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

Proposed title:
Identification of evolutionary trajectories shared across human betacoronaviruses

Authors listed according to affiliation:
Marina Escalera-Zamudio [1,2*]
Sergei L. Kosakovsky Pond [3]
Natalia Martínez de la Viña [1]
Bernardo Gutiérrez [1,2]
Rhys P. D. Inward [1]
Julien Thézé [4]
Lucy van Dorp [5]
Hugo G. Castelán-Sánchez [2,6]
Thomas A. Bowden [7]
Oliver G. Pybus [1,8]
Ruben J.G. Hulswit [7*]

Affiliations:

1. Department of Biology, University of Oxford, Oxford, OX1 3PS, UK
2. Consorcio Mexicano de Vigilancia Genómica (CoViGen-Mex)
3. Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, USA
4. Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, 63122, Saint-Genès-Champanelle, France
5. UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, Gower Street, WC1E 6BT, London
6. Programa de Investigadoras e Investigadores por México, Consejo Nacional de Ciencia y Tecnología, CP 03940, CDMX, México
7. Division of Structural Biology, Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
8. Department of Pathobiology, Royal Veterinary College, NW1 0TU, London, United Kingdom

Email addresses (*corresponding authors):
marina.escalerazamudio@biology.ox.ac.uk *
spond@temple.edu
nataliamv@gmail.com
bernardo.gutierrez@biology.ox.ac.uk
rhys.inward@biology.ox.ac.uk
julien.theze@inrae.fr
lucy.dorp.12@ucl.ac.uk
hugo.castelan@conacyt.mx
thomas.bowden@strubi.ox.ac.uk
oliver.pybus@biology.ox.ac.uk
ruben.hulswit@strubi.ox.ac.uk *

Preprint available at: bioRxiv. 2021 May 25:2021.05.24.445313. doi: 10.1101/2021.05.24.445313

ABSTRACT (238)

Comparing the evolution of distantly related viruses can provide insights into common adaptive processes related to shared ecological niches. Phylogenetic approaches, coupled with other molecular evolution tools, can help identify mutations informative on adaptation, whilst the structural contextualization of these to functional sites of proteins may help gain insight into their biological properties. Two zoonotic betacoronaviruses capable of sustained human-to-human transmission have caused pandemics in recent times (SARS-CoV-1 and SARS-CoV-2), whilst a third virus (MERS-CoV) is responsible for sporadic outbreaks linked to animal infections. Moreover, two other betacoronaviruses have circulated endemically in humans for decades (HKU1 and OC43). To search for evidence of adaptive convergence between established and emerging betacoronaviruses capable of sustained human-to-human transmission (HKU1, OC43, SARS-CoV-1 and SARS-CoV-2), we developed a methodological pipeline to classify shared non-synonymous mutations as putatively denoting homoplasy (repeated mutations that do not share direct common ancestry) or stepwise evolution (sequential mutations leading towards a novel genotype). In parallel, we look for evidence of positive selection, and draw upon protein structure data to identify potential biological implications. We find 30 mutations, with four of these [codon sites 18121 (nsp14/residue 28), 21623 (spike/21), 21635 (spike/25) and 23948 (spike/796); SARS-CoV-2 genome numbering] displaying evolution under positive selection and proximity to functional protein regions. Our findings shed light on potential mechanisms underlying betacoronavirus adaptation to the human host and pinpoint common mutational pathways that may occur during establishment of human endemicity.

KEYWORDS

Molecular Evolution, Phylogenomics, Convergence, Adaptation, Betacoronaviruses

/MAIN TEXT (6589)

INTRODUCTION (882)

Mutation is a fundamental process of evolution, and a main source of variability enabling genetic diversity (Loewe and Hill 2010). However, most mutations are expected to be either eliminated through purifying selection, or to be selectively neutral and thus fixed randomly. Although RNA viruses change rapidly due to their relatively small genomes and high mutation rates, the mutational pathways leading to adaptation are often limited by functional and evolutionary constraints (such as epistasis, referring to the adaptive effect of a mutation being dependent on the genetic background in which it appears) (Dolan et al. 2018). Following a nearly-neutral evolutionary theory, only the smallest proportion of mutations are expected to be adaptive and subsequently fixed through positive selection (Ohta 1973, Pond et al. 2012). Thus, given a reduced admissible genetic variability in viruses, evolutionary trajectories are limited, and may sometimes display recurrent mutational patterns denoting adaptive convergence, at least for viruses sharing ecological niches (Gutierrez, Escalera-Zamudio, and Oliver G Pybus 2019).

Coronaviruses are well known for their propensity to switch host species, evidenced by multiple zoonotic introductions into the human population. There are four betacoronavirus species capable of sustained human transmission: OC43, HKU1, SARS-CoV-1, and SARS-CoV-2. These viruses are further classified into two subgenera: *Embecovirus* and *Sarbecovirus*. The *Embecovirus* subgenus includes OC43 first described in 1967 (McIntosh et al. 1967), and HKU1 identified in 2005 (Woo et al. 2005). Both were introduced into the human population through independent zoonotic events estimated to have occurred at least 50 years ago (Vijgen et al. 2005; Lau et al. 2015; Kistler and Bedford 2021; Pollett et al. 2021), and are associated with mild respiratory disease (Su et al. 2016). On the other hand, the two viruses belonging to the *Sarbecovirus* subgenus (SARS-CoV-1 and SARS-CoV-2) were introduced independently but much more recently, and are associated with contemporary pandemic outbreaks (W. Li et al. 2005; Vijaykrishna et al. 2007; Andersen et al. 2020; Boni et al. 2020; Banerjee et al. 2021).

Following its first detection in 2002 (Peiris et al. 2003), SARS-CoV-1 spread to more than 20 countries in six months, resulting in a short-lived but severe outbreak characterized by sustained human-to-human transmission (Cheng et al. 2007). Even though SARS-CoV-1 circulation was successfully halted, the virus underwent adaption to the human population, resulting in enhanced transmissibility (He et al. 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, despite an apparently low rate of adaptive change during the early stages of the pandemic (van Dorp et al. 2020; MacLean et al. 2021). The constant circulation of OC43 and HKU1 in humans has been accompanied by continuous host-specific adaptation, a

process now evident for SARS-CoV-2, represented by the ongoing emergence of virus lineages across time and space (with sub-lineages now reflecting regional 'endemic' patterns) (O'Toole et al. 2021). Thus, as SARS-CoV-2 becomes established in humans (Shaman and Galanti 2020), it will continue to adapt to overcome the selective pressures exerted by the collective immune response of the human host population (Kissler et al. 2020).

We hypothesize that adaptive convergence may occur across distantly related betacoronaviruses sharing the same ecological niche (in this case represented by the human host environment) (Gutierrez, Escalera-Zamudio, and Oliver G Pybus 2019). To test this hypothesis, we undertook a comparative analysis to search for evidence of shared mutational pathways between emerging sarbecoviruses that display sustained transmission in humans (particularly, SARS-CoV-2) and established human-endemic embecoviruses. Although MERS-CoV (*Merbecovirus* subgenus) can also infect humans, outbreaks have been the result of recurrent zoonotic events characterized by limited transmission chains, with no further evidence of adaptation (Fehr et al. 2017; WHO 2021). Thus, MERS-CoV and related viruses were excluded from our analysis.

Our approach involved the development of a methodological pipeline that allows the identification of amino acid changes putatively associated with adaptive convergence across multiple virus species. This is done by detecting non-synonymous mutations shared across virus taxa that display putative evidence of homoplasy (*i.e.,* repeated mutations that do not share direct common ancestry) and/or stepwise evolution (*i.e.,* sequential mutations leading towards a novel genotype) (for details see Supplementary Text 1 and Supplementary Figure 1). In parallel, we assess the impact of positive selection acting upon specific sites and/or branches (using dN/dS estimating methods), and determine the proximity of the mutations identified to viral protein surfaces with established function. Under these circumstances, structural contextualization of mutations may provide complementary information (to that derived from phylogenetic approaches), helping reveal the potential functional relevance for sites inferred as associated with adaptation across distinct virus species (Ellegren 2008; Hulswit et al. 2016; Avanzato et al. 2019; Escalera-Zamudio et al. 2020).

We identify 30 non-synonymous mutations shared across betacoronaviruses capable of sustained human-to-human transmission that display evolutionary patterns denoting putative homoplasy and/or stepwise evolution. The mutations identified were further ranked according to their selective relevance, and to their proximity to protein regions of known function. Under this approach, we subsequently identify four mutations (sites 18121 [nsp14/27], 21623 [spike/21], 21635 [spike/25] and 23948 [spike/796], in SARS-CoV-2 genome coordinates) near the functional surfaces of nsp14 (Ma et al. 2015), the spike protein S1 receptor-interacting subunit, and the S2 fusion machinery subunit. Our results provide a

molecular-level context for common evolutionary trajectories that betacoronaviruses may undergo during their adaptation to the human host.

RESULTS (2552)

Patterns of genetic variability observed in human-infecting betacoronaviruses

Based on an alignment of the Orf1ab and S viral genes comprising approximately 1400 sequences (Methods section 1), we reconstructed the genomic diversity of the four distantly related betacoronavirus species studied here (Methods section 2). The patterns we recover following phylogenetic inference (represented by the expanded tree shown in Figure 1) are consistent with previously published phylogenies of the genus (Woo et al. 2006; Woo et al. 2010; Oong et al. 2017; Zhu et al. 2018; Bedford 2021). In this context, virus sequences belonging to the *Embecovirus* (HKU1, OC43, and related viruses) and the *Sarbecovirus* (SARS-CoV-1, SARS-CoV-2, and related viruses) (ICTV et al. 2017) subgenera form four well-supported clades. We were further able to identify all previously described genotypes for HKU1 (A-C) and OC43 (A-H) (Woo et al. 2006; Oong et al. 2017) (Supplementary Data 1). In summary, derived from genomic diversity reconstructions, our observations validate the evolutionary relationship known for these four distantly related betacoronaviruses.

Approximately 2% of the sites within the Orf1ab+S alignment (205/8962 codons) correspond to non-synonymous mutations shared between embeco- and sarbecovirus species (*i.e.,* those present in any of the sarbecovirus clades, and also in HKU1 and/or OC43). For Orf1a, 2.7% of all sites (129/4774 codons) were identified as shared, whilst for Orf1b, only 0.9% of all sites (25/2623 codons) were shared. The highest proportion of shared mutations (3.2% of all sites, 48/1457 codons) was identified within Orf S. When analysing genetic variation within single virus species, a high degree of conservation was observed across Orf S (with >91% identity observed for all species). In general, most conserved sites we detect mapped to the membrane proximal S2 domain, whilst most variable sites mapped to the membrane distal S1 subunit (Figure 2). The predominance of variable sites within S1 compared to S2 was most evident for embecoviruses, and less so for sarbecoviruses; thus, we speculate that this may reflect a differential adaptation stage for the sarbecoviruses to the human host environment, evidenced by a lower degree of genetic divergence observed within Orf S.

When analysing genetic variation across virus species, the Orf S genomic region displayed a greater proportion of variable sites relative to the number conserved sites (for the definitions used for ‘conserved’ and ‘variable’ sites, see Methods section 3), as has been before noted for other coronaviruses (Hulswit et al. 2016). In this context, only 16% of the homologous sites within the Orf S alignment were found to be conserved, whilst the remaining 84% were variable (Supplementary Data 2). The highest proportion of conserved sites locate

111 to the S2 subunit, presumably reflecting functional constraints of the viral membrane fusion
112 machinery shared across coronavirus species (Li 2016). Within the S1 subunit, a greater
113 number of variable sites were observed, particularly within the $S1^A$ domain (also known as the
114 N-terminal domain, or NTD). In contrast, the $S1^B$ domain did not display any conserved sites
115 across virus species, likely associated with differences in receptor engagement process
116 between embeco- and sarbecoviruses (mediated by the S1 subunit), as embecoviruses use
117 the $S1^A$ domain to interact with sialoglycan-based receptors, whilst sarbecoviruses use their
118 $S1^B$ domain to bind to angiotensin-converting enzyme 2 (ACE2) (Hulswit et al. 2019; Lan et
119 al. 2020). Additionally, a conserved R residue at site 685 within the S1/S2 cleavage site
120 (numbering according to the SARS-CoV-2 protein, codon sites 23615-23617) was found to be
121 shared within and across all virus species (Supplementary Data 2), reflecting the conserved
122 proteolytic maturation mechanism of the spike protein across virus taxa belonging to the
123 *Orthocoronavirinae* family (Millet and Whittaker 2015).
124
125 Sites displaying evidence of homoplasy and/or stepwise evolution
126 Although not all non-synonymous mutations putatively displaying homoplasy and/or stepwise
127 evolution may derive from positive section, such mutational patterns are most likely to result
128 from adaption (Stern et al. 2017; Gutierrez, Escalera-Zamudio, and Oliver G. Pybus 2019;
129 Escalera-Zamudio et al. 2020). Thus, amongst the non-synonymous mutations identified as
130 shared across virus species, we further searched for those displaying putative evidence for
131 homoplasy and/or stepwise evolution (Supplementary Text 1) under our pipeline (Methods
132 section 3). After visual validation, we confirmed that 30 sites (representing 0.3% within the
133 Orf1ab+S alignment) display evolutionary patterns indicative of homoplasy and/or stepwise
134 evolution (see Supplementary Text 2, Supplementary Figure 2 and 3). Two of these were
135 found within Orf1a, nine within Orf1b, and 19 within Orf S (Table 1). The evolutionary
136 trajectories of different amino acid states observed for three illustrative sites (18121, 21623
137 and 23948) are highlighted below and in Figure 3.
138 Derived from the global, expanded and the re-sampled SARS-CoV-2 trees (Methods
139 section 1, 3 and 6, Supplementary Text 3), our results show that site 18121 (codon 18121-
140 18123 in Orf1b/nsp14, corresponding to amino acid state 'S') is homoplasic between HKU1
141 genotype B and the sarbecoviruses (Table 1, Figure 3, Supplementary Data 3). Comparably,
142 site 21623 (codon 21623-21626 in Orf S, corresponding to amino acid state 'R') was identified
143 as homoplasic between SARS-CoV-2 and OC43 genotypes D, F, G and H. This site also
144 displayed evidence for stepwise evolution within a single virus clade (OC43), exemplified by
145 the sequential amino acid replacement pattern of V→ I→ K→ R (Figure 3).
146 For site 23948 (codon 23948-23950 in Orf S), initial observations based on the global
147 tree revealed that amino acid state 'D' was present in all virus species, except for OC43

148 (displaying amino acid state 'N'). However, when replicating our analyses (expanded tree),
149 the distribution of amino acid state 'D' was now present in some embecoviruses (including
150 OC43 but excluding HKU1) and most sarbecoviruses. These discrepancies are likely due to
151 alignment uncertainty across genome regions of highly divergent virus taxa. Nonetheless,
152 based on consensus protein sequences and structural comparison, the structural
153 contextualization of 23948 and adjacent sites confirmed the presence of 'D' in SARS-CoV-1,
154 SARS-CoV-2 and HKU1, and 'N' in OC43, (Supplementary Figure 6). Thus, amino acid state
155 'D' at site 23948 shows evidence of homoplasy between the SARS-CoV-1, SARS-CoV-2 and
156 HKU1.
157 For this same site (23948), an additional amino acid change from 'D' to 'Y' was
158 identified as homoplasic between some SARS-CoV-1 and SARS-CoV-2 sequences (data
159 derived from the global, expanded and the re-sampled SARS-CoV-2 trees) (Table 1). For
160 SARS-CoV-2, amino acid state 'Y' emerged and was lost repeatedly during the early stage of
161 the pandemic (represented by independent minor clusters that became extinct). However,
162 following emergence and global spread of the B.1.1.529 virus lineage (Omicron variant of
163 concern [VOC], and descending sub-lineages), amino acid state 'Y' replaced amino acid state
164 'D', displaying a prevailing trend associated with the dominance of the B.1.617.2 lineage (Delta
165 VOC, and descending sub-lineages) (Table 1, Figure 3, Supplementary Data 3) (also
166 confirmed by independently sampled SARS-CoV-2 data available up to December 2022:
167 https://nextstrain.org/groups/neherlab/ncov/global?c=gt-S_796).
168
169 Quantifying the effects of positive selection
170 The dN/dS estimates we obtained across complete virus genomes and upon specific coding
171 regions (see Methods section 4) indicates that positive selection is acting upon the Orf1ab
172 and Orf S of SARS-CoV-2, compared to other viruses studied here. Specifically, the effect of
173 episodic diversifying selection was detected upon 5/14 non-recombinant fragments (three in
174 Orf1b and two in Orf S, for details see https://observablehq.com/@spond/beta-cov-analysis).
175 Using the Contrast-FEL method to detect the effect of a differential selection across branches
176 separating lineages (see Methods section 4), we found 36 sites (0.4%) evolving under
177 differential selective pressure across distinct virus clades. Furthermore, we found 0.7% of all
178 sites (67 codons within the Orf1ab+S alignment) to be evolving under episodic diversifying
179 positive selection (scored under MEME with a $p \leq 0.05$ as *positively-selected sites*, or PSS)
180 (Supplementary Table 2). In contrast, we found 5% of all sites (461 codons within the
181 Orf1ab+S alignment) to be evolving under pervasive negative selection (scored under FEL
182 with a $p \leq 0.05$ as *negatively-selected sites,* or NSS). We subsequently mapped the identified
183 PSS and NSSs onto the SARS-CoV-2 S protein structure (Methods section 5). We observe
184 that out of a total of 22 PSSs, 18 locate within the S1 subunit (11 in $S1^A$, 5 in $S1^B$, 1 in $S1^C$

185 and 1 in $S1^D$ domains), whilst the remaining four mapped onto the S2 subunit. Conversely, out
186 of a total of 82 of NSSs, 46 locate within S1 (18 in $S1^A$, 21 in $S1^B$, 3 in $S1^C$ and 4 in $S1^D$), whilst
187 the remaining 36 mapped onto S2 (Supplementary Figure 7).
188 From the 30 non-synonymous mutations we identify as displaying evolutionary
189 patterns putatively denoting homoplasy and/or stepwise evolution (Table 1), sites 19048,
190 21623, 21635, 22124 and 23048 were further scored as PSS (under different methods). Sites
191 21623 and 21635 were inferred as PSSs along ancestral branches leading to the HKU1, OC43
192 and SARS-CoV-2 clades. Sites 19048 and 22124 were inferred as PSSs along the OC43
193 ancestral branch, whilst 23048 was inferred as a PSS along the HKU1 ancestral branch (Table
194 1, Supplementary Table 2). Further analysis under the branch and site model in the MEME
195 method (Methods section 4) revealed site 18121 to be evolving under positive selection for
196 the HKU1 clade/branch (relative to the sarbecoviruses), in agreement with our observations
197 made on putative homoplasy detected for this site between HKU1 genotype B, SARS-CoV-1
198 and SARS-CoV-2 (Table 1, Figure 3). Similarly, site 23948 was also inferred to be evolving
199 under positive selective for the SARS-CoV-1 branch, relative to other virus clades
200 (Supplementary Table 2).
201 For validation purposes, we compared our results with selection analysis available for
202 independently sampled SARS-CoV-2 genome data available as of December 2022
203 (https://observablehq.com/@spond/evolutionary-annotation-of-sars-cov-2-covid-19-
204 genomes-enab) (Kosakovsky Pond). Of the 30 mutations we identify, 16 of these are currently
205 scored as PSS or NSSs, with 13 of these mapping directly onto potential T cell epitopes
206 derived from HLA class I and HLA-DR binding peptides in SARS-CoV-2 (Campbell et al. 2020;
207 Nelde et al. 2021) (Table 1). Additionality, up to December 2022, sites 7478, 21614, 23948,
208 24620 and 25166 were detected as evolving under positive selection, whilst sites 21635,
209 24863, and 25037 were detected as evolving under negative selection.
210
211 Contextualization of mutations using protein structural and functional information
212 For this purpose, we mapped the 30 mutations identified (Table 1) onto corresponding protein
213 structures (Methods section 5). Below, we focus on the potential functional impact identified
214 for four sites (18121, 21623, 21635 and 23948) that meet three criteria of: displaying evidence
215 of homoplasy and/or stepwise evolution, showing evidence of evolution under positive
216 selection, and being proximal to regions of established protein function. A description for all
217 other identified mutations is available in the Supplementary Text 4 and Supplementary Table
218 3.
219
220 **Site 18121 in Orf1ab**

221 The Orf1ab gene encodes 16 non-structural proteins (nsp1-16), primarily involved in viral RNA
222 synthesis and processing (V'kovski et al. 2020). Site 18121 is located within Orf1ab and
223 corresponds to an 'S' to 'A' mutation at residue 28 within the exonuclease domain of the nsp14
224 protein (numbering according to the SARS-CoV-1 protein) (Supplementary Table 3, Figure 4).
225 Nsp14 functions as a methyltransferase, and is involved in 5′-capping of the viral mRNA (Ma
226 et al. 2015). The cap core structure is essential for viral mRNA transcription, but is also
227 implicated in protecting the 5′-triphosphate from activating the host innate immune response
228 (Wang et al. 2015). An 'S' to 'A' mutation within this region is expected to result in the loss of
229 an intra-protein hydrogen-bond (formed with the main chain of residue T25 of nsp14, and
230 between nsp14 and its activator nsp10) (Ma et al. 2015), potentially modulating protein-protein
231 interaction (Figure 4) (assessed under PISAebi;
232 http://www.ebi.ac.uk/pdbe/prot_int/pistart.html) (Krissinel and Henrick 2007).
233
234 **Sites 21623 and 21635 in S1**
235 The S1 subunit of S (spike) protein mediates attachment of the virus to the host cell (Li 2016).
236 Humans-infecting embecoviruses bind to glycan-based cell receptors terminating in 9-*O*-
237 acetylated sialic acids, and receptor recognition is accomplished via two hydrophobic pockets
238 separated by a conserved 'W' residue within the $S1^A$ region of the protein (Hulswit et al. 2019;
239 Tortorici et al. 2019). In contrast, the receptor-binding site for human-infecting sarbecoviruses
240 consists of an extended loop located within the $S1^B$ domain of the protein (Li et al. 2005; Lan
241 et al. 2020; Shang et al. 2020). Despite a limited degree of conservation amongst contact
242 residues within the RBD of SARS-CoV-1 and SARS-CoV-2, both virus species recognize the
243 ACE2 molecule to enter the host cell (Li et al. 2005; Lan et al. 2020).
244 Site 21623 displays several non-synonymous mutations ('R', 'V', 'K' and 'I') mapping
245 to residue 29 of the OC43 S protein, located within the $S1^A$ domain of the S1 subunit
246 (Supplementary Table 3, Figure 5). Site 21635 again shows multiple non-synonymous
247 mutations 'P', 'V', 'S', 'L' and 'H' mapping to residue 33 of the OC43 S protein. Residue 29
248 (site 21635) is just four positions downstream of residue 33 (site 21623) (Figure 5). Within the
249 OC43 S protein, both residues 29 and 33 are located within a loop neighbouring the
250 hydrophobic pockets in $S1^A$. This area is instrumental for receptor recognition, and changes
251 within the region can modulate receptor affinity (Hulswit et al. 2019). Our observations on both
252 these sites displaying mutational patters putatively denoting homoplasy/stepwise evolution
253 and evidencing positive selection (Table 1), are congruent with antigenic drift shaping the
254 evolution of human-endemic coronaviruses (Kistler and Bedford 2021).
255 As a note, in the case of SARS-CoV-2, mutations in both these sites have been
256 observed for two VOCs (B.1.351 and P.1, 'Beta' and 'Gamma') (Faria et al. 2021; Tegally et
257 al. 2021). Particularly, site 21623 corresponds to variable non-synonymous mutations ('R', 'I',

258 'T, 'K and 'G') mapping to residue 21 in the S protein (as confirmed by independently sampled
259 SARS-CoV-2 genome data available up to December 2022:
260 https://nextstrain.org/groups/neherlab/ncov/global?c=gt-S_21). However, given that
261 sarbecoviruses engage with the ACE2 receptor via domain $S1^B$, mutations observed within
262 domain $S1^A$ may reflect relaxed constraints in the local protein surface, unlikely to be related
263 to the receptor's functionality.
264
265 **Site 23948 in S2**
266 Virus binding to the cell surface is followed by fusion of the viral envelope with the host cell
267 membrane, thereby delivering the viral genome into the target cell. The S2 subunit harbours
268 the viral fusion machinery and displays characteristic structural features of class I fusion
269 proteins. This includes a fusion peptide that anchors the virus to the host membrane, initiating
270 fusogenic conformational changes that result in merging of the two membranes (Li 2016).
271 However, for this process to occur, the S protein needs to be primed through cleavage initiated
272 by host cell proteases (Millet and Whittaker 2015). In coronaviruses, cleavage occurs at a
273 conserved site referred to as the S1-S2 junction (consensus RRAR|S in SARS-CoV-2), and is
274 further mediated by an additional R|S site located immediately upstream of the fusion peptide
275 in the S2 subunit (referred to as the S2' cleavage site) (Millet and Whittaker 2015).
276 For site 23948, the observed mutation 'D' to 'Y' (identified as homoplasic between
277 some SARS-CoV-1 and SARS-CoV-2 sequences) corresponds to a non-conservative amino
278 acid replacement at residue 796 of the S2 subunit (SARS-CoV-2 numbering) (Table 1,
279 Supplementary Table 3). This residue is located immediately upstream of the S2' cleavage
280 site, within the region that crucially regulates the release of the fusion peptide. Residues
281 located between position 796 and the fusion peptide form a loop, displaying variability across
282 betacoronavirus species (Supplementary Figure 6), suggesting that the apparent relaxed local
283 constraints may serve to secure loop accessibility for cleavage activation. In agreement with
284 this hypothesis, the corresponding protein region within the HKU1 structure remains
285 unresolved (Kirchdoerfer et al. 2016).
286
287 **DISCUSSION (1260)**
288 In this study, we searched for signatures of adaptive convergence across distantly related
289 human-infecting betacoronaviruses, represented by shared non-synonymous mutations that
290 putatively denote homoplasy and/or stepwise evolution. The mutations we identified were
291 further ranked according to their selective relevance, and to their proximity to protein regions
292 of known function. The majority of the mutations we observe locate to the receptor binding
293 region of the S protein (*i.e.*, S1 subunit), whilst a smaller proportion of these were found within
294 non-structural proteins encoded by Orf1ab (site 18121 in the exonuclease domain of nsp14,

295 and site 20344 in the endonuclease domain of nsp15). Further *in-silico* analyses revealed four
296 genomic sites (18121, 21623, 21635 and 23948) that display cumulative evidence of: *i*) a
297 mutational pattern putatively denoting homoplasy and/or stepwise evolution, *ii*) evolution
298 under positive selection, and *iii*) being structurally proximal to regions of known protein
299 function. Below, we discuss our findings in light of three key evolutionary processes: antigenic
300 drift, epistasis and adaptive convergence.

301 The host humoral immune response is an evolutionary force driving antigenic drift in
302 viruses. In the case of betacoronaviruses, this is reflected by cumulative mutations in the S
303 protein (particularly within the S1 subunit) that may allow frequent reinfections of the host
304 population (Forni et al. 2021; Kistler and Bedford 2021; Yewdell 2021). In agreement with this
305 observation, the emergence of some SARS-CoV-2 lineages (particularly VOC) has been
306 associated with high levels of infection in pre-exposed human populations across different
307 geographic regions (as an example on P.1, see Faria et al. 2021). Our results evidence
308 antigenic drift upon the S1 subunit of distinct betacoronaviruses as a major component of the
309 adaptation process to the human host environment, further evidenced by Orf S also being the
310 least conserved genome region across distinct virus species (Li 2016). On the other hand,
311 mutations found within Orf1ab could have a potential impact on viral fitness related to an
312 enhanced replication efficacy in the human host (Menachery et al. 2017). As the evolution of
313 Orf1ab is also driven by immune responses such as cytokine signalling cascades and antigen
314 presentation (Wang et al. 2015; Hackbart et al. 2020; Taefehshokr et al. 2020; Yuen et al.
315 2020), these mutations may also be the result of concerted selective pressure(s), following
316 the principle that single mutational changes can have pleiotropic effects on distinct viral
317 phenotypes and fitness components (de Wilde et al. 2018).

318 Identifying adaptive convergence raises the possibility of predicting mutational
319 pathways in viruses relevant to global health (Gutierrez, Escalera-Zamudio, and Pybus 2019).
320 When applied to SARS-CoV-2, our results reveal that some of the mutations we had initially
321 identified as potentially relevant back in May 2021 (see (Escalera-Zamudio et al. 2021) had
322 already been observed in other betacoronaviruses that circulate endemically in humans (Table
323 1), and some now display dominant trends in SARS-CoV-2 (as analysed up to December
324 2022). For example, amino acid state 'R' at residue 21 of the S protein (sites 21623)
325 (https://nextstrain.org/groups/neherlab/ncov/global?c=gt-S_21) and 'P' at residue 25 (site
326 21635) (https://nextstrain.org/groups/neherlab/ncov/global?c=gt-S_25) have dominated
327 across time. Moreover, mutation 'D' to 'Y' observed at residue 796 of the S protein (site 23948)
328 has proven to be a successful mutational pathway, evidenced the replacement of state 'D'
329 (previously observed for the B.1.617.2 lineage, Delta VOC and descending sub-lineages) by
330 state 'Y' (now observed for the B.1.1.529 lineage, Omicron VOC and descending sub-
331 lineages) (https://nextstrain.org/groups/neherlab/ncov/global?c=gt-S_796). Of interest,

332 mutations at residue 796 of the S protein have been linked to the emergence viral variants
333 that display reduced susceptibility to neutralizing antibodies (Kemp et al. 2021).
334 Epistasis is thought to have played a central role in the emergence of human-infecting
335 betacoronaviruses (Holmes and Rambaut 2004). However, inferring epistasis across
336 diverging viruses is difficult given the functional differences between homologous genes and
337 proteins. Through our methodological approach we cannot measure epistasis *per se*, but we
338 can aim to identify adaptive convergence and subsequently discuss its possible effects in the
339 light of epistasis. Thus, our results indirectly provide support for epistasis, in the sense that if
340 the same amino acid changes are observed in different virus species, then associated
341 epistatic interactions are expected to be shared. This is of particular importance when
342 considering the potential role of epistasis in antigenic drift, where the combined effect of
343 independent mutations could contribute to antigenic escape (Rochman et al 2022). In the
344 context of our findings, sites 21623 and 21635 are presumed to be involved in the antigenic
345 drift of embecoviruses. As these residues are in close proximity to each other (displaying a
346 linked evolution), these could thus reflect epistatic interactions. Nevertheless, within the
347 SARS-CoV-2 $S1^B$-ACE2 interface, epistasis seems to play a limited role, as the effect of
348 multiple mutations seems to be additive rather than epistatic (Zahradník et al. 2021; Rochman
349 et al. 2022; Starr et al. 2022).
350 The mutational spectrum of SARS-CoV-2 is known to be impacted by the human host
351 apolipoprotein B mRNA-editing enzyme (APOBEC) family (Di Giorgio et al. 2020). The activity
352 of APOBEC induces C → U/T mutations in the viral genome through a cytidine deaminase
353 activity, likely resulting in a high degree of apparent homoplasy, reflected in emerging
354 mutations across distinct virus sub-populations (De Maio, et al. 2020; Worobey, et al. 2020;
355 Wang, et al. 2021). Relative to more commonly used strategies for identifying homoplasy
356 within single virus species, our methodology poses an alternative approach that aims to
357 identify homoplasy *across and within* virus taxa, as represented by shared mutations fixed
358 under an evolutionary scenario mostly driven by selection (see Supplementary Text 5). Given
359 that candidate mutations are observed over longer evolutionary times, this approach
360 represents a useful tool that decreases the likelihood of erroneously scoring mutations as
361 homoplasic (such as those resulting from mutational biases inherent to the SARS-CoV-2
362 genome).
363 However, identifying adaptive convergence faces several important limitations. First,
364 the methodology we use is conservative, as it is based on strict homology. In this context, we
365 only consider sites robustly identifiable as homologous that can be traced back to ancestral
366 nodes with confidence (consequently excluding highly divergent genes). Therefore, our
367 approach may result in an underestimation of sites that may putatively denote adaptive
368 convergence across divergent virus species. Moreover, a limited virus genome sampling

369 across time and space (in particular for HKU1 and SARS-CoV-1), coupled with a relatively low
370 genetic diversity observed for SARS-CoV-2 (Rausch et al. 2020), further restricts the potential
371 to identify shared mutations across virus species (van Dorp et al. 2020). In addition, there is
372 some uncertainty associated with mutations that may derive from processes such as founder
373 effects, mutational hitchhiking, and toggling at hypervariable sites (Delport et al. 2008; Pond
374 et al. 2012; Simmonds 2020; Wang et al. 2020; De Maio et al. 2021). Finally, whilst our
375 analysis provides insights into coronavirus evolution in humans, our approach renders us
376 unable to identify mutations that may result from host switching events. This is due to analyses
377 on nodes representing ancient host switching events (Corman et al. 2018) being constrained
378 by long divergence times, differences in mutation rates across virus taxa in different hosts,
379 mutational saturation, and by a considerable under-sampling of betacoronaviruses circulating
380 in non-human hosts (Holmes and Rambaut 2004; De Maio et al. 2021).
381 In this sense, additional/future experimental data could help reveal the impact of
382 mutations on viral fitness. However, performing such studies may be difficult, as these concern
383 potential gain-of-function experiments. Alternatively, enhanced genomic surveillance of
384 betacoronaviruses infecting the human population and of those ciruclating in other animal host
385 may confirm whether the mutational pathways we identify here represent evolutionary
386 trajectories on which betacoronaviruses converge in their adaptation process to the human
387 host.
388
389 **MATERIAL AND METHODS (1888)**
390 **1. Initial data collation**
391 When this manuscript was first deposited as a preprint (May 2021) (Escalera-Zamudio et al.
392 2021), complete genomes for all HKU1, OC43 and SARS-CoV-1 viruses sampled across
393 different geographical regions and time were downloaded from the Virus Pathogen Resource
394 (ViPR-NCBI 2021) (Supplementary Data 4). Sequences were removed if meeting any of the
395 following criteria: (i) being >1000nt shorter than full genome length, (ii) being identical to any
396 other sequence, or (iii) if showing >10% of site were ambiguities (including N or X). A total of
397 53 HKU1, 136 OC43 and 40 SARS-CoV-1 sequences were initially retained for analyses. For
398 SARS-CoV-2, to better reflect an early zoonotic process into the human population (MacLean
399 et al. 2021), we originally aimed to limit the genetic diversity of the sampled virus population
400 to the first wave of infection recorded during the pandemic. For this, ~23000 full genomes
401 sampled worldwide before May 2021 available in the GISAID platform (GISAID 2021) were
402 downloaded and aligned as part of an initial public dataset provided by the COG-UK
403 consortium (COG-UK 2021) (Supplementary Data 4). To make local analyses computationally
404 feasible, the original SARS-CoV-2 dataset was randomly subsampled to ~5% of its original
405 size, keeping the earliest genomes, and further reducing the dataset under the quality criteria

406 stated above. In total, 1120 SARS-CoV-2 sequences were retained. For all virus species
407 considered in this study, we focused only on genomes derived from human cases, in order to
408 reflect host-specific adaptation processes.
409
410 **2. Phylogenetic analysis**
411 Only the main viral ORFs (Orf1ab and S) were used for further phylogenetic analysis, as these
412 are homologous amongst the four viral species studied, and also encode proteins that are
413 essential for viral function (*i.e.,* the genome replication machinery, receptor binding, and the
414 virus-host membrane fusion apparatus) (Yoshimoto 2020). For each virus species, individual
415 ORFs were extracted and aligned using MAFFT v7.471 (Katoh and Standley 2013), and were
416 concatenated to generate Orf1ab+S alignments (Supplementary Data 1). The concatenated
417 alignments were then combined to generate a global alignment dataset that was then re-
418 aligned at amino acid level using a profile-to-profile approach following taxonomic relatedness
419 (Wang and Dunbrack 2004). This global alignment comprised in total 1314 sequences and
420 26883 sites.
421 Maximum Likelihood phylogenies were estimated for the individual and global
422 alignments using RAxML v8 (Stamatakis 2015), under a general time reversible nucleotide
423 substitution model and a gamma-distributed among-site rate variation (GTR+G). Branch
424 support was assessed using 100 bootstrap replicates. All trees were midpoint-rooted, whilst
425 general phylogenetic patterns observed amongst these distantly related virus species were
426 validated by comparing to previously published phylogenies (Woo et al. 2006; Woo et al. 2010;
427 Lau et al. 2011; Oong et al. 2017; Zhu et al. 2018; Bedford 2021). Recombination is known to
428 be common amongst betacoronaviruses (Woo et al. 2006; Su et al. 2016; Oong et al. 2017),
429 including SARS-CoV-2 (Gutierrez et al. 2022; Turakhia et al. 2022). However, recombinant
430 sequences were not removed at this step, as it was important to detect potentially recombinant
431 isolates that could display relevant mutations. Putative recombinant sequences were
432 eventually removed for subsequent analyses (when identified, see Methods sections 6 and
433 7).
434
435 **3. Identifying homoplasy and/or stepwise evolution**
436 Following the pipeline described by Escalera & Golden (Escalera-Zamudio et al. 2020),
437 variable sites across different virus taxa were identified within the global alignment as those
438 displaying non-synonymous mutations (rendering amino acid changes) occurring in at least
439 ≥1% of the sampled sequences. Variable sites were extracted by masking columns across the
440 alignment showing identical sites and at least 50% gaps, followed by the 'Find
441 Variations/SNPs' function used to compare each site to a consensus generated under a 95%
442 threshold with Geneious Prime v2020.0.4 (Kearse et al. 2012). A total of 6681 variable sites

were identified and used to infer ancestral amino acid state reconstructions onto the nodes/internal branches of the global tree (see Methods section 2 above). This was done using TreeTime (Sagulenko et al. 2018) under a ML approach (RAS-ML) under a time-reversible model (GTR) for state transitions. The genetic variability observed within leaves/tips of the tree was deliberately excluded, in order to only analyse changes occurring within nodes or internal branches of the tree. In parallel, conserved sites were identified as those present in ≥ 99% of the sampled virus sequences. Conserved sites were extracted by reversing the 'variable site masking', to obtain only identical sites identified across the global alignment (Supplementary Data 2).

The resulting 6681 'Ancestral Reconstruction Trees' (named here ARTs) were then classified under a computational algorithm developed to sort mutational patterns based on whether or not they support homoplasy and/or stepwise evolution. Briefly, homoplasy can occur within nodes of single clade or across clades, in which the same amino acid change must be present in at least one internal node of any given clade, and in another internal node of the same/another clade. Clades with the same amino acid states must not share direct common ancestry. Conversely, stepwise evolution is represented as sequential mutations occurring at the same sites within a single clade. Any given site scored under putative 'stepwise evolution' must display changes between at least two different states (A→B), but without any immediate reversion (B→A). A full description of the definitions used here for homoplasy and/or stepwise evolution are available as Supplementary Text 1 and Supplementary Figure 1. A full description of the basic steps used in our algorithm, including a schematic representation, is available in the Supplementary Text 2, Supplementary Figure 2, and Supplementary Figure 3. Associated code is publicly available at https://github.com/nataliamv/SARS-CoV-2-ARTs-Classification.

4. Estimating dN/dS

Derived from the global alignment and tree, we estimated dN/dS (ω, the ratio between the non-synonymous substitution rate per non-synonymous site and the synonymous substitution rate per synonymous site) using site, branch and branch-site 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. 2021). For this, the alignment was partitioned into 14 putatively non-recombinant regions using the Genetic Algorithm for Recombination Detection (GARD) (Kosakovsky Pond, Posada, et al. 2006), with all subsequent analyses conducted on the partitioned data. As dN/dS models use the GTR component for the nucleotide evolutionary rate, biased mutation rates are handled. Further, to mitigate the inflation in dN/dS estimates that results from unresolved and/or maladaptive evolution, testing for selection was again restricted to internal

480 nodes/branches of the phylogeny (Kosakovsky Pond, Frost, et al. 2006). Genome-wide
481 comparison of dN/dS estimates across viral genome regions was performed using the Branch-
482 Site Unrestricted Statistical Test for Episodic Diversification method (BUSTED) (Murrell et al.
483 2015). Finally, the impact of changing biochemical properties at selected sites was further
484 assessed under the Property Informed Models of Evolution method (PRIME) (HyPhy 2013).
485 Our results were further compared to the selection analysis available for independently
486 sampled SARS-CoV-2 genome data available as of December 2022
487 (https://observablehq.com/@spond/evolutionary-annotation-of-sars-cov-2-covid-19-
488 genomes-enab) (Kosakovsky Pond).
489
490 **5. Mapping mutations onto betacoronavirus protein structures**
491 To locate the non-synonymous mutations identified on viral protein regions of known function,
492 corresponding residues were mapped to available structural data using PyMOL v 2.4.0
493 (https://pymol.org/2/) (Supplementary Table 3, see Data Availability section). Mutations were
494 analysed in the context of their relative proximity to previously reported functional regions, and
495 to each other. N-linked glycosylation sites in S protein sequences were identified by searching
496 for the N-[not P]-[S or T] consensus sequence (Watanabe et al. 2019). None of the mutations
497 identified in this study resulted in generation or deletion of N-linked glycosylation sequons. In
498 parallel, the identified conserved and variable sites as well as the 30 mutations evidencing
499 homoplasy and/or stepwise evolution across virus species were mapped onto published
500 protein structures available for the S proteins of the four human-infecting betacoronaviruses
501 studied here (Figure 5, Supplementary Figure 7). Finally, to compare dN/dS distributions
502 between specific domains of the S protein within and across virus species, sites inferred to be
503 under positive or negative selection (PSS, NSS) were mapped onto S protein structures
504 (Supplementary Data 2).
505
506 **6. Validation through resampling and by comparing mutational distributions**
507 To validate our initial observations derived from virus genomes sampled up to May 2021, we
508 sought to determine if the 30 mutations that had been identified initially were also present in
509 the expanded embecov- and sarbecovirus diversity sampled up to July 25th 2022
510 (corresponding to the final sampling date of this study). Virus diversity now included genome
511 sequences derived from more recently collected human isolates (only made publicly available
512 after our initial sampling), and from other closely related embeco- and sarbecoviruses from
513 non-human hosts. The expanded alignment comprises 1455 sequences (~700 embecovirus
514 + SARS-CoV and ~700 SARS-CoV-2), resulting in 27503 columns that were re-aligned under
515 a progressive profile-to-profile approach based on taxonomic relatedness to be further used
516 to estimate an expanded 'Maximum Likelihood' tree (following Methods section 2). To

517 additionally explore if the mutations identified were also present in a larger dataset
518 representing an expanded SARS-CoV-2 diversity (sampled up to July 25th 2022), a set of 1400
519 SARS-CoV-2 genomes denoting 'evolutionary successful' virus lineages (Supplementary
520 Table 1) was examined independently (Supplementary Text 3, Supplementary Figure 5). Both
521 datasets were analysed following the steps described in Methods Section 2 and 3, specifically
522 searching for the mutations listed in Table 1. Virus taxa included in both re-sampled datasets
523 are listed in Supplementary Data 5. A full description of the subsampling methodological
524 approach used is available as Supplementary Text 3, and Supplementary Figure 4.
525 We further sought to explore if the proportion of mutational patterns we classified as
526 putatively denoting homoplasy and/or stepwise evolution were more likely to arise from an
527 evolutionary scenario mostly driven by selection, compared to 'random' mutational patterns
528 derived from evolutionary scenarios generally driven by genetic drift. For this purpose, the
529 expanded alignment was translated to amino acid sequences and used to simulate three
530 alignments with 'AliSim' (http://www.iqtree.org/doc/AliSim) under the 'mimick real alignment'
531 function (mimicking a 'real' evolutionary process based on amino acid evolution under a LG
532 model, and applied to the inputted original tree). To compare the corresponding proportion of
533 sites scored under homoplasy and/or stepwise evolution, each dataset (the expanded and
534 three simulated alignments) was analysed following the steps described in Methods section
535 3. The classification of mutation patterns within expanded and simulated datasets also serves
536 the purpose of validating our algorithm, originally developed for analysing the global dataset
537 (including merely OC43, HKU1, SARS-CoV-1 and SARS-CoV-2). Associated results and a
538 brief discussion are available as Supplementary Text 5.
539
540 **7. Reconstruction of amino acid evolution for selected sites**
541 For those mutations displaying cumulative evidence of being potentially informative on
542 adaptive convergence (sites 18121, 21623 and 23948) (Table 1), we used the expanded
543 dataset to infer ancestral states under a Bayesian framework. For this, we first estimated an
544 MCC (maximum clade credibility) tree from the resampled codon alignment using a SRD06
545 substitution model (Shapiro et al. 2006) and a strict molecular clock. For each site of interest,
546 coded amino acid traits were mapped onto the nodes of the MCC tree by performing
547 reconstructions of ancestral states under an asymmetric discrete trait evolution model (DTA)
548 in BEAST v1.8.4 (Lemey et al. 2009; Suchard et al. 2018). The DTA model was run using a
549 Bayesian Skygrid tree prior for $100X10^6$ generations and sampled every 10,000 states until all
550 DTA-relevant parameters reached an ESS >200. The resulting trees were used as an input
551 for Figure 3.
552
553 **FIGURE LEGENDS**

554
555 **Figure 1. Phylogenetic tree of human-infecting betacoronaviruses**. The expanded tree
556 estimated from the Orf1ab+S alignment comprising 1455 sequences (see Methods section 6),
557 summarizing the phylogenetic pattern observed for four distantly related human-infecting
558 betacoronaviruses: HKU1, OC43, SARS-CoV-1 and SARS-CoV-2. MERS and related virus sequences
559 were included in the tree for rooting purposes only. Both the *Embecovirus* subgenus (HKU1 and OC43
560 and related viruses) and the *Sarbecovirus* subgenus (SARS-CoV-1 and SARS-CoV-2 and related
561 viruses) are indicated, showing the positioning of the most closely related virus genome sequences
562 derived from animal isolates (when available). The different genotypes identified for the HKU1 (A, B
563 and C) and for the OC43 (A–H) are shown in Supplementary Data 3.
564
565 **Figure 2. Distribution of conserved/variable sites with S across different virus species**
566 (a) Top-down (upper panel) and side view (bottom panel) of a cartoon representation of the multidomain
567 architecture of the trimeric SARS-CoV-2 S ectodomain (PDB: 6VXX). The S1 subunit is coloured
568 according to the different protein domains: $S1^{A}$ in cream, $S1^{B}$ in teal, $S1^{C}$ in orange, and $S1^{D}$ in blue,
569 whilst the S2 subunit is shown in grey. (b) Top-down and side views of sphere-based representations
570 of trimeric S protein ectodomains for the viruses studied here: SARS-CoV-2 (PDB: 6VXX), SARS-CoV-
571 1 (PDB: 6ACC), OC43 (PDB: 6OHW) and HKU1 (PDB: 5I08). The sphere-based representation shows
572 conserved (shown in grey; residues present ≥99% of all sequences) and variable sites (blue; residues
573 present in ≥1% of all sequences) across virus species. Variable sites identified as denoting homoplasy
574 or stepwise evolutionary patterns are shown in red (see Methods section 3). The asparagine residues
575 of N-linked glycosylation sequons are indicated in purple.
576
577 **Figure 3. Reconstruction of amino acid evolution at selected sites**. Maximum clade
578 credibility (MCC) trees showing the evolutionary trajectories for different amino acid states observed for
579 three illustrative sites (18121, 21623 and 23948) that (i) display evidence of homoplasy and/or stepwise
580 evolution, (ii) show evidence of positive selection, and (iii) are proximal to regions of established protein
581 function. The reconstructions of ancestral states for these sites show different amino acid states at
582 nodes (represented by circles in different colours). The posterior probability for a given amino acid state
583 occurring at a given node of interest is indicated. Sites 18121 display evidence of homoplasy across
584 virus lineages, site 21623 shows evidence of both homoplasy across species and stepwise evolution
585 within single virus species (*i.e.* OC43), and site 23948 shows evidence of stepwise evolution within
586 single virus species (*i.e.* SARS-CoV-1), and also of homoplasy across virus species (*i.e.* SARS-1 and
587 SARS-CoV-2).
588
589 **Figure 4. Residue Ser^{28} of nsp14 is situated near the nsp14-nsp10 interface.** Cartoon
590 representation of the SARS-CoV-1 nsp14-nsp10 protein complex (PDB: 5C8S) with Ser^{28}
591 (corresponding to site 18121 in SARS-CoV-2 genome coordinates) shown as a red sphere. This residue
592 is located within the nsp14 ExoN domain (cream) and is approximately 9 Å from the interface with nsp10
593 (light blue, the proximal nsp10 residue Cys^{41} was used to calculate the distance and is indicated as a

594 sphere). The distance between nsp14's Ser^{28} and nsp10's Cys^{41} is annotated and indicated by a dashed
595 black line. Zoomed-in panel: detailed representation of the intra-nsp14 hydrogen-bond between the
596 side chain of Ser^{28} and the main chain of Thr^{25} (identified with the PISAebi server). The side chain of
597 Ser^{25} is indicated as a red stick and Thr^{25} is indicated in sticks and coloured according to atom (C,
598 cream; O, red; N, blue). The hydrogen-bond is indicated as a dashed black line.
599
600 **Figure 5. Mapping of mutations exhibiting homoplasy onto the S protein structure of**
601 **SARS-CoV-2.** Top-down (left) and side view (right) of a cartoon representation of the multidomain
602 architecture of the trimeric SARS-CoV-2 S ectodomain (PDB: 6VXX). The S2 subunit is highlighted in
603 grey and the S1 ectodomain is divided into $S1^A$ (highlighted in cream), $S1^B$ (teal), $S1^C$ (orange), and
604 $S1^D$ (blue) domains, following the colour scheme in Figure 3. Homoplasic mutations that co-localize to
605 known functional surfaces (see Supplementary Table 4) are indicated in the structure and coloured in
606 groups: Arg^{21} (corresponding to site 21623 in SARS-CoV-2 genome coordinates, in green), Pro^{25} (site
607 21635, in green), Asp^{796} (site 23948, in yellow), Ile^{1018} (site 24614, in red), Ala^{1020} (site 24620, in red)
608 and Leu^{1024} (site 24632, in red). All representations are shown with a transparent protein surface for
609 clarity.
610
611 **ACKNOWLEDGEMENTS**
612 MEZ was supported by Leverhulme Trust ECR Fellowship (ECF-2019-542). RJGH is
613 supported by the European Molecular Biology Organisation (ALTF 869-2019). RJGH and TAB
614 are supported by the Medical Research Council (MR/S007555/1). The Wellcome Centre for
615 Human Genetics is supported by Wellcome Trust grant 203141/Z/16/Z. SKP is supported in
616 part by the the NIH (grants AI134384, AI140970, GM110749) and the NSF (grant 2027196).
617 JT and RPDI were supported by European Union's Horizon 2020 project MOOD (874850).
618 LvD is supported by a UCL Excellence Fellowship. HGCS is supported by funding through the
619 "Vigilancia Genómica del Virus SARS-CoV-2 en México" grant from the National Council for
620 Science and Technology-México (CONACyT). Author contributions: MEZ and OGP designed
621 research. MEZ and RJGH performed research. MEZ, RJGH, BG, SKP, LvD, JT, RPDI and
622 HGCS analysed data. NM developed the code for implementing the computational pipeline.
623 OGP and TAB supervised data analysis. MEZ and RJGH wrote the manuscript, with
624 comments from all authors. We thank Dr. Louis Du Plessis for his help with the phylogenomic
625 analysis and Dr. Nicola De Maio for constructive comments on our analytical approach.
626
627 **COMPETING INTERESTS**
628 The authors declare no competing interests.
629
630 **DATA AVAILABILITY STATEMENT**

631 Taxa IDs and accession numbers for virus sequences used in this study are provided in the
632 Supplementary Data 4 and 5 files. All SARS-CoV-2 genome sequences and associated
633 metadata used in this study are published in GISAID's EpiCoV database under the EPI SET
634 GISAID Identifier: EPI_SET_230131zy. To view the contributors of each individual sequence
635 with details such as accession number, virus name, collection date, originating and submitting
636 lab, as well as the list of all authors, visit 10.55876/gis8.230131zy. PBD files used are listed
637 as follows: S protein (HKU1 PDB:5I08, OC43 PDB:6OHW, SARS-CoV-1 PDB:6ACC and
638 SARS-CoV-2 PDB:6VXX, 6ZGI). Orf1a (SARS-CoV-1 nsp3 PDB:2W2G). Orf1b (SARS-CoV-
639 2 nsp13 PDB:6XEZ, SARS-CoV-1 nsp14 PDB:5C8S and SARS-CoV-2 nsp15 PDB:6WLC).
640 Full code for our algorithm is available as open source: https://github.com/nataliamv/SARS-
641 CoV-2-ARTs-Classification. An interactive notebook with our full selection analysis results is
642 available at https://observablehq.com/@spond/beta-cov-analysis.
643

Table 1. Potentially relevant sites across human-infecting betacoronaviruses

SARS-CoV-2 genome coordinates †	ORF	Protein/ Residue †	Mutations observed (global tree): Ancestral LinA	OC43	HKU1	Ancestral LinB	SARS-CoV-1	SARS-CoV-2§	Confirmed in re-sampled trees: Expanded	1400 SARS-CoV-2	Homoplasy (H)/ Stepwise Evolution (SWE)	Selection across species (Method, p-value) †#	Selection in SARS-CoV-2 (recent amino acid changes)¶	Epitopes*
2557	Orf1a	nsp2 585	P	P	S	S	A/T	P/S	Yes	Conserved in SARS-CoV-2	H/SWE			0
7478	Orf1a	nsp3 1587	N	S/N	N	N	T	N	Yes	Conserved in SARS-CoV-2	H		PSS, N→S/D (OC43-like and new state)	0
16189	Orf1b	nsp12 917	D	D	E/D	E	E	E	Yes	Conserved in SARS-CoV-2	H/SWE	Overall negative selection (FEL 0.02)		1
17809	Orf1b	nsp13 525	V	V	V/I	I	I	I	Yes	Conserved in SARS-CoV-2	H			0
18121	**Orf1b**	**nsp14 28**	**A**	**A**	**A/S**	**S**	**S**	**S**	Yes	Conserved in SARS-CoV-2	**H**	**Different overall positive selection (CF 0.022)**		**1**
18334	Orf1b	nsp14 100	D	D	E/D	E	D	E	Yes	Conserved in SARS-CoV-2	H/SWE	Overall negative selection (FEL 0.004)		0
18442	Orf1b	nsp14 136	K	K	K/R	R	R	R	Yes	Conserved in SARS-CoV-2	H			0
19048	Orf1b	nsp14 338	A	G/A	G	A	A	A	Yes	Conserved in SARS-CoV-2	H/SWE	OC43 branch (MEME 0.035)		0
20344	Orf1b	nsp15 243	Q	Q	H/Y	H	H	H	Yes	Conserved in SARS-CoV-2	H/SWE			2
20554	Orf1b	nsp15 313	N	N	S/N	S	S	S	No	Conserved in SARS-CoV-2	H	Overall negative selection (FEL 0.04)		0
21400	Orf1b	nsp16 249	A	A	T/S	S	S	S	Yes	Conserved in SARS-CoV-2	H/SWE			2
21614	Orf S	S1 18	F	F/I/L	I	L	F	L	Yes	Variable in SARS-CoV-2	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**	Yes	Conserved in SARS-CoV-2	**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/S**	Yes	Conserved in SARS-CoV-2	**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	Yes	Variable in SARS-CoV-2	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	Yes	Conserved in SARS-CoV-2	H/SWE		PSS, I→V (SARS-CoV-1-like)	0
21920	Orf S	S1 120	V	V	V/I	V	I	V	Yes	Conserved in SARS-CoV-2	H/SWE			0
21926	Orf S	S1 122	T	T	N/T	N	N	N	Yes	Conserved in SARS-CoV-2	H/SWE	Overall negative selection (FEL 0.002)	NSS	0
22004	Orf S	S1 149	N	N/K	K/I	N	G	N	Yes	Conserved in SARS-CoV-2	H		NSS, N→D (new state)	0
22124	Orf S	S1 189	D	T/D/N	D	H	H	N	Yes	Conserved in SARS-CoV-2	H/SWE	OC43 branch (MEME 0.008)	NSS	0
22553	Orf S	S1 332	N	D/N	D/N	N	N	N	Yes	Conserved in SARS-CoV-2	H/SWE			1
23048	Orf S	S1 497	S	A/G/S	D/S	G	G	G	Yes	Variable in SARS-CoV-2	H/SWE	HKU1 branch (MEME 0.044)		2
23948	**Orf S**	**S2 796**	**D**	**N**	**D**	**D**	**Y/D**	**Y/D**	**D/N discrepancy**	Variable in SARS-CoV-2	**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	Yes	Conserved in SARS-CoV-2	H		NSS	1
24620	Orf S	S2 1020	F	F	F/A/L	A	A	A	Yes	Conserved in SARS-CoV-2	H		PSS, A→S/V (new states)	2
24632	Orf S	S2 1024	Q	Q	L/R	L	L	L	Yes	Conserved in SARS-CoV-2	H/SWE			2
24863	Orf S	S2 1101	T	T	H/S	H	S	H	Yes	Conserved in SARS-CoV-2	H/SWE		NSS, H→Y (new state)	1
25037	Orf S	S2 1159	Q	Q	Q/H	H	H	H	Yes	Conserved in SARS-CoV-2	H		NSS, H→Y (new state)	0
25166	Orf S	S2 1202	D	D/Y	D/E	E	E	E	Yes	Conserved in SARS-CoV-2	H		PSS, E→Q/G (new states)	0
25247	Orf S	S2 1230	V	V	V/M	M	M	M	Yes	Conserved in SARS-CoV-2	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 highlighted in the results section
# Sites/branches scored under MEME/FEL and Contrast-FEL (CF); CF tests for differences is selective pressures between clades
§ Representing virus diversity sampled as of May 2021
¶ Representing viral diversity sampled as of December 2022 available from: https://observablehq.com/@spond/sars_cov_2_sites
* 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)

REFERENCES

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nat. Med.* [Internet] 26:450–452. Available from: https://pubmed.ncbi.nlm.nih.gov/32284615/

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.* [Internet] 116:25057–25067. Available from: https://pubmed.ncbi.nlm.nih.gov/31767754/

Banerjee A, Doxey AC, Mossman K, Irving AT. 2021. Unraveling the Zoonotic Origin and Transmission of SARS-CoV-2. *Trends Ecol. Evol.* [Internet] 36:180–184. Available from: https://pubmed.ncbi.nlm.nih.gov/33384197/

Bedford T. 2021. Genetic diversity of betacoronaviruses including novel coronavirus (nCoV) 2020. Available from: https://nextstrain.org/groups/blab/beta-cov

Boni MF, Lemey P, Jiang X, Lam TTY, 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.* [Internet] 5:1408–1417. Available from: https://pubmed.ncbi.nlm.nih.gov/32724171/

C-GU C. 2021. COVID-19 Genomics UK Consortium. Available from: https://www.cogconsortium.uk/

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 Prepr. Serv. Biol.* [Internet]. Available from: https://pubmed.ncbi.nlm.nih.gov/32511325/

Cheng VCC, Lau SKP, Woo PCY, Kwok YY. 2007. Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. *Clin. Microbiol. Rev.* [Internet] 20:660–694. Available from: https://pubmed.ncbi.nlm.nih.gov/17934078/

Corman VM, Muth D, Niemeyer D, Drosten C. 2018. Hosts and Sources of Endemic Human Coronaviruses. *Adv. Virus Res.* [Internet] 100:163–188. Available from: https://pubmed.ncbi.nlm.nih.gov/29551135/

Delport W, Scheffler K, Seoighe C. 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog.* [Internet] 4. Available from: https://pubmed.ncbi.nlm.nih.gov/19096508/

Dolan PT, Whitfield ZJ, Andino R. 2018. Mapping the Evolutionary Potential of RNA Viruses. *Cell Host Microbe* [Internet] 23:435–446. Available from: https://pubmed.ncbi.nlm.nih.gov/29649440/

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.* [Internet] 11. Available from: https://pubmed.ncbi.nlm.nih.gov/33239633/

Ellegren H. 2008. Comparative genomics and the study of evolution by natural selection. *Mol. Ecol.* [Internet] 17:4586–4596. Available from: https://pubmed.ncbi.nlm.nih.gov/19140982/

Escalera-Zamudio M, Golden M, Gutiérrez B, Thézé J, Keown JR, Carrique L, Bowden TA, Pybus OG. 2020. Parallel evolution in the emergence of highly pathogenic avian influenza A viruses. *Nat. Commun.* 11:1–11.

Escalera-Zamudio M, Pond SLK, Viña NM de la, Gutiérrez B, Thézé J, Bowden TA, Pybus OG, Hulswit RJG. 2021. Identification of site-specific evolutionary trajectories shared across human betacoronaviruses. *bioRxiv* [Internet]:2021.05.24.445313. Available from: https://pubmed.ncbi.nlm.nih.gov/34075377/

Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, Crispim MAE, Sales FCS, Hawryluk I, McCrone JT, et al. 2021. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science (80-.).*

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

Forni D, Cagliani R, Arrigoni F, Benvenuti M, Mozzi A, Pozzoli U, Clerici M, de Gioia L, Sironi M. 2021. Adaptation of the endemic coronaviruses HCoV-OC43 and HCoV-229E to the human host. *Virus Evol.* [Internet] 7. Available from: https://academic.oup.com/ve/article/7/2/veab061/6308720

Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 2020. Evidence for host-dependent regular RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* 6:eabb5813.

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

Gutierrez B, Castelán Sánchez HG, Candido D da S, Jackson B, Fleishon S, Houzet R, Ruis C, Delaye L, Faria NR, Rambaut A, et al. 2022. Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America. *Cell Host Microbe* 30:1112-1123.e3.

Gutierrez B, Escalera-Zamudio M, Pybus Oliver G. 2019. Parallel molecular evolution and adaptation in viruses. *Curr. Opin. Virol.* [Internet] 34:90–96. Available from: https://doi.org/10.1016/j.coviro.2018.12.006

Gutierrez B, Escalera-Zamudio M, Pybus Oliver G. 2019. Parallel molecular evolution and adaptation in viruses. *Curr. Opin. Virol.* [Internet] 34:90–96. Available from: https://pubmed.ncbi.nlm.nih.gov/30703578/

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.* [Internet] 117:8094–8103. Available from: https://pubmed.ncbi.nlm.nih.gov/32198201/

He JF, Peng GW, Min J, Yu DW, Liang WJ, Zhang SY, Xu RH, Zheng HY, Wu XW, Xu J, et al. 2004. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* [Internet] 303:1666–1669. Available from: https://pubmed.ncbi.nlm.nih.gov/14752165/

Holmes EC, Rambaut A. 2004. Viral evolution and the emergence of SARS coronavirus. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* [Internet] 359:1059–1065. Available from: https://pubmed.ncbi.nlm.nih.gov/15306390/

Hulswit RJG, De Haan CAM, Bosch B-J. 2016. Coronavirus Spike Protein and Tropism Changes. *Adv. Virus Res.* [Internet] 96:29–57. Available from: http://dx.doi.org/10.1016/bs.aivir.2016.08.004

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.* [Internet] 116:2681–2690. Available from: https://pubmed.ncbi.nlm.nih.gov/30679277/

HyPhy. 2013. PRIME. Available from: http://hyphy.org/w/index.php/PRIME

ICTV, Ziebuhr J, Baric RS, Baker S, de Groot RJ, Drosten C, Gulyaeva A, Haagmans BL, Neuman BW, Perlman S, et al. 2017. Nidovirales. :11.

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* [Internet] 28:1647–1649. Available from: https://pubmed.ncbi.nlm.nih.gov/22543367/
Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, Hosmillo M, Rees-Spear C, Mlcochova P, Lumb IU, et al. 2021. SARS-CoV-2 evolution during treatment of chronic infection. *Nat. 2021 5927853* [Internet] 592:277–282. Available from: https://www.nature.com/articles/s41586-021-03291-y
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.
Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science (80-.).* 368:860–868.
Kistler KE, Bedford T. 2021. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229E. *Elife* 10:1–35.
Kosakovsky Pond SL. Evolutionary annotation of global SARS-CoV-2/COVID-19 genomes enabled by data from GSAID 2021. Available from: https://observablehq.com/@spond/sars_cov_2_sites
Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* [Internet] 22:1208–1222. Available from: https://pubmed.ncbi.nlm.nih.gov/15703242/
Kosakovsky Pond SL, Frost SDW, Grossman Z, Gravenor MB, Richman DD, Leigh Brown AJ. 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput. Biol.* [Internet] 2:0530–0538. Available from: https://pubmed.ncbi.nlm.nih.gov/16789820/
Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* [Internet] 22:3096–3098. Available from: https://pubmed.ncbi.nlm.nih.gov/17110367/
Kosakovsky Pond SL, Wisotsky SR, Escalante A, Magalis BR, Weaver S. 2021. Contrast-FEL-A Test for Differences in Selective Pressures at Individual Sites among Clades and Sets of Branches. *Mol. Biol. Evol.* [Internet] 38:1184–1198. Available from: https://pubmed.ncbi.nlm.nih.gov/33064823/
Krissinel E, Henrick K. 2007. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* [Internet] 372:774–797. Available from: https://pubmed.ncbi.nlm.nih.gov/17681537/
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* [Internet]. Available from: http://dx.doi.org/10.1038/s41586-020-2180-5
Lau SKP, Lee P, Tsang AKL, Yip CCY, Tse H, Lee RA, So L-Y, Lau Y-L, Chan K-H, Woo PCY, 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.* [Internet] 85:11325–11337. Available from: https://pubmed.ncbi.nlm.nih.gov/21849456/
Lau SKP, Woo PCY, Li KSM, Tsang AKL, Fan RYY, Luk HKH. 2015. Discovery of a Novel Coronavirus, China Rattus Coronavirus HKU24, from Norway Rats Supports the Murine Origin of Betacoronavirus 1 and Has Implications for the Ancestor of Betacoronavirus Lineage A. *J. Virol.* 89:3076–3092.
Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* [Internet] 5. Available from: https://pubmed.ncbi.nlm.nih.gov/19779555/
Li F. 2016. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu. Rev. Virol.* 3:237–261.
Li, Li W, Farzan M, Harrison SC. 2005. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* [Internet] 309:1864–1868. Available from: https://pubmed.ncbi.nlm.nih.gov/16166518/
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* [Internet] 310:676–679. Available from: https://pubmed.ncbi.nlm.nih.gov/16195424/
Loewe L, Hill WG. 2010. The population genetics of mutations: good, bad and indifferent. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* [Internet] 365:1153–1167. Available from: https://pubmed.ncbi.nlm.nih.gov/20308090/
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.* [Internet] 112:9436–9441. Available from: https://pubmed.ncbi.nlm.nih.gov/26159422/
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.* [Internet] 19. Available from: https://pubmed.ncbi.nlm.nih.gov/33711012/
De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. 2021. Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biol. Evol.* [Internet] 13. Available from: https://pubmed.ncbi.nlm.nih.gov/33895815/
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.* [Internet] 58:2268–2273. Available from: https://pubmed.ncbi.nlm.nih.gov/4298953/
Menachery VD, Graham RL, Baric RS. 2017. Jumping species—a mechanism for coronavirus persistence and survival. *Curr. Opin. Virol.* [Internet] 23:1–7. Available from: http://dx.doi.org/10.1016/j.coviro.2017.01.002
Millet JK, Whittaker GR. 2015. Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. *Virus Res.* [Internet] 202:120–134. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7189851/pdf/main.pdf
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.* [Internet] 32:1365–1371. Available from: https://pubmed.ncbi.nlm.nih.gov/25701167/
Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* [Internet] 8. Available from: https://pubmed.ncbi.nlm.nih.gov/22807683/
Nelde A, Bilich T, Heitmann JS, Maringer Y, Salih HR, Roerden M, Lübke 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.* [Internet] 22:74–85. Available from: https://pubmed.ncbi.nlm.nih.gov/32999467/
O’Toole Á, Hill V, Pybus OG, Watts A, Bogoch II, Khan K, Messina JP, Tegally H, Lessells RR, Giandhari J, et al. 2021. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. *Wellcome open Res.* [Internet] 6:121. Available from: https://pubmed.ncbi.nlm.nih.gov/34095513/
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* [Internet] 246:96–98. Available from: https://pubmed.ncbi.nlm.nih.gov/4585855/
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.* [Internet] 6. Available from:

https://pubmed.ncbi.nlm.nih.gov/28050020/
Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, Nicholls J, Yee WKS, Yan WW, Cheung MT, et al. 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* 361:1319–1325.
Pollett S, Conte MA, Sanborn M, Jarman RG, Lidl GM, Modjarrad K, Maljkovic Berry I. 2021. A comparative recombination analysis of human coronaviruses and implications for the SARS-CoV-2 pandemic. *Sci. Rep.* [Internet] 11:1–11. Available from: https://doi.org/10.1038/s41598-021-96626-8
Pond SLK, Murrell B, Poon AFY. 2012. Evolution of viral genomes: interplay between selection, recombination, and other forces. *Methods Mol. Biol.* [Internet] 856:239–272. Available from: https://pubmed.ncbi.nlm.nih.gov/22399462/
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.* [Internet] 117:24614–24616. Available from: https://pubmed.ncbi.nlm.nih.gov/32958678/
Rochman ND, Faure G, Wolf YI, Freddolino PL, Zhang F, Koonin E V. 2022. Epistasis at the SARS-CoV-2 Receptor-Binding Domain Interface and the Propitiously Boring Implications for Vaccine Escape. *MBio* 13:e00135-22.
Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* [Internet] 4. Available from: https://pubmed.ncbi.nlm.nih.gov/29340210/
Shaman J, Galanti M. 2020. Will SARS-CoV-2 become endemic? *Science* [Internet] 370:527–529. Available from: https://pubmed.ncbi.nlm.nih.gov/33055131/
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* [Internet]. Available from: http://dx.doi.org/10.1038/s41586-020-2179-y
Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* [Internet] 23:7–9. Available from: https://pubmed.ncbi.nlm.nih.gov/16177232/
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* [Internet] 5. Available from: https://pubmed.ncbi.nlm.nih.gov/32581081/
Stamatakis A. 2015. Using RAxML to Infer Phylogenies. *Curr. Protoc. Bioinforma.* [Internet] 51:6.14.1-6.14.14. Available from: https://pubmed.ncbi.nlm.nih.gov/26334924/
Starr TN, Greaney AJ, Hannon WW, Loes AN, Hauser K, Dillen JR, Ferri E, Farrell AG, Dadonaite B, McCallum M, et al. 2022. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science (80-.).* [Internet] 377:420–424. Available from: https://www.science.org/doi/10.1126/science.abo7896
Stern A, Yeh M Te, 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* [Internet] 169:35-46.e19. Available from: https://pubmed.ncbi.nlm.nih.gov/28340348/
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.* [Internet] 24:490–502. Available from: http://dx.doi.org/10.1016/j.tim.2016.03.003
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.* [Internet] 4. Available from: https://pubmed.ncbi.nlm.nih.gov/29942656/
Taefehshokr N, Taefehshokr S, Hemmat N, Heit B. 2020. Covid-19: Perspectives on Innate Immune Evasion. *Front. Immunol.* [Internet] 11. Available from: https://pubmed.ncbi.nlm.nih.gov/33101306/
Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N, et al. 2021. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* [Internet] 592:438–443. Available from: https://pubmed.ncbi.nlm.nih.gov/33690265/
Tortorici MA, Walls AC, Lang Y, Wang C, Li Z, Koerhuis D, Boons G, Bosch B, Rey FA, Groot RJ De, et al. 2019. Structural basis for human coronavirus attachment to sialic acid receptors. *Nat. Struct. Mol. Biol.* 26:481–489.
Turakhia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, Ye C, Smith K, De Maio N, Haussler D, Lanfear R, et al. 2022. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nat. 2022 6097929* [Internet] 609:994–997. Available from: https://www.nature.com/articles/s41586-022-05189-9
V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. 2020. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat. Rev. Microbiol. 2020 193* [Internet] 19:155–170. Available from: https://www.nature.com/articles/s41579-020-00468-6
Vijaykrishna D, Smith GJD, Zhang JX, Peiris JSM, Chen H, Guan Y. 2007. Evolutionary Insights into the Ecology of Coronaviruses. *J. Virol.* 81:4012–4020.
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. *J. Virol.* 79:1595–1604.
ViPR-NCBI. 2021. Virus Pathogen Resource. Available from: https://www.viprbrc.org/brc/home.spg?decorator=vipr
Wang G, Dunbrack RL. 2004. Scoring profile-to-profile sequence alignments. *Protein Sci.* [Internet] 13:1612–1626. Available from: https://pubmed.ncbi.nlm.nih.gov/15152092/
Wang H, Pipes L, Nielsen R. 2020. Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. *Virus Evol.* [Internet] 7. Available from: https://pubmed.ncbi.nlm.nih.gov/33500788/
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.* [Internet] 89:8416–8427. Available from: https://pubmed.ncbi.nlm.nih.gov/26041293/
Watanabe Y, Bowden TA, Wilson IA, Crispin M. 2019. Exploitation of glycosylation in enveloped virus pathobiology. *Biochim. Biophys. Acta - Gen. Subj.* [Internet]:1–18. Available from: https://doi.org/10.1016/j.bbagen.2019.05.012
WHO. 2021. Middle East respiratory syndrome 2021. Available from: http://www.emro.who.int/health-topics/mers-cov/mers-outbreaks.html
de Wilde AH, Snijder EJ, Kikkert M, van Hemert MJ. 2018. Host Factors in Coronavirus Replication. *Curr. Top. Microbiol. Immunol.* [Internet] 419:1–42. Available from: https://pubmed.ncbi.nlm.nih.gov/28643204/
Woo PCY, Huang Y, Lau SKP, Yuen KY. 2010. Coronavirus genomics and bioinformatics analysis. *Viruses* [Internet] 2:1805–1820. Available from: https://pubmed.ncbi.nlm.nih.gov/21994708/
Woo PCY, Lau SKP, Chu C, Chan K, Tsoi H, Huang Y, Wong BHL, Poon RWS, Cai JJ, Luk W, et al. 2005. Characterization and Complete Genome Sequence of a Novel Coronavirus, Coronavirus HKU1, from Patients with Pneumonia. *J. Virol.* 79:884–895.
Woo PCY, Lau SKP, Yip CCY, Huang Y, Tsoi H-W, Chan K-H, Yuen K-Y. 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.
Yewdell JW. 2021. Antigenic drift: Understanding COVID-19. *Immunity* [Internet] 54:2681–2687. Available from: https://doi.org/10.1016/j.immuni.2021.11.016
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.* [Internet] 39:198–216. Available from: https://pubmed.ncbi.nlm.nih.gov/32447571/
Yuen CK, Lam JY, Wong WM, Mak LF, Wang X, Chu H, Cai JP, Jin DY, To KKW, Chan JFW, et al. 2020. SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists. *Emerg. Microbes Infect.* [Internet] 9:1418–1428. Available from: https://pubmed.ncbi.nlm.nih.gov/32529952/
Zahradník J, Marciano S, Shemesh M, Zoler E, Harari D, Chiaravalli J, Meyer B, Rudich Y, Li C, Marton I, et al. 2021. SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nat. Microbiol. 2021 69* [Internet] 6:1188–1198. Available from: https://www.nature.com/articles/s41564-021-00954-4
Zhou P, Yang X Lou, 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* [Internet] 579:270–273. Available from: http://dx.doi.org/10.1038/s41586-020-2012-7
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.* [Internet] 7. Available from: https://pubmed.ncbi.nlm.nih.gov/30377292/

Figure 1

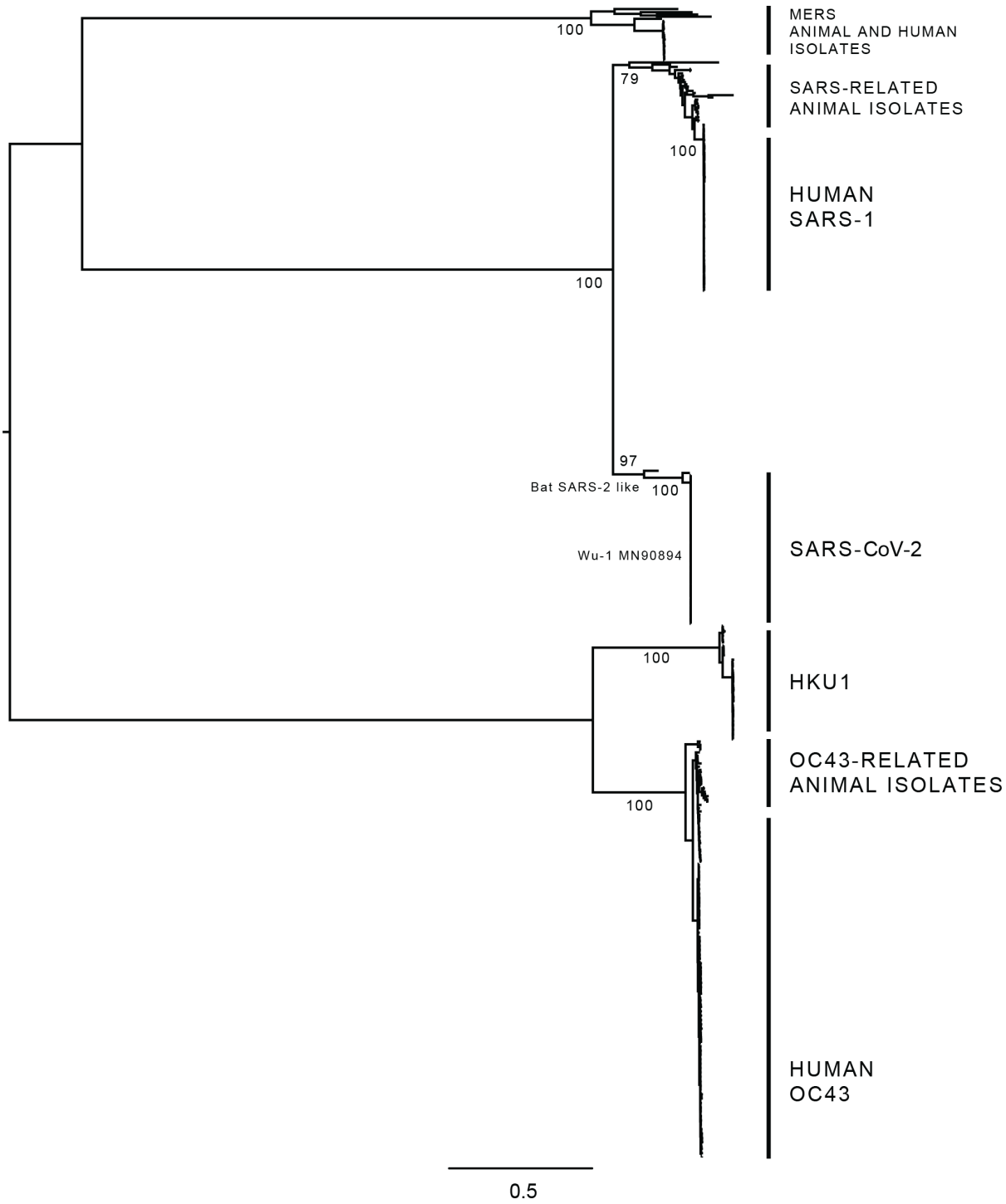

MERS
ANIMAL AND HUMAN
ISOLATES
SARS-RELATED
ANIMAL ISOLATES
HUMAN
SARS-1
SARS-CoV-2
HKU1
OC43-RELATED
ANIMAL ISOLATES
HUMAN
OC43
100
79
100
100
97
Bat SARS-2 like
100
Wu-1 MN90894
100
100
0.5

Figure 2

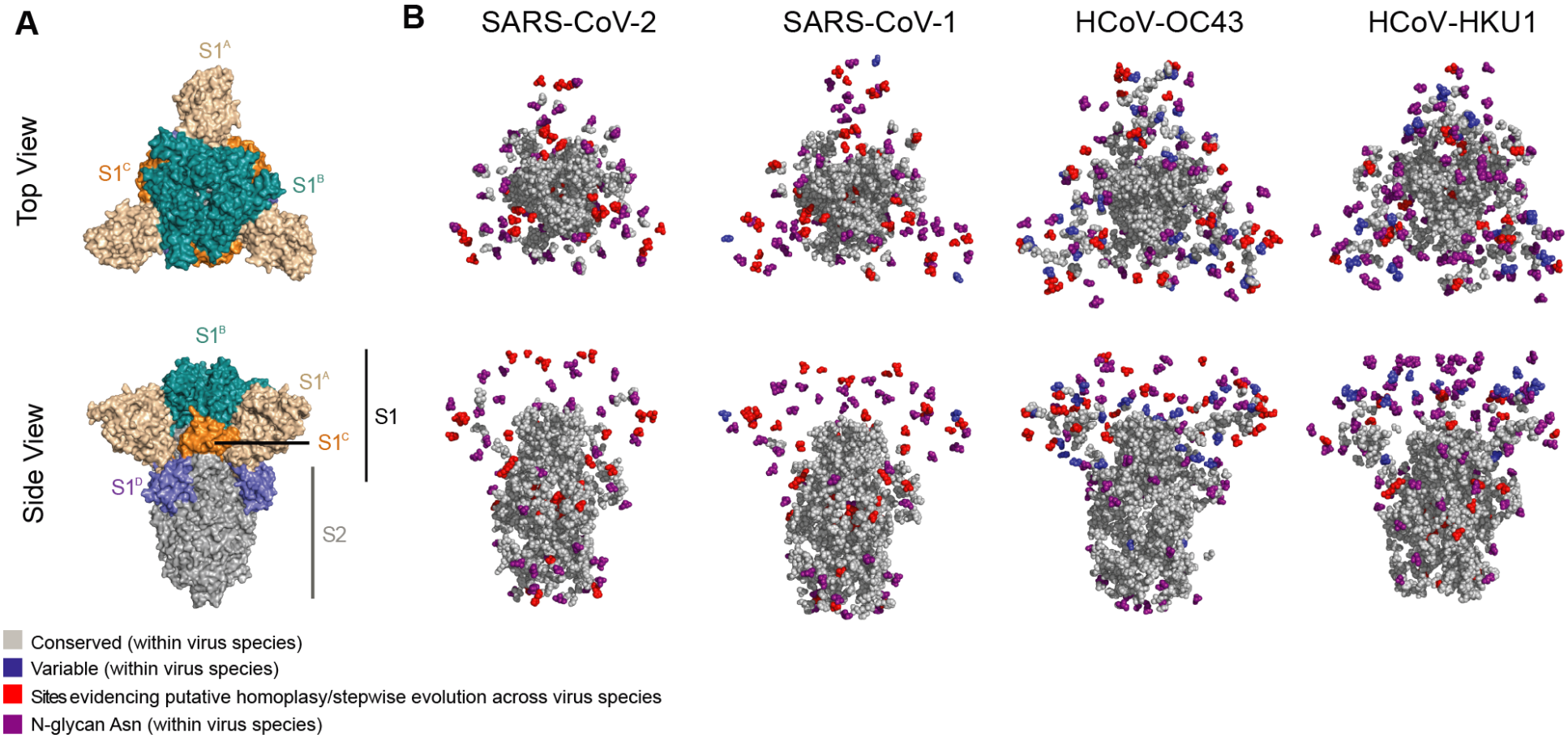

A
B
SARS-CoV-2
SARS-CoV-1
HCoV-OC43
HCoV-HKU1
Top View
Side View
S1A
S1B
S1C
S1D
S1
S2
Conserved (within virus species)
Variable (within virus species)
Sites evidencing putative homoplasy/stepwise evolution across virus species
N-glycan Asn (within virus species)

Figure 3

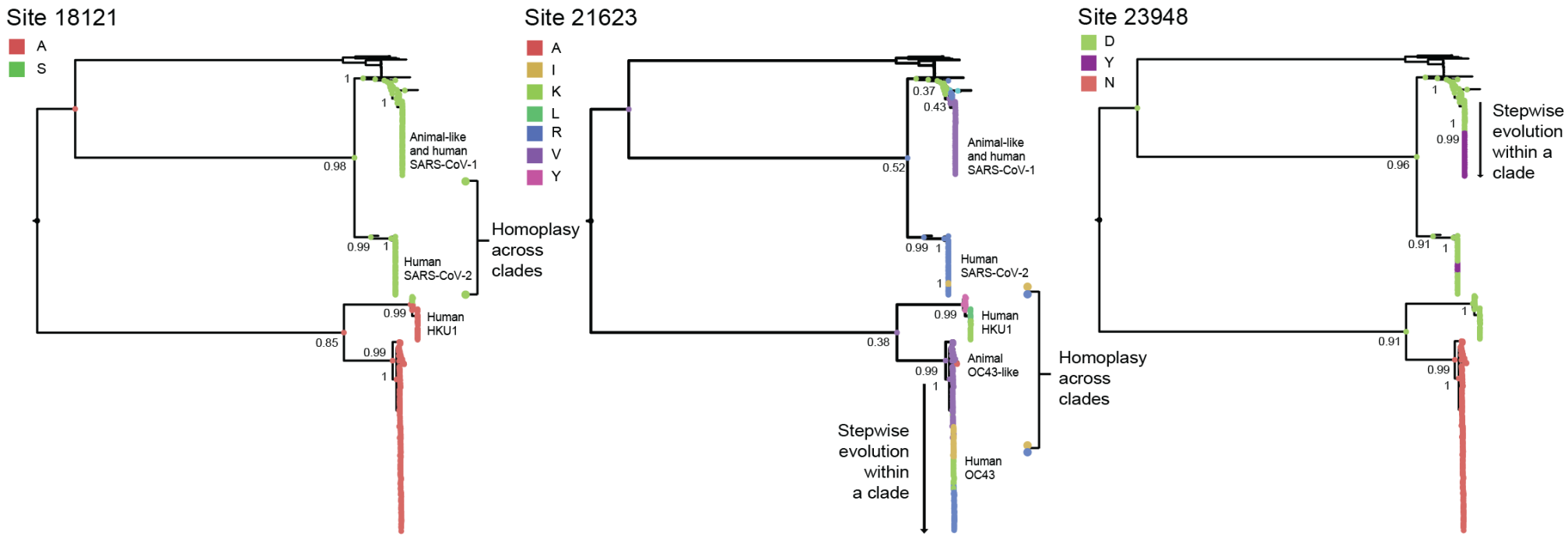

Site 18121
A
S
1
1
0.98
Animal-like and human SARS-CoV-1
0.99
1
Human SARS-CoV-2
Homoplasy across clades
0.99
Human HKU1
0.85
0.99
1
Site 21623
A
I
K
L
R
V
Y
0.37
0.43
0.52
Animal-like and human SARS-CoV-1
0.99
1
Human SARS-CoV-2
1
0.99
Human HKU1
0.38
0.99
1
Animal OC43-like
Homoplasy across clades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

Figure 4

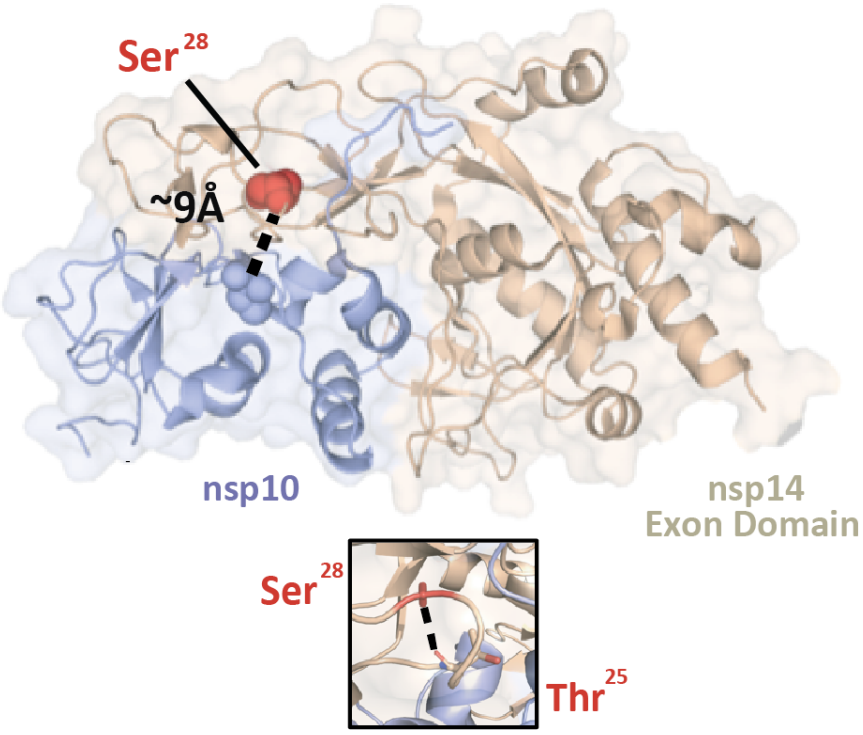
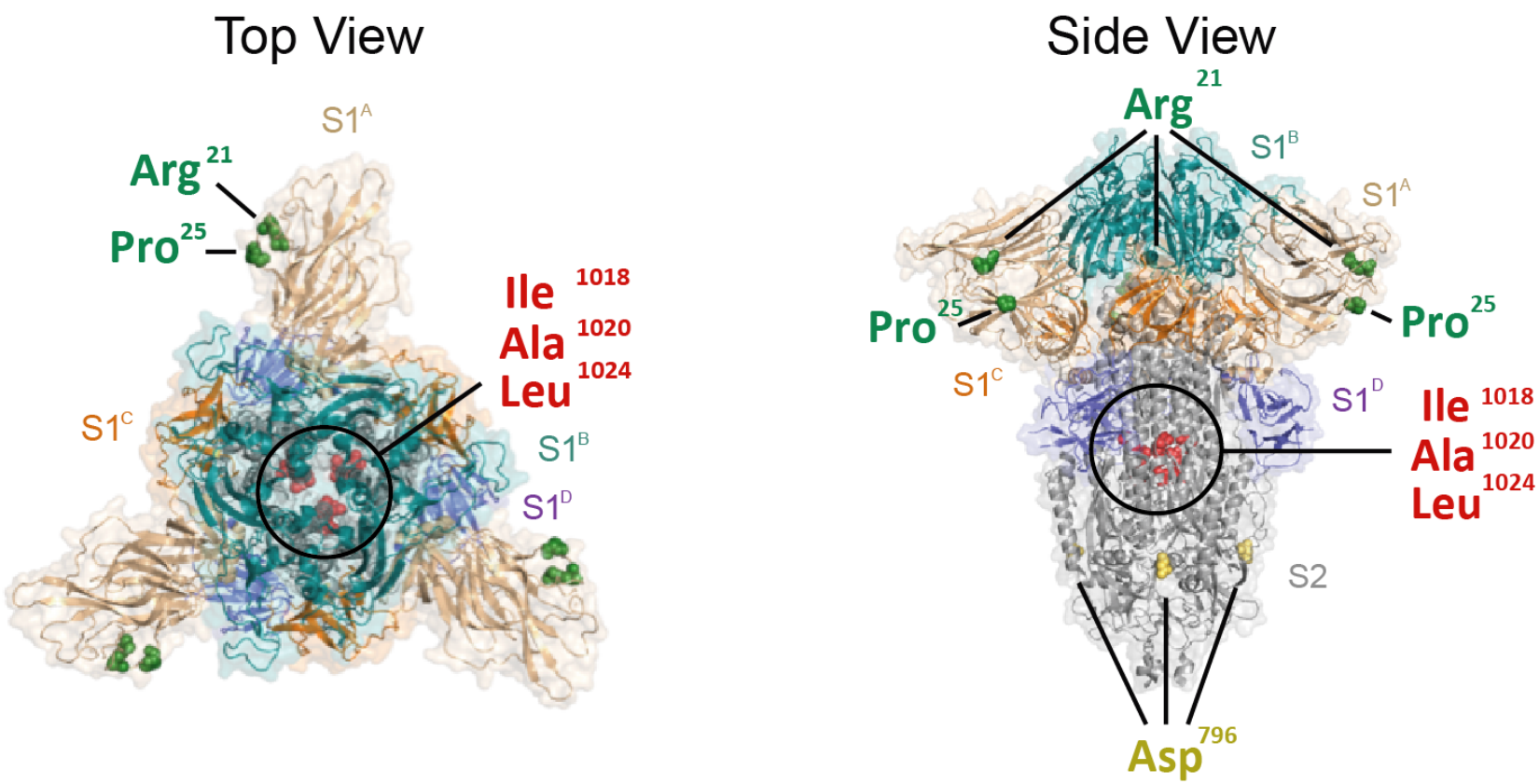


Figure 5


Top View
S1A
Arg21
Pro25
Ile 1018
Ala 1020
Leu 1024
S1C
S1B
S1D
Side View
Arg21
S1B
S1A
Pro25
Pro25
S1C
S1D
Ile 1018
Ala 1020
Leu 1024
S2
Asp796