Microbes Infect 7:173.

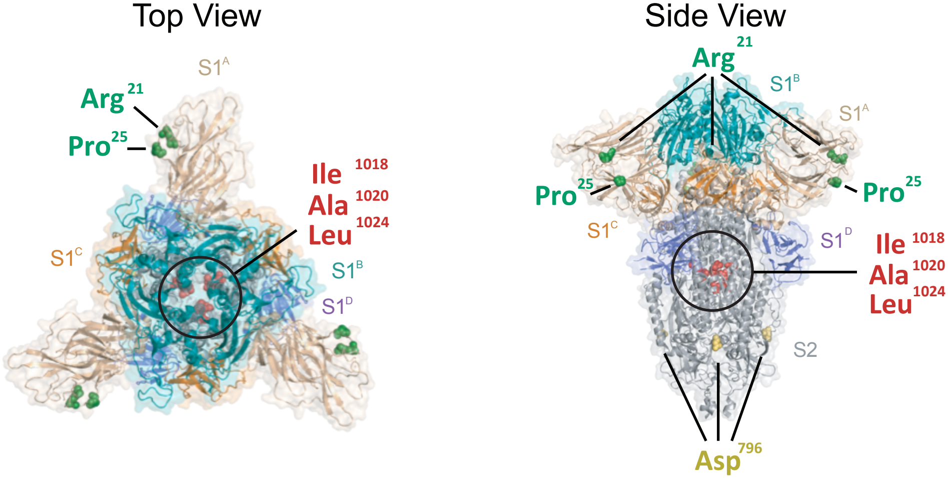

(a) Homoplasy
across
taxa
(interspecies)
within
taxa
(intraspecies)
amino acid state A
amino acid state B
(b) Stepwise Evolution
within
taxa
(intraspecies)
A
↓
B
amino acid state A
amino acid state B


MERS
ANIMAL AND HUMAN
ISOLATES
LinC
100
79
SARS-RELATED
ANIMAL ISOLATES
100
100
HUMAN
SARS-1
LinB
97
Bat SARS-2 like
100
HUMAN
SARS-2
Wu-1 MN90894
100
HUMAN
HKU1
OC43-RELATED
ANIMAL ISOLATES
100
HUMAN
OC43
LinA
0.5


A
Top View
S1^A
S1^C
S1^B
Side View
S1^B
S1^A
S1^C
S1^D
S1
S2
B
SARS-CoV-2
SARS-CoV-1
HCoV-OC43
HCoV-HKU1
Conserved (within virus species)
Variable (within virus species)
Sites evidencing putative homoplasy/stepwise evolution across virus species
N-glycan Asn (within virus species)


Site 18121
🟥 A
🟩 S
1
1
0.98
Animal-like
and human
SARS-CoV-1
0.99
1
Human
SARS-CoV-2
Homoplasy
across species
0.99
Human
HKU1
0.85
0.99
1
Site 21623
🟥 A
🟨 I
🟩 K
🟩 L
🟦 R
🟪 V
🟪 Y
0.37
0.43
Animal-like
and human
SARS-CoV-1
0.52
0.99
1
Human
SARS-CoV-2
1
0.99
Human
HKU1
0.38
Animal
OC43-like
0.99
1
Homoplasy
across species
Human
OC43
Stepwise
evolution
within species
Site 23948
🟩 D
🟪 Y
🟥 N
1
1
0.99
Stepwise
evolution
within species
0.96
0.91
1
1
0.91
0.99
1


A
Ser 28
~9Å
nsp10
nsp14
Exon Domain
Ser 28
Thr 25
B
UMP
~12Å
nsp15
His 243


Top View
S1^{A}
Arg^{21}
Pro^{25}
Ile^{1018}
Ala^{1020}
Leu^{1024}
S1^{C}
S1^{B}
S1^{D}
Side View
Arg^{21}
S1^{B}
S1^{A}
Pro^{25}
Pro^{25}
S1^{C}
S1^{D}
Ile^{1018}
Ala^{1020}
Leu^{1024}
S2
Asp^{796}