The emergence of highly fit SARS-CoV-2 variants accelerated by recombination

Michael R. Garvin[1,2+*], Erica T. Prates[1,2+], Jonathon Romero[3], Ashley Cliff [3], Joao Gabriel Felipe Machado Gazolla[1,2], Monica Pickholz[4,5], Mirko Pavicic[1,2], Daniel Jacobson[1,2,*]

Affiliations:

[1]Oak Ridge National Laboratory, Computational Systems Biology, Biosciences, Oak Ridge, TN; [2]National Virtual Biotechnology Laboratory, US Department of Energy; [3]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee Knoxville, Knoxville, TN; [4]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina; [5]Instituto de Física de Buenos Aires (IFIBA), CONICET-Universidad de Buenos Aires, Buenos Aires, Argentina.

***Correspondence:** garvinmr@ornl.gov, jacobsonda@ornl.gov

[+]**Contributed equally**

Key words: SARS-CoV-2; recombination; haplotype; Delta variant; variants of concern; COVID-19, syncytia

Running title: Highly fit SARS-CoV-2 generated from recombination

Abstract

The SARS-CoV-2 pandemic has entered an alarming new phase with the emergence of the variants of concern (VOC), P.1, B.1.351, and B.1.1.7, in late 2020, and B.1.427, B.1.429, and B.1.617, in 2021. Substitutions in the spike glycoprotein (S), such as $Asn^{501}Tyr$ and $Glu^{484}Lys$, are likely key in several VOC. However, $Asn^{501}Tyr$ had been circulating for months in earlier strains and $Glu^{484}Lys$ is not found in B.1.1.7, indicating that they do not fully explain those fast-spreading variants. Here we use a computational systems biology approach to process more than 900,000 SARS-CoV-2 genomes and map spatiotemporal relationships, revealing other critical attributes of these variants. Comparisons to earlier dominant mutations and protein structural analyses indicate that the increased transmission is promoted by the combination of functionally complementary mutations in S and in other regions of the SARS-CoV-2 proteome. We report that the VOC have in common mutations in non-S proteins involved in immune-antagonism and replication performance, such as the nonstructural proteins 6 and 13, suggesting a convergent evolution of the virus. Critically, we propose that recombination events among divergent coinfecting haplotypes greatly accelerates the emergence of VOC by bringing together cooperative mutations and explaining the remarkably high mutation load of B.1.1.7. Therefore, extensive community distribution of SARS-CoV-2 increases the probability of future recombination events, further accelerating the evolution of the virus. This study reinforces the need for a global response to stop COVID-19 and future pandemics.

Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

"Nothing in Biology Makes Sense Except in the Light of Evolution" -Theodosius Dobzhansky

1. Introduction

In late 2020, three SARS-CoV-2 variants of concern, VOC; B.1.1.7, B.1.351, and P.1 (also called Alpha, Beta, and Gamma respectively) rapidly became the predominant source of infections due to enhanced transmission rates and have since been linked to increased hospitalizations and mortalities (Alpert et al. 2021; Challen et al. 2021; Faria et al. 2021; Funk et al. 2021; Sabino et al. 2021; Washington et al. 2021; Davies et al.; Volz et al.). In early 2021, several new VOC appeared including, B.1.427 (Epsilon), B.1.526 (Iota), and B.1.617 (Delta). B.1.617 is of immediate concern because it is responsible for the COVID-19 crisis that recently began in India (Singh et al. 2021), is causing the majority of new infections in the United Kingdom (UK), and the United States (USA), and has now been observed in more than 70 countries worldwide. Notably, several of these VOC have rapidly spread even in regions such as the UK that depend on robust sampling efforts for early detection. There is therefore a critical need to identify accurate predictors and biological causes for the increased transmission of the next VOC, which will inevitably emerge if the viral spread is not globally restrained.

Although extensive efforts are underway to achieve these ends, integrating new findings is critical to unravel the multiple biomolecular and environmental factors influencing viral evolution. Toward a holistic understanding of VOC emergence, two major weaknesses need to be addressed: (1) currently, the mutations used to identify VOC and potentially explain the altered biology of the virus are predominantly focused on the changes observed in the spike glycoprotein (S) whereas those in other genomic regions are largely ignored, and (2) the molecular models used to reconstruct the evolutionary history of the virus employ phylogenetic trees that are useful for species-level but not population-based analyses, which is the case with SARS-CoV-2 (Huson and Bryant 2006; Velasco 2013).

72 For example, the Asn^{501}Tyr substitution in S is likely key because it increases affinity for
73 the host receptor, angiotensin-converting enzyme 2 (ACE2) (Liu et al. 2021), and is often used to
74 identify the late 2020 VOC (Fratev; Luan et al.), but this mutation has been circulating widely at
75 low frequency and only expanded seven months after being first detected. Similarly, the
76 Glu^{484}Lys substitution in S is often discussed in the context of P.1 and B.1.351 VOC and may
77 allow escape from neutralizing antibodies (Starr et al. 2021; Greaney et al.), but is not found in
78 B.1.1.7 and therefore does not explain the increased transmission of all three late 2020 VOC.
79 These characteristics suggest that several mutations including those in S are being transmitted as
80 a linked set, i.e., a haplotype and their combined effects (i.e., epistasis) may be contributing to
81 the rapid viral spread.

82 Widely used molecular evolutionary models based on phylogenetic trees are also
83 problematic because the *algorithms* that are applied assume that mutations appearing in different
84 SARS-CoV-2 haplotypes are due to repeated, independent mutations and the *scientific*
85 *community* is interpreting this as evidence for the same; i.e., the logic is circular. Alternatively,
86 these apparent repeated mutations may represent recombination, which is a common mechanism
87 to accelerate evolution compared to single site mutations in positive strand RNA viruses such as
88 SARS-CoV-2 (Bentley and Evans 2018). Furthermore, phylogenetic trees are unable to
89 incorporate important molecular events and metadata such as geospatial and temporal data that
90 would be highly informative for detecting current and future VOC.

91 In contrast, median-joining networks (MJN) are an efficient and accurate means to
92 analyze haploid genomic data at the population level (Bandelt et al. 1999) such as SARS-CoV-2,
93 (Garvin, Prates, et al. 2020). Unlike independently segregating sites represented in phylogenetic
94 trees, the unit of interest in an MJN is the haplotype, which more accurately reflects the biology

95 of coronaviruses and enables the detection of important evolutionary events such as
96 recombination. Furthermore, a network can be annotated with information including frequency,
97 geospatial location, demographic, or clinical outcomes associated with a unique haplotype to
98 create interpretable patterns of genome variation.

99 Here, we processed more than 900,000 SARS-CoV-2 genomes using a computational
100 workflow that combines MJN and protein structural analysis (Garvin, Prates, et al. 2020; Prates
101 et al. 2020) to identify critical attributes of these VOC and provide substantial evidence that the
102 genome-wide mutation load of the late 2020 VOC results from recombination between divergent
103 strains. Via focused structural analysis and molecular dynamics simulations, we explore the
104 individual effects of key mutations in S and other proteins of SARS-CoV-2 that are shared
105 among different VOC. We propose that linked mutations in VOC act cooperatively to enhance
106 viral spread and our results emphasize the major role of community spread in generating future
107 VOC (Sheikh et al. 2021).

108

109 **2. Results and Discussion**

110 ***Molecular evolution of populations is best represented by a network***

111 The COVID-19 pandemic is both an unprecedented tragedy and an opportunity to study
112 molecular evolution given the abundant and global sampling of the mutational space of SARS-
113 CoV-2. The MJN is a valuable method of integrating these data to understand viral evolution
114 because the model assumes single mutational steps in which each node represents a haplotype
115 and the edge between nodes is a mutation leading to a new one. Typically, a subsample of extant

116 haplotypes for a taxon is obtained and unsampled, or extinct lineages are inferred. In contrast,
117 SARS-CoV-2 sequence data repositories provide extensive sampling of haplotypes and
118 collection dates (the calendar date of the sample). Given that in an MJN, the temporal
119 distribution of haplotypes is inherent (the model assumes time-ordered sets of mutations), the
120 mutational history of the virus can be traced as a genealogy that can incorporate both the relative
121 *and* absolute times of emergence of SARS-CoV-2 variants. Importantly, when the single-
122 mutational step MJN model fails, it produces features such as loops or clusters of inferred
123 haplotypes that can indicate biologically important processes such as recombination events, back
124 mutations, or repeat mutations at a site that may be under positive selection.

125 To make a direct comparison, we generated a network and a phylogenetic tree of SARS-
126 CoV-2 haplotypes that were identified from sequences sampled during the first four months of
127 the pandemic and deposited into GISAID (A Global Initiative on Sharing Avian Flu Data,
128 gisaid.org) (Figure 1). Clearly, important metadata such as haplotype frequency, date of
129 emergence, and mutations of interest are easily displayed on the network but are not on the
130 phylogenetic tree. Likewise, at day 96, reticulations (i.e., homoplasy loops) begin to appear in
131 the MJN, indicating reverse mutations to the ancestral states at specific sites or possibly
132 recombination events that can be explored further. Another important feature identified when
133 using networks, but is lost when using phylogenetic trees, is the presence of polytomies. So-
134 called soft polytomies often indicate unsampled genomic information at the species level and
135 hard polytomies are molecular events often found in rapidly expanding populations. For
136 example, haplotype H04 in the MJN (Figure 1) represents a hard polytomy and indicates that a
137 frequent variant is further undergoing multiple independent mutational events, but the
138 phylogenetic tree is unable to convey this information.

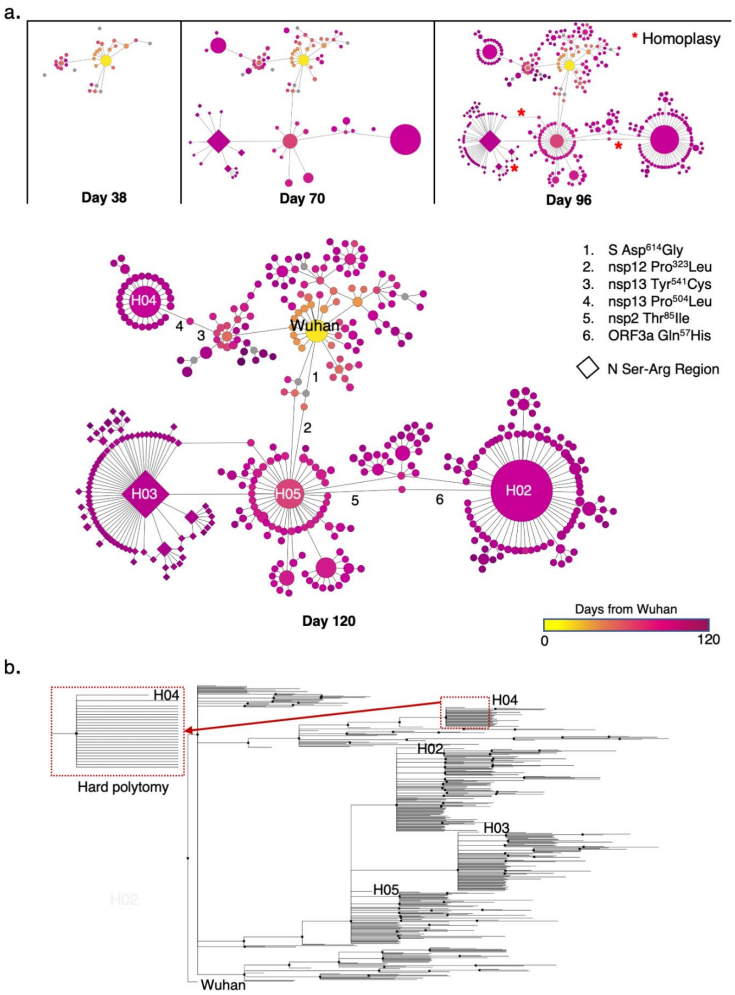


Fig. 1. Comparison of a median-joining network (MJN) and phylogenetic tree generated with SARS-CoV-2 sequences sampled through April 2020. **a**. MJN of SARS-CoV-2 haplotypes, 96, and 120 days. Node sizes in the MJN correspond to sample sizes for a given haplotype and node colors indicate the time of its first report relative to the putative origin of the pandemic in Wuhan. The most abundant haplotypes are named H02 - H05 and numerals 1 - 6 identify important mutations (Garvin, Prates, et al. 2020). Diamond shape nodes denote haplotypes that harbor a 3-bp mutation in the nucleocapsid gene (N) that is highly conserved and directly affects viral replication *in vitro (Tylor et al. 2009; Thorne et al. 2021)*. **b.** The phylogenetic tree is unable to convey the same information. For example, rapidly expanding populations often display polytomies, i.e., single mutations from a common central haplotype. Those events are readily identified on the MJN but difficult to interpret on a tree because they are usually visualized as a multi-pronged fork (outlined in the dashed-line box) rather than a star pattern (compare H04 in (a) and (b)). These true biological processes also cause tree algorithms to perform poorly because they violate their

151 assumptions, slowing convergence. Additionally, MJN are able to indicate reticulations (i.e., loops) that could
152 denote recombination, reverse mutations, or other biologically important events whereas the forced bifurcation of
153 phylogenetic tree algorithms is unable to display these. Reference sequence: NC_045512, Wuhan, December 24,
154 2019.

155

156 ***Lineage-defining mutations of Variants of Concern***

157 We processed more than 900,000 SARS-CoV-2 genomes from human and mink, built a MJN
158 network using the 640,211 genomes that survived our quality control workflow, and annotated
159 with PANGO lineages defined in the GISAID database (Figure 2, *see* Methods). This
160 genealogy-based approach to molecular evolution identifies the mutations that define a given
161 haplotype based on the edge between nodes. Here, they define the variants of concern/interest
162 based on the edge that initiates their corresponding clusters of nodes (Table 1).

163 This approach also enables the identification of all the common features acquired in
164 different VOC, which can elucidate the set of molecular features underlying their rapid spread.
165 For example, the B.1.1.7, B.1.351, and P.1 variants, here referred to as late_2020_VOC, can all
166 be defined by a triple amino acid deletion in the nonstructural protein 6 (nsp6; Ser^{106}_Phe^{108}del)
167 as well as Asn^{501}Tyr in S. Notably this latter mutation has received considerable attention
168 compared to the former (Figure 2a, Table 1), but both are likely key to the biology of these VOC.
169 The MJN also reveals what appears to be a previous dominant but now rare variant (D.2) in
170 Australia; an Ile^{120}Phe mutation in nsp2 was followed immediately by S Ser^{477}Asn, which seems
171 to have led to its rapid expansion in April 2020, indicated by increased node size. This
172 underscores the usefulness of the MJN approach because it is able to convey the sequence of

173 timing of mutational events and the number of individuals carrying those haplotypes
174 simultaneously.

175 In order to account for sampling bias (there are a disproportionate number of sequences
176 contributed to GISAID by the UK and Australia as well as the high frequency of B.1.1.7 and D.2
177 in those two geographic locations), we plotted the number of daily samples of selected variants
178 relative to their respective total number of cases to date (April, 2021) and compared the resulting
179 slopes of the linear range of the curves (Figure 2c). The late_2020_VOC, early_2021_VOC, and
180 early_2021_VOI (Table 1) display higher daily accumulation rates, between 1% and 3% of total
181 observed cases per day, compared to other variants (e.g., B.1.177), which show less than 1%
182 accumulation per day. Notably, the rapid increase in D.2 (2%) supports the MJN view of this as
183 a likely VOC. This analysis and the MJN confirm the importance of monitoring these variants
184 closely and identify both S and non-S mutations that define the current and potential SARS-
185 CoV-2 VOC (Table 1).

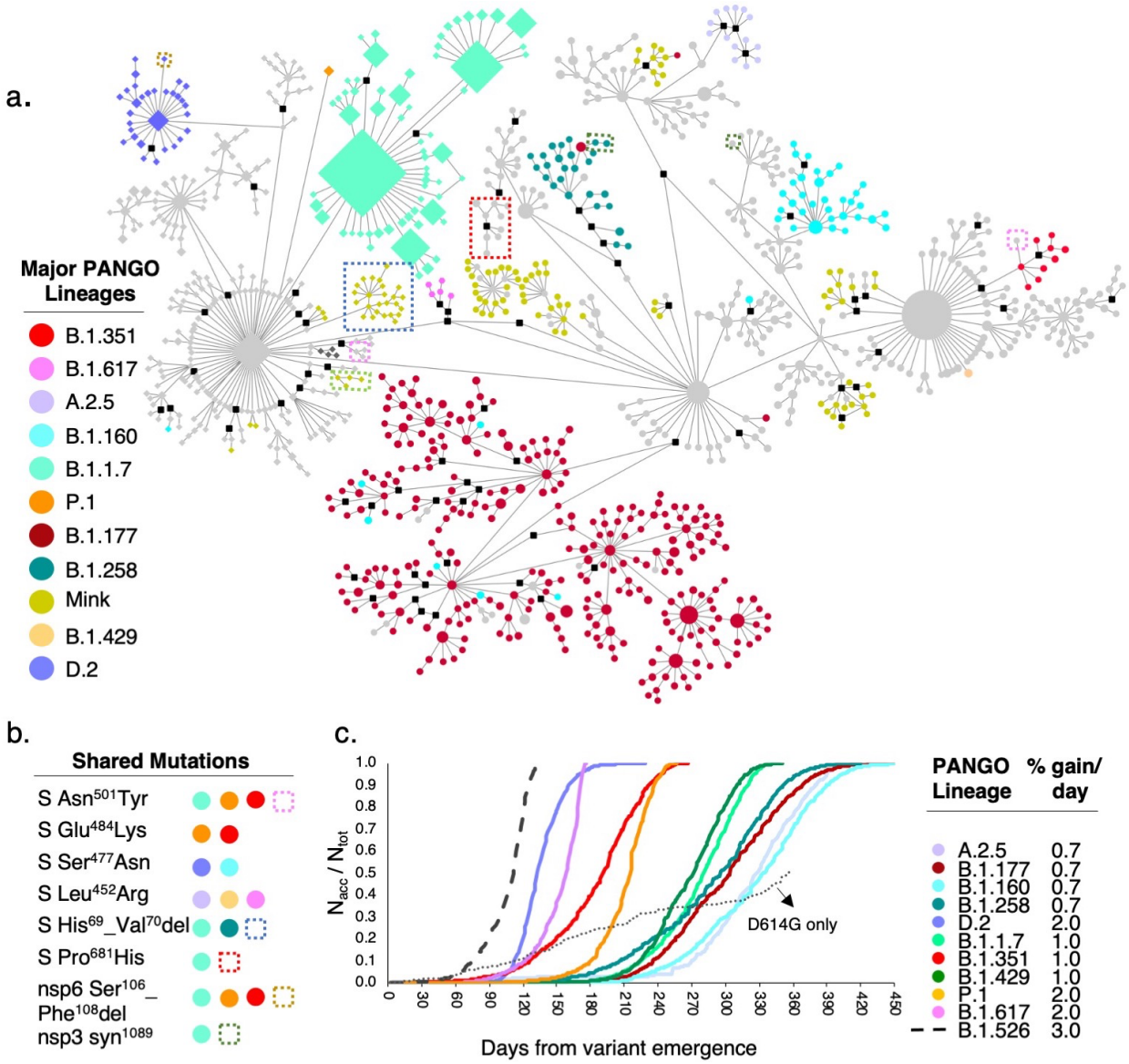


Fig. 2. Median-joining network (MJN) of SARS-CoV-2 genomes. a. MJN of haplotypes found in more than 30 individuals (N=640,211 sequences) using 2,128 variable sites. Colors identify PANGO lineages from GISAID. Diamond-shaped nodes correspond to haplotypes carrying a three base pair deletion in the nucleocapsid gene (N) at sites 28881-28883 (Arg^{203}Lys, Gly^{203}Arg). Black square nodes are inferred haplotypes, dashed-line box defines a subgroup of haplotypes within a lineage with a disjoint mutation that is also found in B.1.1.7. Several lineages show introgression from others (e.g., cyan nodes, B.1.160, into brick red, B.1.177). **b.** Several important mutations in S and non-S proteins appear in multiple lineages. For example, the B.1.1.7 variant carries four mutations that are in disjoint nodes: S Asn^{501}Tyr, S Pro^{681}His, a silent mutation in the codon for amino acid 1089 in nsp3, and the S

His^{69}_Val^{70}del that is also found in a clade of haplotypes from mink (blue dashed-line box in (a)). **c.** Accumulation rate for common GISAID lineages including VOC represented by the ratio between the accumulated number of reported sequences of a given lineage per day since the appearance of that haplotype (N_{acc}) divided by the corresponding total number (N_{tot}) at the final sample date for this study. Colors of curves correspond to node colors in (a). All VOC display accumulation rates of at least 1% of the total for that variant per day. The remaining are less than 1% except for the VOI B.1.526 (not displayed in MJN) that is the highest with 3% per day, indicating further scrutiny of this variant is warranted. We also plotted the accumulation rate for lineages that carry the widely reported S Asp^{614}Gly mutation but without the nsp12 Pro^{323}Leu commonly found with it, supporting our previous hypothesis (Garvin, Prates, et al. 2020) that mutations in S alone are not responsible for the rapid transmission of these VOC/VOI but is a function of epistasis among S and non-S mutations. Reference sequence: NC_045512, Wuhan, December 24, 2019.

Table 1. Major lineages shown in the median-joining network and their defining mutations. Center for disease control (CDC)-defined variants and their timing are listed under *Lineages* and discussed in the text. L-VOC denotes likely variants of concern, that is, those that we propose to have strong potential to become VOC. Non-VOC (N-VOC) are not identified by CDC as VOC. Potential epistatic non-S mutations lineage-defining mutations are listed for the VOC, L-VOC, and VOI. Sites in red font are discussed in the text.

Lineage	Class	Spike Mutation(s)	Likely non-S Epistatic Partner(s)	First major detection
B.1	Early_2020_VOC	D614G	nsp12 P323L	Germany
B.1.1.7 (alpha)	Late_2020_VOC	N501Y, del 69-70, P681H*, T716I,D1118H	nsp6 del 106-108, N L3D, N S235Y	United Kingdom
B.1.351 (beta)	Late_2020_VOC	N501Y, E484K, K417N	nsp6 del 106-108	South Africa
P.1 (gamma)	Late_2020_VOC	N501Y, E484K, K417N	nsp6 del 106-108	Brazil
B.1.427 (epsilon)	Early_2021_VOC	L452R, S13I	nsp13 D260Y	United States, California
B.1.429	Early_2021_VOC	L452R, W152C	nsp13 D260Y	United States, Washington
B.1.617 (delta)	Early_2021_VOC	L452R, E484Q, P681R*	N R203M, ORF7a V82G, ORF3a S26L	India
A.2.5	L-VOC	L452R, del 142-145	nsp1 L4P, nsp3 K839E, nsp4 P308Y	Panama
D.2	L-VOC	S477N	nsp2 I120F	Australia
B.1.160	N-VOC	S477N	na	Denmark
B.1.177	N-VOC	A222V	na	United Kingdom/Denmark
B.1.258	N-VOC	N434K, del 69-70	na	Denmark
B.1.526 (iota)	Early_2021_VOI	L5P, T95I, D253G	nsp6 del 106-108, nsp4 L438P, nsp13 Q88H	United States, New York

* multibasic furin cleavage site

215 ***Recombination is the likely source for the rapidly expanding variants***

216 Haploid, clonally replicating organisms such as SARS-CoV-2 are predicted to eventually
217 become extinct due to the accumulation of numerous slightly deleterious mutations over time,
218 i.e., Muller's ratchet (Muller 1964). Recombination is not only a rescue from Muller's ratchet, it
219 can also accelerate evolution by allowing for the union of advantageous mutations from
220 divergent haplotypes (Bentley and Evans 2018). In SARS-CoV-2, recombination manifests as a
221 template switch during replication when more than one haplotype is present in the host cell, i.e.
222 the virus replisome stops processing a first RNA strand and switches to a second one from a
223 different haplotype, producing a hybrid virus (Simon-Loriere and Holmes 2011). In fact,
224 template switching is a necessary step during the negative-strand synthesis of SARS-CoV-2
225 when the replisomes functions as an RNA-dependent RNA polymerase and pauses at
226 transcription-regulatory motifs of the sub-genomic template to add the leader sequence from the
227 5' end of the genome (this "recombination" is not detected if only a single strain is present, i.e.
228 there is no variation) (Kim et al. 2020),). Given this and the fact that recombination is a major
229 mechanism of coronavirus evolution (Boni et al. 2020) it would be improbable for this process
230 *not* to occur in the case of multiple strains infecting a cell (Gribble et al. 2021).

231 The late_2020_VOC exhibits large numbers of new mutations relative to any closely
232 related sequence indicating rapid evolution of SARS-CoV-2 (Figure 3a). For example, the
233 original node of B.1.1.7 differs from the most closely related node by 28 mutations. However,
234 the majority of this total (15) corresponds to deletions that could be considered two single
235 mutational events, as does a 3-bp change in N (28280-22883) since they occur in factors of three
236 (a codon), maintaining the coding frame. By summing the two deletions and the full codon
237 change 3-bp change in N with the 10 remaining single site mutations, a conservative estimate

238 would be 13 distinct mutational events leading to B.1.1.7. The plot of the accumulating
239 mutations in the 640,211 haplotypes sampled to date reveals a linear growth of roughly 0.05
240 mutations per day (Figure 3b) and therefore, given this pace, it would be expected to take 260
241 days for these 13 mutational events to accumulate in a haplotype. For the B.1.1.7 to appear in
242 October 2020 as reported, the genealogy would have to have been initiated in January 2020 and
243 yet the nearest node harboring the S $Asn^{501}Tyr$ mutation was not sampled until June 2020 and no
244 intermediate haplotypes have been identified to date.

245 Alternatively, it could be that the 13 mutational events occurred between June and
246 October (122 days), but the probability of this is about one in 10^{15} (Supplemental Methods).
247 Furthermore, all 28 differences between the Wuhan reference sequence and B.1.1.7 appeared in
248 earlier haplotypes (Table S1) and therefore, if rapid evolution were the cause, it would require
249 the extremely unlikely process of 28 independent, repeat mutations to the same nucleotide state.
250 In order to test this, we plotted the population-level mutations per day (including repeat
251 mutations at variable sites), which did not reveal any increase in mutation rate at the time of the
252 B.1.1.7 and in fact displayed a *decrease* with the emergence of the late_2020_VOC (Figure 3c,
253 Figure S1). Possible explanations are either a large increase in mutations in a small number of
254 individuals over a short time period (that would have to occur on multiple continents to explain
255 B.1.1.7, B.1.351, and P.1), or recombination between two or more divergent haplotypes carrying
256 the VOC mutations (Figure 3d).

257 Recombination is the most parsimonious explanation given (1) the absence of a
258 substantial increase in mutation rate at any time prior to the appearance of the VOC, (2) the
259 widespread and early circulation of the majority of the mutations associated with them in other
260 haplotypes and, (3) that several mutations appear disjointly across the MJN (Figure 2a). The

261 first notable disjoint mutation is Ser^{477}Asn in S that defines D.2 along with nsp2 Ile^{120}Phe
262 (Figure 2a), which then appears in B.1.160. Likewise, Asn^{501}Tyr and Pro^{681}His in S appear in
263 divergent haplotypes, including one mink subgroup from Denmark and a basal node to B.1.351
264 (without the nsp6 deletion). It could be argued that those in S (Asn^{501}Tyr, Ser^{477}Asn, and
265 Pro^{681}His) are the result of multiple independent mutation events because they are under positive
266 selection (Martin et al. 2021), but we also identified a mutation in nsp3 that is one of the lineage-
267 defining mutations for B.1.1.7 and appears in disjoint nodes, but is unlikely to be under selection
268 because it is synonymous. It should also be noted that recombination can generate a high
269 number of false-positives when testing for signs of positive selection (Anisimova et al. 2003),
270 and the complexity of coronavirus recombinants compared to those generated in diploid
271 organisms through homologous chromosome crossovers (Figure 3d) makes that process difficult
272 to detect. Therefore, analyses that test for positive selection based on multiple independent
273 mutations at a site may, in fact, be false positives that result from recombination events. The
274 majority of the mutations found in B.1.1.7 could be explained by the admixture and
275 recombination among lineages and a random scan of 100 FASTQ files from B.1.1.7 available in
276 the NCBI SRA database identified two co-infected individuals in further support of this
277 hypothesis (Table S2). Large-scale analyses of these data may enable the detection of
278 recombinants.

279

280

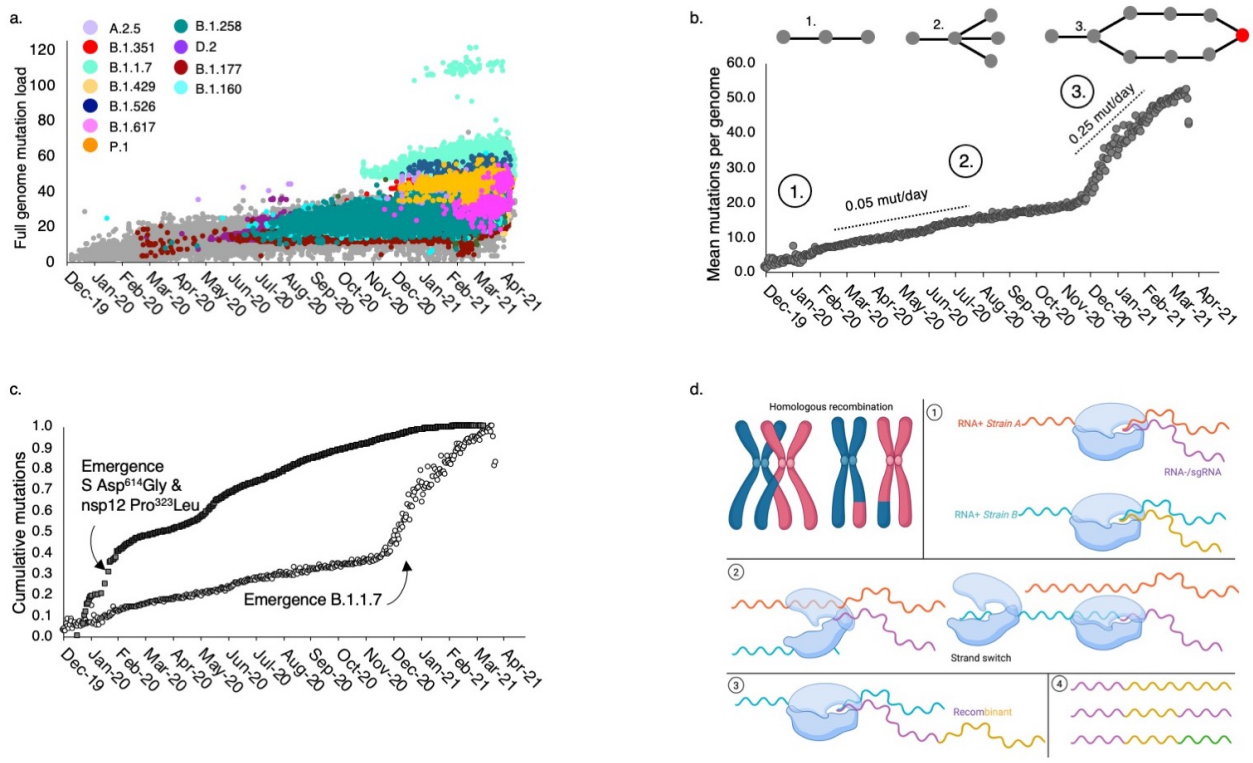


Fig. 3 Mutation rates and genomic mutation load of SARS-CoV-2. a. A rapid increase in the number of mutations per individual genome is evident in the late_2020_VOC. The outliers of the B.1.1.7 lineages (mint green) are a subset of that lineage due to a single, 57 base pair deletion in ORF7a (amino acids 5-23). **b.** Mean mutation load per individual, based on 2,128 high-confidence sites by date. The SARS-CoV-2 virus accumulated an estimated 0.05 mutations per day until the appearance of B.1.1.7, when it increased five-fold. Circles with numbers denote three processes occurring at different timepoints: (1) emergence, (2) haplotype expansion, and (3) recombination of divergent lineages. **c.** A population-level analysis of new mutations per day over the same time period (dark squares) displays a declining rate of mutations with a slight increase around the emergence of D.2 in Australia but not an increase with the emergence of B.1.1.7 that could explain the rapid accumulation of mutations shown in (b) plotted as percent accumulation (unfilled circles). **d**. Recombination in a diploid organism results from the crossover of homologous chromosomes during meiosis. In RNA+ betacoronaviruses, recombination occurs when two or more strains (haplotypes) infect a single cell (1). The replisome dissociates (2) from one strand and switches to another, (3) generating a hybrid recombinant. The resulting chimera (4) can be as simple as a section of *strain A* fused to a section of *strain B* or more complex recombinants if strand switching occurs more than once or there are multiple strains per cell (green section). Reference sequence: NC_045512, Wuhan, December 24, 2019.

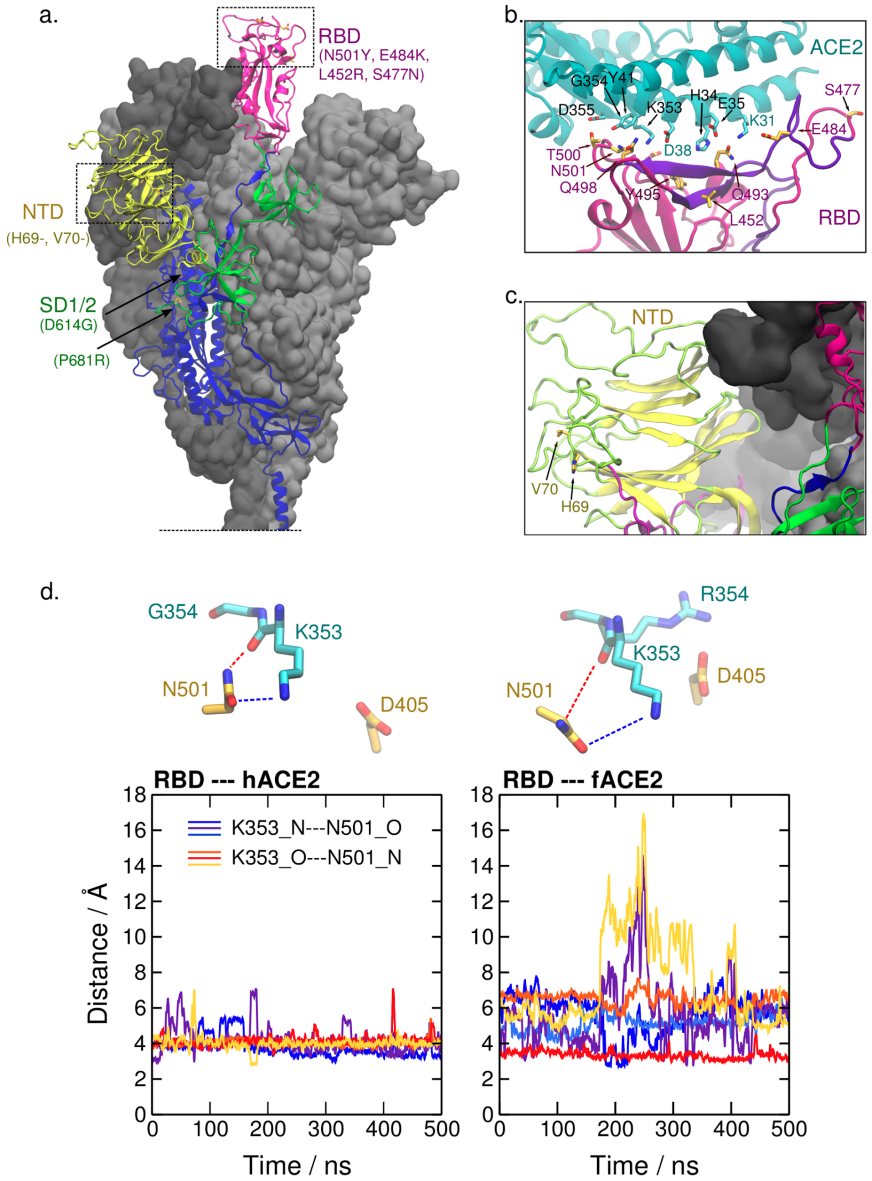


Fig. 4. Location of mutation sites of SARS-CoV-2 VOC on the structure of the spike glycoprotein. a. Several mutations associated with dominant haplotypes are located in the receptor-binding domain (RBD, aa. 331-506), N-terminal domain (NTD, aa. 13-305), and subdomains 1 and 2 (SD1/2, aa. 528-685) of S. The structure of S in the prefusion conformation derived from PDB ID 6VSB (Wrapp et al. 2020) and completed *in silico* (Casalino et al. 2020) is shown. Glycosyl chains are not depicted, and the S trimer is truncated at the connecting domain for visual clarity. The secondary structure framework of one protomer is represented and the neighboring protomers are shown as a gray surface. **b.** Mutation sites in the S RBD of SARS-CoV-2 VOC, such as 484, 452, 477, and 501 are located

306 at or near the interface with ACE2. Notably, site 452 and 484 reside in an epitope that is a target of the adaptive
307 immune response in humans (aa. 480-499, in violet) and site 501 is also located near it (B.-Z. Zhang et al. 2020).
308 Dashed lines represent relevant polar interactions discussed here. PDB ID 6M17 was used (R. Yan et al. 2020). **c.**
309 The sites 69 and 70 on the NTD, which are deleted in the VOC B.1.1.7, are also found near an epitope (aa. 21-45, in
310 violet) (B.-Z. Zhang et al. 2020). **d.** Time progression of N---O distances between atoms of Asn^{501} in RBD and
311 Lys^{353} in human and ferret ACE2 (hACE2 and fACE2, respectively) from the last 500 ns of the simulation runs.
312 Colors in the plots correspond to the distances Lys^{353}_N---Asn^{501}_O (cold colors) and Lys^{353}_O---Asn^{501}_N (warm
313 colors) in three independent simulations of each system. These distances are represented in the upper part of the
314 figure.

315

316 ***The potential functional impact of key mutations in S and non-S proteins***

317 Given the results from the MJN analysis and our previous hypothesis (Garvin, Prates, et al.
318 2020) that the cooperative effects of mutations in S and non-S proteins (i.e., epistasis) define and
319 are responsible for the increased transmission of prevalent SARS-CoV-2 variants (Lauring and
320 Hodcroft 2021), we performed protein structural analyses and discuss below the functional
321 effects of these individual and combined mutations in SARS-CoV-2 VOC. We analyze ten likely
322 key mutation sites (red font, Table 1) in S and non-S proteins.

323 *S Asn^{501}Tyr* - Located in the receptor-binding domain (RBD) of SARS-CoV-2 S,
324 immunoprecipitation assays reveal that site 501 plays a major role in the affinity of the virus to
325 the host receptor, ACE2 (Shang et al. 2020). Via structural analysis and extensive molecular
326 dynamics simulations, Ali et al. highlighted the importance of the interactions with human ACE2
327 (hACE2) near the site 501 of the receptor-binding domain of S, particularly via a sustained
328 hydrogen bond between RBD Asn^{498} and hACE2 Lys^{353} (Ali and Vijayan 2020). Deep

329 mutational scanning of SARS-CoV-2 RBD reveals that the naturally occurring mutations at site
330 501, Asn^{501}Tyr and Asn^{501}Thr, lead to an increased affinity to hACE2 (Starr et al.). Additionally,
331 this site is located near a linear B cell immunodominant site (B.-Z. Zhang et al. 2020), and
332 therefore the mutation may allow SARS-CoV-2 variants to escape neutralizing antibodies
333 (Figure 4, Figure 5). Indeed, neutralizing antibodies derived from vaccinations and natural
334 infection have significantly reduced activity against pseudotyped viruses carrying this mutation
335 (Wang et al. 2021).

336

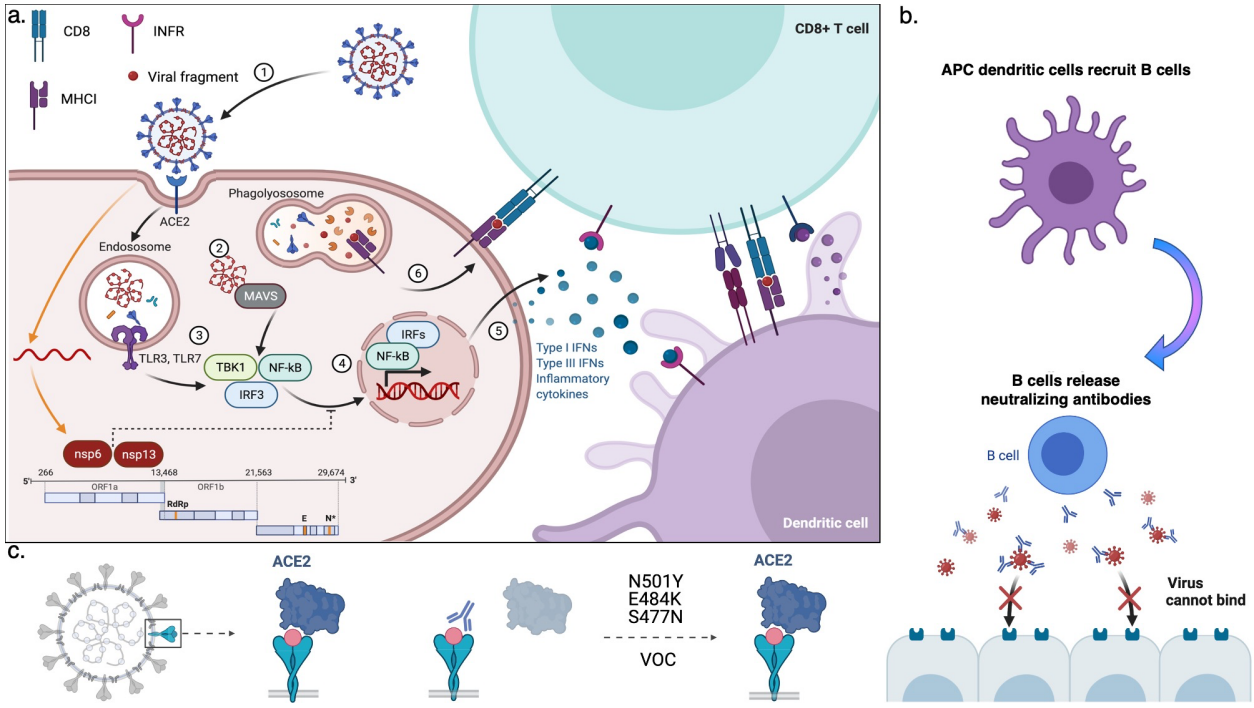
337 **Table 2. Surface exposed residues of ACE2 orthologues forming the region of contact with site 501 of SARS-**
338 **CoV-2 S**. Relative to the human sequence, almost all these residues are either conserved ("|") or replaced by a nearly
339 equivalent amino acid in mouse, American mink, European mink, ferret, and pangolin. Notably, there is a
340 nonconservative substitution of Gly^{354} to a bulky positively charged amino acid in most species. Our structural
341 analyses suggests that this substitution contributes to a putative host-dependent selective pressure at site 501 of
342 SARS-CoV-2 S. Prevalent residues reported at this site are informed in order of frequency.

Species	**Residues in ACE2**									**S 501**
***Homo sapiens* (Human)**	**D38**	**Y41**	**Q42**	**L45**	**K353**	**G354**	**D355**	**R357**	**I358**	**Y, N**
Mus musculus (House mouse)	|	|	|	|	H	|	|	|	|	Y, N
Neovison vison (American mink)	E	|	|	|	|	H	|	|	|	N, T
Mustela lutreola (European mink)	E	|	|	|	|	R	|	|	|	N, T
Mustela putorius furo (Ferret)	E	|	|	|	|	R	|	|	|	T, N
Manis pentadactyla (Pangolin)	E	|	|	|	|	H	|	|	|	N, T

343

344 Transmission between human and non-human hosts for SARS-CoV-2 provides further

information on the evolutionary selectivity of site 501 in S. Repeated infection of mice with human SARS-CoV-2 resulted in the selection of a mouse-adapted strain carrying S Tyr^{501} (Gu et al. 2020). It is possible that $Asn^{501}Tyr$ results in an additional stabilization of the RBD-ACE2 interaction via π-stacking of Tyr^{501} with Tyr^{41} in ACE2 (Figure 4a-b). In contrast, several introductions into farmed mink (*Neovison vison*), which caused a substantial increase in their mortality (Oude Munnink et al. 2021), have not led to the same selection. To date, reported sequences in GISAID of SARS-CoV-2 in this host carry either S Asn^{501}, which is prevalent, or S Thr^{501}, which appeared independently in mink farms (Table S3) (Oude Munnink et al. 2021). In ACE2 of these taxa, Tyr^{41} is conserved, but near this site, a larger, positively charged amino acid, His^{354}, replaces Gly^{354}. Table 2 shows that the amino acids in the RBD 501-binding region of the ACE2 orthologues are conserved, except for Gly^{354}, indicating that this site may play a key role in viral fitness.



359 **Fig. 5. Response to viral infection. a**. As part of the innate immune response, (Step 1) the SARS-CoV-2 virus is
360 internalized into endosomes and degraded. (Step 2) viral RNA activates the mitochondrial antiviral innate immunity
361 (MAVS) pathway and (Step 3) degraded proteins activate the toll receptor pathway (TLR3/TLR7), which result in
362 the (Step 4) phosphorylation of TBK1 and translocation of NF-*k*B and IRF3 to the nucleus, where they regulate the
363 transcription of immune genes including interferons (IFNs, Step 5). IFNs recruit CD8+ T cells that, (Step 6)
364 recognize fragments of the virus on the cell surface via their class I major histocompatibility complex (MHC I)
365 receptors and are activated by dendritic cells (antigen processing cells, or APC). If the virus bypasses innate
366 immunity (orange arrows) nonstructural proteins (nsp6 and nsp13) block the IRF3 nuclear translocation. **b.** APCs
367 recruit B lymphocytes and stimulate the production of antibodies that recognize SARS-CoV-2 S (whereas T cells
368 recognize fragments of S bound to MHC I). **c.** The neutralizing antibodies block binding of the virus to the ACE2
369 receptor and can prevent re-infection but mutations in the receptor-binding domain (RBD), e.g., S Asn^{501}Tyr,
370 prevent binding of the antibodies and the virus is then able to bind the receptor again even if individuals experienced
371 exposure to an earlier strain or were vaccinated. *Created with BioRender.com.*

372

373 Similarly, ferrets (*Mustela putorius furo*) and pangolins (*Manis pentadactyla)*, relevant
374 potential reservoirs of SARS-CoV-2, carry a large basic residue at site 354 (an arginine and
375 histidine, respectively). Sawatzki et al. reported that the constant exposure of ferrets to infected
376 humans did not result in natural transmission in a domestic setting, suggesting that ferret
377 infection may require improved viral fitness (Sawatzki et al. 2021). In agreement with that,
378 Richard et al. (Richard et al. 2020) reported that the adaptive substitution Asn^{501}Thr was detected
379 in all experimentally infected ferrets in the laboratory. In order to further investigate the role of
380 the Gly^{354} versus Arg^{354} in the adaptive mutation of site 501 in S RBD, we performed extensive
381 molecular dynamics simulations of the truncated complexes of Asn^{501}-carrying RBD of SARS-
382 CoV-2 and ACE2, from human (hACE2) and ferret (fACE2). The simulations indicate that there
383 is a remarkable difference in the interaction pattern between the two systems in the region

384 surrounding site 501 of RBD. Firstly, we identified the main ACE2 contacts with Asn^{501}, which
385 were the same for both species, namely, Tyr^{41}, Lys^{353}, and Asp^{355}, and we also show that the
386 intensity of these contacts is lower in the simulations of fACE2 (Table S4).

387 To investigate further, we analyzed structural features in the interaction between ACE2
388 Lys^{353} and RBD Asn^{501}. Distances between polar atoms computed from the simulations indicate
389 a weaker electrostatic interaction between this pair of residues in ferret compared to human
390 (Figure 4d). This effect is accompanied by a conformational change of fACE2 Lys^{353}. Figure S2
391 shows that, in ferret, the side chain of Lys^{353} exhibits more stretched conformations, i.e., a higher
392 population of the *trans* mode of the dihedral angle formed by the side chain carbon atoms. This
393 conformational difference could be partially attributed to the electrostatic repulsion between the
394 two consecutive bulky positively charged amino acids in ferrets, Lys^{353} and Arg^{354}. Additionally,
395 the simulations suggest a correlation, in a competitive manner, between other interactions that
396 these residues display with the RBD. For example, Figure S3 shows that the salt bridge
397 fACE2_Arg^{354}---RBD_Asp^{405} and the HB interaction fACE2_Lys^{353}---RBD_Tyr^{495} (backbone)
398 alternate in the simulations. This also suggests that the salt bridge formed by fACE2 Arg^{354} drags
399 Lys^{353} apart from RBD Asn^{501}, weakening the interaction between this pair of residues in ferrets
400 relative to humans.

401 Altogether, these analyses indicate that site 354 in ACE2 significantly influences the
402 interactions with RBD in the region of site 501 and is likely playing a major role in the
403 selectivity of the size and chemical properties of this residue in SARS-CoV-2. We propose that,
404 in contrast to Tyr^{501}, a smaller HB-interacting amino acid at site 501 of RBD, such as the
405 threonine reported in farmed mink and ferrets, may ease the interactions on the region, e.g., the
406 salt bridge between fACE2 Arg^{354} and RBD Asp^{405}. The differences in the region of ACE2 in

407 contact with site 501 seem to have a key role for host adaptation and are worth further
408 investigation as it may also reveal details of the origin of this zoonotic pandemic.

409 *S His^{69}_Val^{70} deletion* - The His^{69}_Val^{70} deletion (in B.1.1.7) is adjacent to a linear epitope at the
410 N-terminal domain of S (Figure 4a,c) (B.-Z. Zhang et al. 2020), suggesting it too may improve
411 fitness by reducing host antibody effectiveness.

412 *S Leu^{452}Arg* - The Leu^{452}Arg mutation in S is a core change in the early_2021_VOC (Table 1,
413 Figure 4a-b). Although Leu^{452} does not interact directly with ACE2, this mutation was shown to
414 moderately increase infectivity in cell cultures and lung organoids using Leu^{452}Arg-carrying
415 pseudovirus (Deng et al. 2021). It is possible that the substitution of the leucine, hydrophobic, to
416 arginine, a positively charged residue, creates a direct binding site with ACE2 via the
417 electrostatic interaction with Glu^{35}. However, in Starr et al., experiments with the isolated RBD
418 expressed on the cell surface of yeast show that this mutation is associated with enhanced
419 structural stability of RBD, while it only slightly improves ACE2-binding (Starr et al.). An
420 alternative but not mutually exclusive hypothesis is that it causes a local conformational change
421 that impacts the complex dynamic interchange between interactions of RBD with the spike
422 trimer itself and with the host receptor. Noteworthy, site 452 resides in a significant
423 conformational epitope in RBD and Leu^{452}Arg was shown to decrease binding to neutralizing
424 antibodies (Figure 4b) (Deng et al. 2021; Li et al. 2021).

425 As noted in Deng et al., S Leu^{425}Arg has been reported in rare variants starting in March 2020
426 from Denmark, i.e., several months before the surge of the VOC that carry this mutation
427 (B.1.427, B.1.429, and B.1.617) (Deng et al. 2021). This indicates that the high transmissibility
428 of the early_2021_VOC is not fully explained by the increased infectivity caused by Leu^{425}Arg

429 and combined mutations may be essential for the rapid spread. Besides the other mutations in the
430 spike in these VOC, the substitution Asp^{260}Tyr in the SARS-CoV-2 helicase (nsp13, below) is
431 especially interesting, as it was identified in the MJN analysis as a defining mutation of both
432 B.1.427 and B.1.429 variants.

433 *S Ser^{477}Asn* - Variants carrying the S Ser^{477}Asn mutation spread rapidly in Australia (Figure 1,
434 Figure 3b). This site, located at the loop β4-5 of the RBD, is predicted not to establish persistent
435 interactions with ACE2 (Ali and Vijayan 2020). However, deep scanning shows that this
436 mutation is associated with a slight enhancement of ACE2-binding. Molecular dynamics
437 simulations suggest that Ser^{477}Asn affects the local flexibility of the RBD at the ACE2-binding
438 interface, which could be underlying the highest binding affinity with ACE2 reported from
439 potential mean force calculations (Singh et al.). Additionally, this site is located near an epitope
440 and may alter antibody recognition and counteract the host immune response (Figure 4b).

441 *S Glu^{484}Lys* - A recent computational study suggests that Glu^{484} exhibits only intermittent
442 interactions with Lys^{31} in ACE2 (Ali and Vijayan 2020). Deep scanning shows that this mutation
443 is associated with higher affinity to ACE2 (Starr et al.) and may be explained by its proximity to
444 Glu^{75} in ACE2, which would form a salt bridge with Lys^{484}. Aside from the potential impact of
445 Glu^{484}Lys between virus-host cell interaction, this site is part of a linear B cell immunodominant
446 site (B.-Z. Zhang et al. 2020) and this mutation was shown to impair antibody neutralization
447 (Wang et al. 2021).

448 *S Pro^{681}Arg and Pro^{681}His* - These mutations in the multibasic furin cleavage site are particularly
449 relevant given the importance of this region for cell-cell fusion (Hoffmann et al. 2020; Papa et
450 al.). The presence of the multibasic motif of SARS-CoV-2 has shown to be essential to the

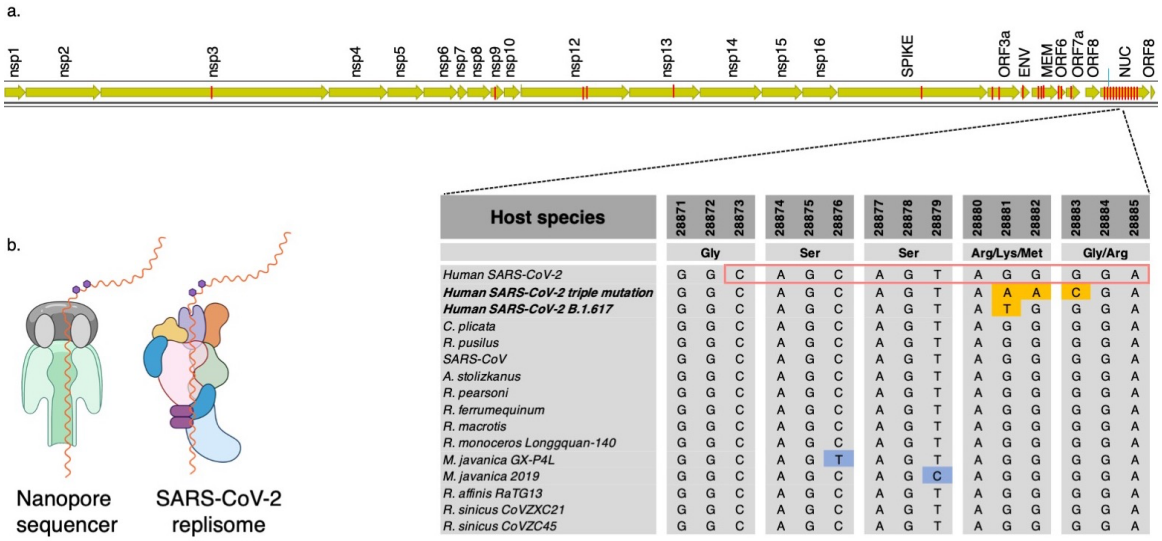
451 formation of syncytia (i.e., multinucleate fused cells), and thus it is thought to be a key factor
452 underlying pathogenicity and virulence differences between SARS-CoV-2 and other related
453 betacoronaviruses. Hoffmann et al. recognized the importance of the furin cleavage site in
454 SARS-CoV-2 and its biochemically basic signature and generated mutants to determine the
455 effects of specific amino acids. Notably, they showed that pseudotyped virion particles bearing
456 mutant SARS-CoV-2 S with additional basic residues in this region, including the substitution
457 $Pro^{681}Arg$ (present in B.1.617), exhibits a remarkable increase in syncytium formation in lung
458 cells *in vitro* (Hoffmann et al. 2020), which may explain the increased severity of the disease
459 (Sheikh et al. 2021). Hoffman et al did not include a $Pro^{681}His$ change that is a defining mutation
460 of B.1.1.7, and so it is not known if this too increases syncytium formation given that it is a basic
461 amino acid, but should be the target of future studies.

462 *Nucleocapsid* $Arg^{203}Met$ - The main function of the nucleocapsid (N) protein in SARS-CoV-2 is
463 to act as a scaffold for the viral genome and it is also the most antigenic protein produced by the
464 virus (Dutta et al. 2020). In a previous study, we reported that the Ser-Arg-rich motif of this
465 protein (a.a. 183-206), shown *in vitro* to be necessary for viral replication (Tylor et al. 2009;
466 Garvin, Prates, et al. 2020), displays a high number of amino acid changes during the COVID-
467 19 pandemic and is likely under positive selection. We propose that the RNA gene segment
468 coding this particular subsequence may be linked to improved fitness of specific SARS-CoV-2
469 haplotypes including the rapidly spreading Delta variant and is linked to epigenetic alterations
470 (Figure 6a).

471 A recent deep transcriptome sequencing study used Oxford Nanopore™ technology to
472 detect epigenetic modifications at 41 sites in the RNA genome that are associated with leader
473 sequence addition to sub-genomic RNA transcripts, a recombination-like process of SARS-CoV-

474 2 (Kim et al. 2020). Nanopore instruments can detect epigenetic modifications based on
475 disruptions of the electrical current as the RNA molecule passes through the molecular pore
476 (Rand et al. 2017; Simpson et al. 2017), which Kim et al propose is responsible for the pause that
477 occurs before leader sequence addition. Twenty-five of the 41 modified sites reside in the N gene
478 and the majority of the sites in this subset are found near the Ser-Arg-rich motif (Figure 6a).

479 Furthermore, one specific epigenetic site is linked to two highly successful SARS-CoV-2
480 haplotypes. The first is a triple mutation at sites 28881-28883 (GGG to AAC, $Arg^{203}Lys$) that is
481 now found in nearly half of all sequences sampled across the globe (diamond nodes, Figure 1)
482 and the second is $Arg^{203}Met$, which is a defining mutation for the rapidly spreading B.1.617.
483 Notably, this region of the genome is highly conserved across several hundred years of
484 coronavirus evolution (Boni et al. 2020) (Figure 6a). Given that these epigenetic sites were
485 discovered because the RNA pauses as it passes across the pore of the molecular nanopore
486 sequencer, one interesting hypothesis is that mutations at this region remove the epigenetic
487 modification and speed the SARS-CoV-2 genome through the replisome (Figure 6b), increasing
488 the production of virions, which is consistent with the more than 1000-fold higher virion count in
489 those infected with B.1.617 (Lu et al.).



Host species	28871	28872	28873	28874	28875	28876	28877	28878	28879	28880	28881	28882	28883	28884	28885
	Gly			Ser			Ser			Arg/Lys/Met			Gly/Arg		
Human SARS-CoV-2	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
Human SARS-CoV-2 triple mutation	G	G	C	A	G	C	A	G	T	A	A	A	C	G	A
Human SARS-CoV-2 B.1.617	G	G	C	A	G	C	A	G	T	A	T	G	G	G	A
C. plicata	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
R. pusilus	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
SARS-CoV	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
A. stolizkanus	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
R. pearsoni	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
R. ferrumequinum	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
R. macrotis	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
R. monoceros Longgquan-140	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
M. javanica GX-P4L	G	G	C	A	G	T	A	G	T	A	G	G	G	G	A
M. javanica 2019	G	G	C	A	G	C	A	G	C	A	G	G	G	G	A
R. affinis RaTG13	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
R. sinicus CoVZXC21	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A
R. sinicus CoVZC45	G	G	C	A	G	C	A	G	T	A	G	G	G	G	A

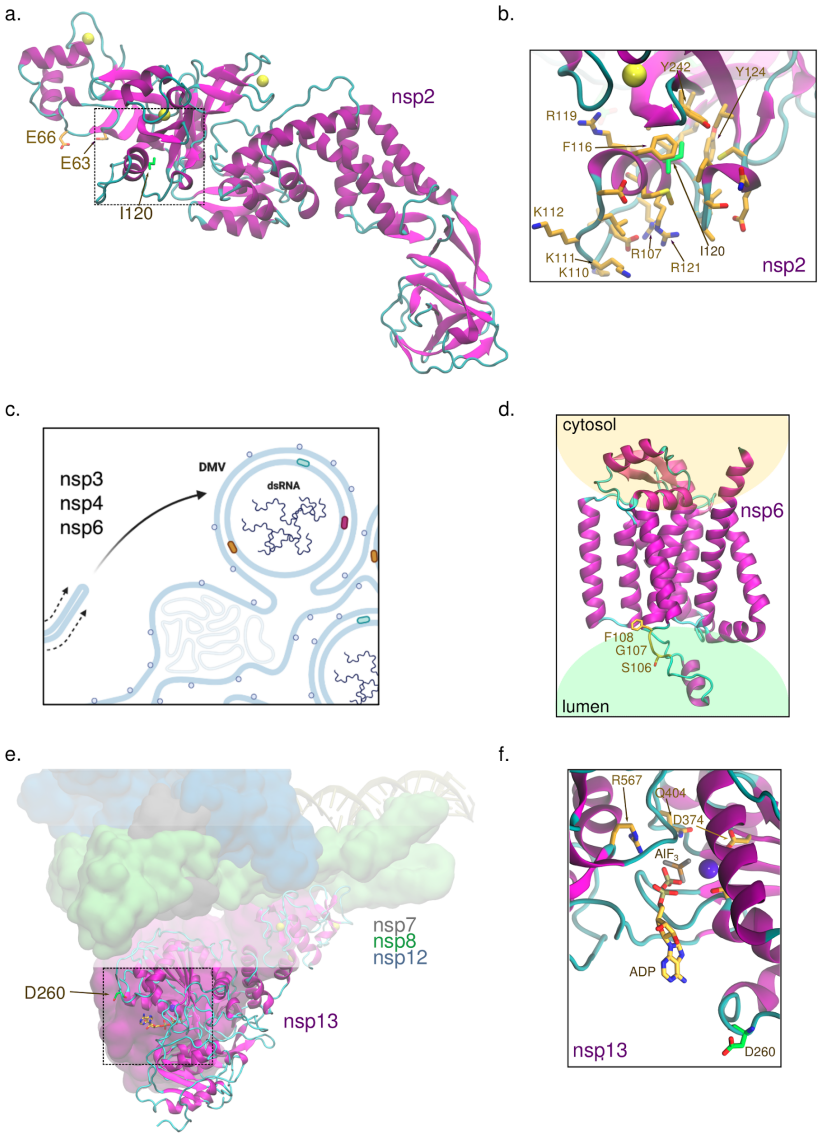
Fig. 6. Modifications at the Ser-Arg-rich region of N may affect replication speed. a. Location of 41 epigenetic sites reported in Kim et al. 2020 (red bars on SARS-CoV-2 genome). One of the sites in the nucleocapsid gene (nucleotides in red box of aligned sequences) is highly conserved across diverse host-defined coronaviruses. All bats and human coronavirus species from China are completely conserved at the epigenetic site 28881-28883, except for a 3-bp mutation in SARS-CoV-2 that occurred early in the pandemic and now corresponds to ~50% of all sequences globally (diamond nodes in Figure 1). **b.** Kim et al. proposed that N^6-methyladeonsine modification of the genome (purple hexagons), common in RNA viruses, caused the strand to pause while traversing the nanopore sequencing apparatus. We propose that loss of this site via mutations at site 203 in N may increase the replication rate of the RNA strand through the SARS-CoV-2 replisome. *Aselliscus stoliczkanus* - Stoliczka's trident bat, *Chaerephon plicata* - wrinkle-lipped free-tailed bat, *Rhinolophus pusillus* - least horseshoe bat, *R. pearsoni* - Pearson's horseshoe bat, *R. macrotis* - big-eared horseshoe bat, *R. ferrumequinum* - greater horseshoe bat, *R. monoceros* - Formosan lesser horseshoe bat, *R. affinis* intermediate horseshoe bat, *R. sinicus* Chinese rufous horseshoe bat, *R. mayalanis* - Mayalan horseshoe bat, *SARS* - Severe Acute Respiratory Syndrome, *Manis javanica* - Malayan pangolin. *Created with BioRender.com.*

507 *nsp2 Ile^{120}Phe* - The main role of the nonstructural protein 2 (nsp2) in viral performance is not
508 yet defined. Instead, this protein appears to be part of multiple interactions with host proteins
509 involved in a range of processes including the regulation of mitochondrial respiratory function,
510 endosomal transport, and ribosome biogenesis (Verba et al. 2021). Very recently, deep learning-
511 based methods of structure prediction and cryo-electron microscopy density were combined to
512 provide the atomic model of nsp2 (PDB id 7MSW). In a preprint from Verba et al., structural
513 information was used to localize the surfaces that are key for protein-protein interaction with
514 nsp2 (Verba et al. 2021). From structural analysis and mass spectrometry experiments, the
515 authors pose the interesting hypothesis that nsp2 interacts directly with ribosomal RNA via a
516 highly conserved zinc ribbon motif to bring ribosomes close to the replication-transcription
517 complexes.

518 Here we are particularly interested in the functional impact of the mutation Ile^{120}Phe in nsp2
519 present in the D.2 variant. Site 120, identified in the nsp2 structure on Figure 7a, is a point of
520 hydrophobic contact between a small helix, rich in positively charged residues, and a zinc
521 binding site. The positively charged surface of the helix may be especially relevant for a putative
522 interaction with the phosphate groups from ribosomal RNA. Normal mode analysis from
523 DynaMut2 predicts that the substitution has a destabilizing effect in the protein structure
524 (estimated $\Delta\Delta G^{stability}$ = -2 kcal/mol) (Rodrigues et al. 2021). Possibly, this could be caused by π-
525 π stacking interactions of the tyrosine with aromatic residues in the same helix that would disrupt
526 the contacts anchoring it to the protein core (Figure 7b). Additionally, site 120 is spatially close
527 to Glu^{63} and Glu^{66}, which were shown to be relevant for interactions with the endosomal/actin
528 machinery via affinity purification mass spectrometry in HEK293T cells. Remarkably, upon

529 mutation of these glutamates to lysines, there is increased interactions with proteins involved in

530 ribosome biogenesis (Verba et al. 2021).



531

532 **Fig. 7. Location of mutations of prevalent SARS-CoV-2 variants on the structure of the nonstructural**

533 **proteins nsp2, nsp6, and nsp13. a.** Site 120 in nsp2 is located in a small helix near a zinc-binding site and residues

534 Glu^{63} and Glu^{66}, which play a role in the interaction with proteins involved in ribosome biogenesis and in the

535 endosomal/actin machinery (Verba et al. 2021). PDB id 7MSW was used. **b.** Ile^{120} forms some of the hydrophobic

536 contacts that anchor the helix at the surface of nsp2, where this site resides, to the protein core. **c.** Nsp6 participates

537 in generating double-membrane vesicles (DMV) for viral genome replication. Natural selection for the biological

538 traits of viral entry and replication may explain the increased transmission of variants with adaptive mutations in
539 both S and nsp6. DMVs isolate the viral genome from host cell attack to provide for efficient genome and sub-
540 genome replication and generate virions. **d.** Sites 106-108 are predicted to be located at/near the protein region of
541 nsp6 embedded in the endoplasmic reticulum lumen (structure generated by AlphaFold2 (Jumper et al.)). **e.** Nsp13 is
542 the SARS-CoV-2 helicase, and it is part of the replication complex. **f.** Asp^{260} in nsp13 is mutated to tyrosine in
543 B.1.427 and B.1.429 and it is located at the entrance of the NTP-binding site. PDB id 6XEZ was used (Chen et al.
544 2020).

545 *nsp6 Ser^{106}_Phe^{108}deletion* - The nsp6 protein plays critical roles in viral replication and
546 suppression of the host immune response (Figure 5a and Figure 7c) (Gupta et al. 2020). Along
547 with nsp3 and nsp4, nsp6 is responsible for producing double-membrane vesicles from the
548 endoplasmic reticulum (ER) to protect the viral RNA from host attack and increase replication
549 efficiency (Figure 7c) (Santerre et al. 2020). The nsp6 Ser^{106}_Phe^{108}del is predicted to be located
550 at a loop in the interface between a transmembrane helix and the ER lumen based on a
551 preliminary structural analysis of the model generated by the AlphaFold2 system (Figure 7d),
552 and we hypothesize that the deletion may affect functional interactions of nsp6 with other
553 proteins. In addition, in agreement with the enhanced suppression of innate immune response
554 reported for B.1.1.7 (Thorne et al. 2021), changes in immune-antagonists, such as nsp6
555 Ser^{106}_Phe^{108}del, may be key to prolonged viral shedding (Calistri et al. 2021).

556 *nsp13 $Asp^{260}Tyr$* - The nonstructural protein 13 is a component of the viral replication-
557 transcription complex, (nsp13; or SARS-CoV-2 helicase) and plays an essential role in
558 unwinding the duplex oligonucleotides into single strands in a NTP-dependent manner (L. Yan et
559 al. 2020). Hydrogen/deuterium exchange mass spectrometry demonstrates that the helicase and
560 NTPase activities of SARS-CoV nsp13 are highly coordinated, and mutations at the NTPase
561 active site impair both ATP hydrolysis and the unwinding process (Jia et al. 2019). Here we note

562 that the substitution Asp^{260}Tyr, present in B.1.427 and B.1.429, is located at the entrance of the
563 NTPase active site and may favor π-π stacking interactions with nucleobases (Figure 7e-f). Given
564 that, at high ATP concentrations, SARS-CoV nsp13 exhibits increased helicase activity on
565 duplex RNA (Jang et al. 2020), it is possible that, similarly, the putative optimization on NPT
566 uptake in nsp13 Asp^{260}Tyr favors RNA unwinding.

567 Additionally, nsp13 was shown to play an important role as an innate immune antagonist
568 (Figure 5a). It contributes to the inhibition of the type I interferon response by directly binding to
569 TBK1 and, with that, it impedes IRF3 phosphorylation (Guo et al. 2021). The dual role of nsp6
570 and nsp13 in immune suppression and viral replication may suggest a convergent evolution of
571 SARS-CoV-2 manifested in most of the VOC, which carries either nsp6 Ser^{106}_Phe^{108}del or
572 nsp13 Asp^{260}Tyr.

573

574 **3. Concluding Remarks**

575 From our thorough analysis of the spatiotemporal relationships of SARS-CoV-2 variants, we
576 propose that the rapid increase of mutations in the late 2020 VOC is likely a consequence of the
577 recombination of haplotypes carrying adaptive mutations in S and in non-S proteins that act
578 cooperatively to enhance viral fitness. For example, as indicative of that, we call attention to five
579 mutations that occur independently in disjoint clusters of our MJN, four of which (S Asn^{501}Tyr,
580 S His^{69}_Val^{70}del, S Phe^{681}His/Arg, and nsp6 Ser^{106}_Phe^{108}del) are shared by different VOC,
581 including B.1.1.7. Notably, S His^{69}_Val^{70}del appeared in human and mink populations
582 simultaneously in August 2020, prior to the emergence of B.1.1.7, indicating that mink should be
583 further investigated as a possible component of a recombination event. In turn, our molecular

584 dynamics simulations indicate that the molecular forces at site 501 in S and how they are altered
585 upon mutation (S Asn^{501}Tyr in B.1.1.7) are a key component to describe the history of
586 transmission among other putative zoonotic reservoirs, such as farmed minks, ferrets, and
587 pangolins.

588 The S Asp^{614}Gly mutation has been shown to increase infectivity and is now predominant
589 in the circulating virus (L. Zhang et al. 2020), and S Asn^{501}Tyr is associated with higher
590 virulence (Gu et al. 2020). We show that the expansion of the strains carrying these mutations
591 only occurred upon the additional substitutions in nsp12 Leu^{323}Pro (Figure 2b) (Garvin, Prates, et
592 al. 2020) and nsp6 Ser^{106}_Phe^{108}del, respectively. A hypothesis consistent with these
593 observations is that the changes in S enhance viral entry into the host cells but they do not easily
594 transmit due to rapid suppression by a robust innate immune response. A secondary mutation is
595 able to counteract the immune-driven suppression. In the case of S Asp^{614}Gly, the nsp12
596 Leu^{323}Pro may have increased the replication rate of the virus, which was supported by
597 quantitative PCR from clinical samples with different viral strains in Korber et al. (Garvin,
598 Prates, et al. 2020; Korber et al. 2020). However, the separate effects of S Asp^{614}Gly and nsp12
599 Leu^{323}Pro could not be described in the referred study because it did not include individuals
600 infected with variants harboring only one of the mutations.

601 For the late 2020 VOC, nsp6 Ser^{106}_Phe^{108}del may affect viral replication in DMVs or
602 suppress the interferon-driven antiviral response (Xia et al. 2020). It is likely that other mutations
603 also enhance viral mechanisms that impair the host immune response. For example, Thorne et al.
604 recently showed that the B.1.1.7 VOC suppresses the innate immune response by host cells *in*
605 *vitro* and attributed it to the increased transcription of the *orf9b* gene, nested within the gene
606 coding the nucleocapsid protein. (Thorne et al. 2021), although they could not rule out the

607 possibility that this was due to nsp6 Ser^{106}_Phe^{108}del.

608 Via focused protein structural analysis, we identify other mutations shared among
609 different VOC that reside in key locations of proteins involved in viral replication and/or in
610 suppressing the innate immune suppression, such as nsp13, suggesting a convergent evolution of
611 SARS-CoV-2. This emphasizes the importance of tracking mutations in a genome-wide manner
612 as a strategy to avoid the emergence of future VOC. For example, an earlier dominant variant in
613 Australia (D.2) that carried the mutations Ser^{477}Asn in S and Ile^{120}Phe in nsp6 was successfully
614 restrained. However, we note that variants harboring only the S Ser^{477}Asn substitution are
615 currently circulating in several European countries (Figure 2, Table S5) and may only need to
616 recombine with a variant carrying an advantageous complementary mutation to become the next
617 VOC.

618 A second and equally significant outcome from recombination-driven haplotypes is the
619 generation of variants that allow escape from neutralizing antibodies produced by an adaptive
620 immune response (Garvin, T Prates, et al. 2020) (Figure 5c). As a case in point, the resurgence of
621 COVID-19 in Manaus, Brazil, in January 2021, where seroprevalence was above 75% in October
622 2020, is due to immune escape of new SARS-CoV-2 lineages (Sabino et al. 2021). Broad disease
623 prevalence and community spread of COVID-19 increase the probability that divergent
624 haplotypes may come in contact, thereby dramatically accelerating the evolution and
625 transmission of the virus. This emphasizes that regions with low sequence surveillance can be
626 viral breeding grounds for the next SARS-CoV-2 VOC. Lastly, it is apparent that the adaptive
627 evolution of the SARS-CoV-2 virus to vaccinated individuals is generating forms that are
628 harmful to those who are unvaccinated, making it clear that a multi-pronged approach that
629 includes increased vaccination rates, accurate predictive models of VOC, and more effective

630 treatments against disease will be necessary if we are to put this pandemic behind us.

631

632 **4. Methods**

633 *Sequence data pre-processing*

634 We downloaded SARS-CoV-2 sequences in FASTA format and corresponding metadata from
635 GISAID and processed as we have reported previously (Garvin, Prates, et al. 2020; Prates et al.
636 2021). To ensure that deletions were accounted for, full genome sequences were aligned with
637 MAFFT (Katoh et al. 2002) to the established reference genome (accession NC_045512),
638 uploaded into CLC Genomics Workbench, and trimmed to the start and stop codons (nsp1 start
639 site and ORF10 stop codon). Aligned sequences in tab-delimited format were imported into R to
640 count the number of variable accessions at each of the 29,409 sites.

641 Variable sites were determined with all sequences downloaded up through the end of
642 January, 2021. In order to reduce false-positive mutation sites (those that were due to technical
643 error), we selected sites that were variable in 25 or more individuals (0.01%) compared to the
644 reference (all 25 were required to be the same state: A, G, T, C, or -). We further pruned these
645 by removing sites in which 20% or more of the accessions harbored an unknown character state
646 ("N"), leaving 2,128 variable sites for downstream analyses. After removing sequences with an
647 "N" at any of these sites, we retained 280,409 individuals. Prior to submission, we updated the
648 number of sequences through April 19, 2021, keeping the same 2128 variable sites, which
649 allowed us to capture the most up-to date metadata and produced 640,211 for analysis. We kept
650 haplotypes that occurred in more than 35 individuals to remove rare or artifact-derived

651 haplotypes. For the comparison of median-joining networks and phylogenetic trees, we used
652 sequences from the pandemic sampled through the end of April 2020. We used variable sites
653 found in more than ten individuals and haplotypes found in five or more individuals as we had in
654 previous work (Garvin, Prates, et al. 2020). This produced 410 unique haplotypes based on 467
655 variable sites.

656

657 *Median-joining network (MJN)*

658 Haplotypes were coded in NEXUS format and uploaded to PopArt (Leigh and Bryant 2015). An
659 MJN was produced with the epsilon parameter set to 0. The networks were exported as a table
660 and visualized in Cytoscape (Shannon et al. 2003) with corresponding metadata. The date of
661 emergence of each haplotype was defined by the sample date subtracted from the report date for
662 the Wuhan reference sequence (December 24, 2019) and then one day was added to remove
663 zeros. For samples that only reported the month but no day, we recorded the day as the 15th of
664 that month. We excluded samples with no sampling date.

665

666 *Phylogenetic tree*

667 We used the program MrBayes to generate a phylogenetic tree (Ronquist and Huelsenbeck
668 2003). Parameters were set to *Nucmodel=4by4*, *Nst=6*, *Code=Universal*, and *Rates=Invgamma*.
669 We performed 5,000,000 mcmc generations, which produced a stable standard deviation of split
670 frequencies of 0.014. A consensus tree was generated using the 50% majority rule and visualized
671 using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

672

673 *Estimation of genome mutation load*

674 We estimated the mutation load using two data sets. First, we used the 640,211 sequences based
675 on 2,128 variable sites used for the MJN because these represent high-confidence mutations. For
676 each of the 640,211 accessions, we counted the number of differences of the 2,128 variable sites
677 compared to the reference genome (accession NC_045512) and recorded the day of emergence.
678 The mutational load for all accessions for a given day was then averaged and this was plotted
679 across time. For the second estimate of mutation rate, we used all variable sites across the full
680 genome (29,409 sites) to include rare variants and removed all sequences with at least one
681 ambiguous site, leaving 584,119 accessions.

682 For the population-level estimate of mutation accumulation, we applied the filters used to
683 identify the 2,128 variable sites that were used for the MJN for all sequences up through April
684 19, 2021. We did not include new mutations because the B.1.1.7 VOC and its downstream
685 haplotypes had become the predominant variants globally at that time and, consequently, much
686 early information of the molecular evolution is lost when applying frequency filters on the entire
687 GISAID database. This is exacerbated with the MJN approach because the software algorithm
688 used to generate the network is computationally intractable with greater than 1,000 haplotypes
689 and therefore future efforts will either need to ignore early molecular events or use new methods
690 that can handle the large datasets and any recombination events that occur (an alternative
691 approach would be to now use the Alpha or Delta variant as the reference sequence because they
692 are now the predominant strains globally).

For calculations of population-level mutation accumulation, it is possible (and necessary) to include all sequences to determine if mutation or recombination are the cause of the high mutation load seen in the late 2020 VOC. After applying the frequency and haplotype filters, we retained 5,011 variable sites that define 12,282 unique haplotypes for further analysis. Mutations to five possible states (A, G, T, C, and -) were counted at each site on the first date that they appeared and their appearance at later dates were excluded. Multiple mutations at a site to different states were counted with this method.

For lineage-specific mutation curves, we extracted all sequences based on their PANGO lineage listed in the metadata from GISAID that also had a sample data and plotted the cumulative number over time, where time is represented by days from first appearance. To estimate the rate of accumulation, we calculated the slope for the linear portion of each of the curves.

Probability of mutation accumulation

To calculate the chance of accumulating several mutations in a certain period, the probability density function for a normal distribution is used:

$$PDF(x) = exp(-(x-\mu)^2/2\sigma^2)/sqrt(2\,\pi * \sigma^2),$$

where μ is the expected number of mutations for that date, x is the measured value, and σ is the standard deviation of error calculated from the data shown in Fig. 1b, considering the difference between the actual and predicted number of mutations. The expected value of mutations μ for a given time period is computed from the estimated rate of mutations per day (Figure 3, 0.05). c.

714 The period of interest to our discussion (June-October 2020) corresponds to 122 days, for which,
715 the integral of PDF(x=13) gives the probability of $1*10^{-15}$ to accumulate 13 mutational events.

716

717 *Screen for coinfected individuals with UK B.1.1.7*

718 We extracted 25 samples from the Sequence Read Archive at NCBI for each of the months of
719 October, November, December, and January listed as variant B.1.1.7 from the UK (Table S2) for
720 a total of 100 samples to check for coinfection. The reads were mapped to the NC_045512
721 Wuhan reference using CLC Genomics Workbench using the default parameters except for
722 length fraction and similarity fraction were set to 0.9. Three sites specific to UK B.1.1.7 were
723 analyzed for possible heterozygosity. Of the 100 we sampled, two appeared to be cases of
724 coinfection. This supports the hypothesis that the large expansion in overall mutations seen in
725 UK B.1.1.7 are likely due to recombination. In addition, it also supports the case that coinfection
726 is occurring at a baseline sufficient to allow for occasional recombination.

727

728 *Protein structure analysis*

729 VMD was used to visualize the protein structures and analyze the potential functional effects of
730 mutations (Humphrey et al. 1996). Figure 3 was created using Inkscape (https://inskape.org/)
731 and Gimp 2.8 (https://www.gimp.org) (Anon).

732

733 *Molecular dynamics simulations*

734 Molecular dynamics (MD) simulations were used to study interactions between SARS-CoV-2
735 RBD and ACE2 from ferret and human. Three independent extensive MD simulations were
736 performed for each species using GROMACS 2020 package (Lindahl et al. 2020) and the
737 CHARMM36 force field for protein and glycans (Guvench et al. 2011; Huang and MacKerell
738 2013). Each simulation ran up to 800 ns, being the last 500 ns used for analysis. PDB id 6M17
739 was used to build the ACE2-RBD complexes. Given the high sequence identity between human
740 and ferret ACE2 (83%), we performed local modeling of the non-conserved amino acid residues
741 in ferret ACE2 using the human homolog as the template, via RosettaRemodel (Huang et al.
742 2011).

743 The inputs for simulations were generated using CHARMM-GUI (Jo et al. 2008).
744 Counterions were added for electroneutrality (0.1 M NaCl). The complexes were surrounded by
745 TIP3P water molecules to form a layer of at least 10 Å relative to the box borders (Jorgensen et
746 al. 1983). Simulations were performed using the NPT ensemble. The temperature was
747 maintained at 310 K with the Nosé–Hoover thermostat using a time constant of 1.0 ps (Evans
748 and Holian 1985). The pressure was maintained at 1 bar with the isotropic Parrinello–Rahman
749 barostat using a compressibility of 4.5×10^{-5} bar^{-1} and a time constant of 1.0 ps in a rectangular
750 simulation box (Parrinello and Rahman 1981). The particle mesh Ewald method was used for the
751 treatment of periodic electrostatic interactions with a cutoff distance of 1.2 nm (Darden et al.
752 1993). The Lennard–Jones potential was smoothed over the cutoff range of 1.0–1.2 nm by using
753 the force-based switching function. Only atoms in the Verlet pair list with a cutoff range
754 reassigned every 20 steps were considered. The LINCS algorithm was used to constrain all
755 bonds involving hydrogen atoms to allow the use of a 2 fs time step (Hess et al. 1997). The

suggested protocol for nonbonded interactions with the CHARMM36 force field when used in the GROMACS suite was followed.

The Hbonds plugin in VMD was used to identify hydrogen bond interactions along the simulations (Humphrey et al. 1996). The geometric criteria adopted are a cutoff of 3.5 Å for donor-acceptor distance and 30° for acceptor-donor-H angle. The Timeline plugin was used to count contacts formed by a given amino acid residue. We defined the distance of 4 Å between any atom pairs as the cutoff for contact.

5. Data Access

All SARS-CoV-2 sequences used in this study are available from the public repositories Genome Initiative on Sharing Avian Influenza Data (GISAID, gisaid.org), the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/sars-cov-2/) and the COVID-19 Genomics UK Consortium (COG, https://www.sanger.ac.uk/collaboration/covid-19-genomics-uk-cog-uk-consortium/

6. Acknowledgments

The viral evolution research was funded by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC for the US Department of Energy (LOIS:10074) and the structural implication work was funded via the DOE Office of Science through the National Virtual Biotechnology Laboratory (NVBL), a

consortium of DOE national laboratories focused on the response to COVID-19, with funding provided by the Coronavirus CARES Act. This work was also funded by the United States Government. This research used resources of the Oak Ridge Leadership Computing Facility (OLCF) and the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. Figures generated with Biorender and VMD. We gratefully acknowledge the Originating laboratories responsible for obtaining the viral specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based.

Author Contribution

MR Garvin: Conceptualization, Data curation, Funding acquisition, Formal Analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing.

ET Prates: Formal Analysis, Investigation, Visualization, Writing - original draft, Writing - review & editing

J Romero: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing.

A Cliff: Methodology, Software, Writing – review & editing

JGFM Gazolla: Software, Formal Analysis, Investigation, Data Curation, Visualization, Writing - Review and Editing.

M Pickholz: Investigation, Visualization, Writing – original draft, Writing – review & editing

M Pavicic: Investigation, Writing – original draft, Writing – review & editing

DA Jacobson: Conceptualization, Funding acquisition, Formal Analysis, Investigation, Project administration, Supervision, Resources, Writing – original draft, Writing – review & editing

801 **7. References**

802 Ali A, Vijayan R. 2020. Dynamics of the ACE2-SARS-CoV-2/SARS-CoV spike protein
803 interface reveal unique mechanisms. *Sci. Rep.* 10:14214.

804 Alpert T, Brito AF, Lasek-Nesselquist E, Rothman J, Valesano AL, MacKay MJ, Petrone ME,
805 Breban MI, Watkins AE, Vogels CBF, et al. 2021. Early introductions and transmission of
806 SARS-CoV-2 variant B.1.1.7 in the United States. *Cell* 184:2595–2604.e13.

807 Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the
808 likelihood method for detecting positive selection at amino acid sites. *Genetics* 164:1229–
809 1236.

810 Anon. The GIMP Development Team. (2019). GIMP. Retrieved from https://www.gimp.org.
811 *https://www.gimp.org*.

812 Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific
813 phylogenies. *Mol. Biol. Evol.* 16:37–48.

814 Bentley K, Evans DJ. 2018. Mechanisms and consequences of positive-strand RNA virus
815 recombination. *J. Gen. Virol.* 99:1345–1356.

816 Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, Rambaut A, Robertson DL.
817 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the
818 COVID-19 pandemic. *Nat Microbiol* 5:1408–1417.

819 Calistri P, Amato L, Puglia I, Cito F, Di Giuseppe A, Danzetta ML, Morelli D, Di Domenico M,
820 Caporale M, Scialabba S, et al. 2021. Infection sustained by lineage B.1.1.7 of SARS-CoV-

821 2 is characterised by longer persistence and higher viral RNA loads in nasopharyngeal
822 swabs. *Int. J. Infect. Dis.* 105:753–755.

823 Casalino L, Gaieb Z, Goldsmith JA, Hjorth CK, Dommer AC, Harbison AM, Fogarty CA,
824 Barros EP, Taylor BC, McLellan JS, et al. 2020. Beyond Shielding: The Roles of Glycans
825 in the SARS-CoV-2 Spike Protein. *ACS Cent Sci* 6:1722–1734.

826 Challen R, Brooks-Pollock E, Read JM, Dyson L, Tsaneva-Atanasova K, Danon L. 2021. Risk
827 of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched
828 cohort study. *BMJ* 372:n579.

829 Chen J, Malone B, Llewellyn E, Grasso M, Shelton PMM, Olinares PDB, Maruthi K, Eng ET,
830 Vatandaslar H, Chait BT, et al. 2020. Structural Basis for Helicase-Polymerase Coupling in
831 the SARS-CoV-2 Replication-Transcription Complex. *Cell* 182:1560–1573.e13.

832 Darden T, York D, Pedersen L. 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums
833 in large systems. *J. Chem. Phys.* 98:10089–10092.

834 Davies NG, Jarvis CI, John Edmunds W, Jewell NP, Diaz-Ordaz K, Keogh RH, CMMID
835 COVID-19 Working Group. Increased mortality in community-tested cases of SARS-CoV-
836 2 lineage B.1.1.7. Available from: http://dx.doi.org/10.1101/2021.02.01.21250959

837 Deng X, Garcia-Knight MA, Khalid MM, Servellita V, Wang C, Morris MK, Sotomayor-
838 González A, Glasner DR, Reyes KR, Gliwa AS, et al. 2021. Transmission, infectivity, and
839 neutralization of a spike L452R SARS-CoV-2 variant. *Cell* 184:3426–3437.e8.

840 Dutta NK, Mazumdar K, Gordy JT. 2020. The Nucleocapsid Protein of SARS–CoV-2: a Target

841 for Vaccine Development. *Journal of Virology* [Internet] 94. Available from:
842 http://dx.doi.org/10.1128/jvi.00647-20

843 Evans DJ, Holian BL. 1985. The Nose–Hoover thermostat. *J. Chem. Phys.* 83:4069–4074.

844 Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, Crispim MAE, Sales
845 FCS, Hawryluk I, McCrone JT, et al. 2021. Genomics and epidemiology of the P.1 SARS-
846 CoV-2 lineage in Manaus, Brazil. *Science* 372:815–821.

847 Fratev F. The N501Y and K417N mutations in the spike protein of SARS-CoV-2 alter the
848 interactions with both hACE2 and human derived antibody: A Free energy of perturbation
849 study. Available from: http://dx.doi.org/10.1101/2020.12.23.424283

850 Funk T, Pharris A, Spiteri G, Bundle N, Melidou A, Carr M, Gonzalez G, Garcia-Leon A,
851 Crispie F, O'Connor L, et al. 2021. Characteristics of SARS-CoV-2 variants of concern
852 B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021.
853 *Euro Surveill.* [Internet] 26. Available from: http://dx.doi.org/10.2807/1560-
854 7917.ES.2021.26.16.2100348

855 Garvin MR, Prates ET, Pavicic M, Jones P, Amos BK, Geiger A, Shah M, Streich J, Gazolla
856 JGFM, Kainer D, et al. 2020. Potentially adaptive SARS-CoV-2 mutations discovered with
857 novel spatiotemporal and explainable-AI models. *Genome Biol.* in press.

858 Garvin MR, T Prates E, Pavicic M, Jones P, Amos BK, Geiger A, Shah MB, Streich J, Felipe
859 Machado Gazolla JG, Kainer D, et al. 2020. Potentially adaptive SARS-CoV-2 mutations
860 discovered with novel spatiotemporal and explainable AI models. *Genome Biol.* 21:304.

861 Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, Hilton SK, Huddleston J,
862 Eguia R, Crawford KHD, et al. Complete mapping of mutations to the SARS-CoV-2 spike
863 receptor-binding domain that escape antibody recognition. Available from:
864 http://dx.doi.org/10.1101/2020.09.10.292078

865 Gribble J, Stevens LJ, Agostini ML, Anderson-Daniels J, Chappell JD, Lu X, Pruijssers AJ,
866 Routh AL, Denison MR. 2021. The coronavirus proofreading exoribonuclease mediates
867 extensive viral recombination. *PLoS Pathog.* 17:e1009226.

868 Gu H, Chen Q, Yang G, He L, Fan H, Deng Y-Q, Wang Y, Teng Y, Zhao Z, Cui Y, et al. 2020.
869 Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science*
870 369:1603–1607.

871 Guo G, Gao M, Gao X, Zhu B, Huang J, Luo K, Zhang Y, Sun J, Deng M, Lou Z. 2021. SARS-
872 CoV-2 non-structural protein 13 (nsp13) hijacks host deubiquitinase USP13 and counteracts
873 host antiviral immune response. *Signal Transduct Target Ther* 6:119.

874 Gupta R, Charron J, Stenger CL, Painter J, Steward H, Cook TW, Faber W, Frisch A, Lind E,
875 Bauss J, et al. 2020. SARS-CoV-2 (COVID-19) structural and evolutionary dynamicome:
876 Insights into functional evolution and human genomics. *J. Biol. Chem.* 295:11742–11753.

877 Guvench O, Mallajosyula SS, Raman EP, Hatcher E, Vanommeslaeghe K, Foster TJ, Jamison
878 FW 2nd, Mackerell AD Jr. 2011. CHARMM additive all-atom force field for carbohydrate
879 derivatives and its utility in polysaccharide and carbohydrate-protein modeling. *J. Chem.*
880 *Theory Comput.* 7:3162–3180.

881 Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. 1997. LINCS: A linear constraint solver for

882 molecular simulations. *J. Comput. Chem.* 18:1463–1472.

883 Hoffmann M, Kleine-Weber H, Pöhlmann S. 2020. A Multibasic Cleavage Site in the Spike
884 Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cell* 78:779–
885 784.e5.

886 Huang J, MacKerell AD Jr. 2013. CHARMM36 all-atom additive protein force field: validation
887 based on comparison to NMR data. *J. Comput. Chem.* 34:2135–2145.

888 Huang P-S, Ban Y-EA, Richter F, Andre I, Vernon R, Schief WR, Baker D. 2011.
889 RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One*
890 6:e24109.

891 Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.*
892 14:33–38, 27–28.

893 Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol.*
894 *Biol. Evol.* 23:254–267.

895 Jang K-J, Jeong S, Kang DY, Sp N, Yang YM, Kim D-E. 2020. A high ATP concentration
896 enhances the cooperative translocation of the SARS coronavirus helicase nsP13 in the
897 unwinding of duplex RNA. *Sci. Rep.* 10:4481.

898 Jia Z, Yan L, Ren Z, Wu L, Wang J, Guo J, Zheng L, Ming Z, Zhang L, Lou Z, et al. 2019.
899 Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus
900 Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* 47:6538–6550.

901 Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983. Comparison of

902 simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

903 Jo S, Kim T, Iyer VG, Im W. 2008. CHARMM-GUI: a web-based graphical user interface for
904 CHARMM. *J. Comput. Chem.* 29:1859–1865.

905 Jumper J, Tunyasuvunakool K, Kohli P, Hassabis D, AlphaFold Team. Computational
906 predictions of protein structures associated with COVID-19. *Deep Mind* [Internet].
907 Available from: https://deepmind.com/research/open-source/computational-predictions-of-
908 protein-structures-associated-with-COVID-19

909 Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple
910 sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

911 Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. 2020. The Architecture of SARS-CoV-
912 2 Transcriptome. *Cell* 181:914–921.e10.

913 Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi
914 EE, Bhattacharya T, Foley B, et al. 2020. Tracking Changes in SARS-CoV-2 Spike:
915 Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* 182:812–827.e19.

916 Lauring AS, Hodcroft EB. 2021. Genetic Variants of SARS-CoV-2-What Do They Mean? *JAMA*
917 325:529–531.

918 Leigh JW, Bryant D. 2015. popart: full-feature software for haplotype network construction.
919 *Methods Ecol. Evol.* 6:1110–1116.

920 Lindahl, Abraham, Hess, Spoel V der. 2020. GROMACS 2020 Source code. Available from:
921 https://zenodo.org/record/3562495

922 Liu Y, Liu J, Plante KS, Plante JA, Xie X, Zhang X, Ku Z, An Z, Scharton D, Schindewolf C, et
923 al. 2021. The N501Y spike substitution enhances SARS-CoV-2 transmission. *bioRxiv*
924 [Internet]. Available from: http://dx.doi.org/10.1101/2021.03.08.434499

925 Li Y, Ma M-L, Lei Q, Wang F, Hong W, Lai D-Y, Hou H, Xu Z-W, Zhang B, Chen H, et al.
926 2021. Linear epitope landscape of the SARS-CoV-2 Spike protein constructed from 1,051
927 COVID-19 patients. *Cell Rep.* 34:108915.

928 Luan B, Wang H, Huynh T. Molecular Mechanism of the N501Y Mutation for Enhanced
929 Binding between SARS-CoV-2's Spike Protein and Human ACE2 Receptor. Available
930 from: http://dx.doi.org/10.1101/2021.01.04.425316

931 Lu J, Li B, Deng A, Li K, Hu Y, Li Z, Xiong Q, Liu Z, Guo Q, Zou L, et al. Viral infection and
932 transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant.
933 Available from: http://dx.doi.org/10.21203/rs.3.rs-738164/v1

934 Martin DP, Weaver S, Tegally H, San EJ, Shank SD, Wilkinson E, Giandhari J, Naidoo S, Pillay
935 Y, Singh L, et al. 2021. The emergence and ongoing convergent evolution of the N501Y
936 lineages coincides with a major global shift in the SARS-CoV-2 selective landscape.
937 *medRxiv* [Internet]. Available from: http://dx.doi.org/10.1101/2021.02.23.21252268

938 Muller HJ. 1964. THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE.
939 *Mutat. Res.* 106:2–9.

940 Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R,
941 van der Spek A, Tolsma P, Rietveld A, Brouwer M, et al. 2021. Transmission of SARS-
942 CoV-2 on mink farms between humans and mink and back to humans. *Science* 371:172–

943 177.

944 Papa G, Mallery DL, Albecka A, Welch L, Cattin-Ortolá J, Luptak J, Paul D, McMahon HT,
945 Goodfellow IG, Carter A, et al. Furin cleavage of SARS-CoV-2 Spike promotes but is not
946 essential for infection and cell-cell fusion. Available from:
947 http://dx.doi.org/10.1101/2020.08.13.243303

948 Parrinello M, Rahman A. 1981. Polymorphic transitions in single crystals: A new molecular
949 dynamics method. *J. Appl. Phys.* 52:7182–7190.

950 Prates E, Garvin M, Jones P, Miller JI, Kyle S, Cliff A, Gazolla JGFM, Shah M, Walker A, Lane
951 M, et al. 2021. Antiviral Strategies Against SARS-CoV-2 – For a Bioinformatics Approach.
952 In: Hann JJ, Bintou A, Keng C, editors. SARS-CoV-2 Methods and Protocols. Spriinger.

953 Prates ET, Garvin MR, Pavicic M, Jones P, Shah M, Demerdash O, Amos BK, Geiger A,
954 Jacobson D. 2020. Potential pathogenicity determinants identified from structural
955 proteomics of SARS-CoV and SARS-CoV-2. *Mol. Biol. Evol.* [Internet]. Available from:
956 http://dx.doi.org/10.1093/molbev/msaa231

957 Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B. 2017.
958 Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods*
959 14:411–413.

960 Richard M, Kok A, de Meulder D, Bestebroer TM, Lamers MM, Okba NMA, Fentener van
961 Vlissingen M, Rockx B, Haagmans BL, Koopmans MPG, et al. 2020. SARS-CoV-2 is
962 transmitted via contact and via the air between ferrets. *Nat. Commun.* 11:3496.

963 Rodrigues CHM, Pires DEV, Ascher DB. 2021. DynaMut2: Assessing changes in stability and
964 flexibility upon single and multiple point missense mutations. *Protein Sci.* 30:60–69.

965 Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed
966 models. *Bioinformatics* 19:1572–1574.

967 Sabino EC, Buss LF, Carvalho MPS, Prete CA Jr, Crispim MAE, Fraiji NA, Pereira RHM, Parag
968 KV, da Silva Peixoto P, Kraemer MUG, et al. 2021. Resurgence of COVID-19 in Manaus,
969 Brazil, despite high seroprevalence. *Lancet* [Internet]. Available from:
970 http://dx.doi.org/10.1016/S0140-6736(21)00183-5

971 Santerre M, Arjona SP, Allen CN, Shcherbik N, Sawaya BE. 2020. Why do SARS-CoV-2 NSPs
972 rush to the ER? *J. Neurol.* [Internet]. Available from: http://dx.doi.org/10.1007/s00415-020-
973 10197-8

974 Sawatzki K, Hill NJ, Puryear WB, Foss AD, Stone JJ, Runstadler JA. 2021. Host barriers to
975 SARS-CoV-2 demonstrated by ferrets in a high-exposure domestic setting. *Proc. Natl.*
976 *Acad. Sci. U. S. A.* [Internet] 118. Available from:
977 http://dx.doi.org/10.1073/pnas.2025601118

978 Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, Geng Q, Auerbach A, Li F. 2020. Structural
979 basis of receptor recognition by SARS-CoV-2. *Nature* 581:221–224.

980 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B,
981 Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular
982 interaction networks. *Genome Res.* 13:2498–2504.

983 Sheikh A, McMenamin J, Taylor B, Robertson C. 2021. SARS-CoV-2 Delta VOC in Scotland:
984 demographics, risk of hospital admission, and vaccine effectiveness. *The Lancet* [Internet].
985 Available from: http://dx.doi.org/10.1016/s0140-6736(21)01358-1

986 Simon-Loriere E, Holmes EC. 2011. Why do RNA viruses recombine? *Nat. Rev. Microbiol.*
987 9:617–626.

988 Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA
989 cytosine methylation using nanopore sequencing. *Nat. Methods* 14:407–410.

990 Singh A, Steinkellner G, Köchl K, Gruber K, Gruber CC. Serine 477 plays a crucial role in the
991 interaction of the SARS-CoV-2 spike protein with the human receptor ACE2. Available
992 from: http://dx.doi.org/10.21203/rs.3.rs-106969/v1

993 Singh J, Rahman SA, Ehtesham NZ, Hira S, Hasnain SE. 2021. SARS-CoV-2 variants of
994 concern are emerging in India. *Nat. Med.* [Internet]. Available from:
995 http://dx.doi.org/10.1038/s41591-021-01397-4

996 Starr TN, Greaney AJ, Addetia A, Hannon WW, Choudhary MC, Dingens AS, Li JZ, Bloom JD.
997 2021. Prospective mapping of viral mutations that escape antibodies used to treat COVID-
998 19. *Science* [Internet]. Available from: http://dx.doi.org/10.1126/science.abf9302

999 Starr TN, Greaney AJ, Hilton SK, Crawford KHD, Navarro MJ, Bowen JE, Alejandra Tortorici
1000 M, Walls AC, Veesler D, Bloom JD. Deep mutational scanning of SARS-CoV-2 receptor
1001 binding domain reveals constraints on folding and ACE2 binding. Available from:
1002 http://dx.doi.org/10.1101/2020.06.17.157982

1003 Thorne LG, Bouhaddou M, Reuschl A-K, Zuliani-Alvarez L, Polacco B, Pelin A, Batra J,
1004 Whelan MVX, Ummadi M, Rojc A, et al. 2021. Evolution of enhanced innate immune
1005 evasion by the SARS-CoV-2 B.1.1.7 UK variant. *bioRxiv* [Internet]. Available from:
1006 http://dx.doi.org/10.1101/2021.06.06.446826

1007 Tylor S, Andonov A, Cutts T, Cao J, Grudesky E, Van Domselaar G, Li X, He R. 2009. The SR-
1008 rich motif in SARS-CoV nucleocapsid protein is important for virus replication. *Canadian*
1009 *Journal of Microbiology* [Internet] 55:254–260. Available from:
1010 http://dx.doi.org/10.1139/w08-139

1011 Velasco JD. 2013. Phylogeny as population history. *Philosophy and Theory in Biology* [Internet]
1012 5. Available from: http://dx.doi.org/10.3998/ptb.6959004.0005.002

1013 Verba K, Gupta M, Azumaya C, Moritz M, Pourmal S, Diallo A, Merz G, Jang G, Bouhaddou
1014 M, Fossati A, et al. 2021. CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a
1015 multifunctional protein involved in key host processes. *Res Sq* [Internet]. Available from:
1016 http://dx.doi.org/10.21203/rs.3.rs-515215/v1

1017 Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, Hinsley WR, Laydon DJ,
1018 Dabrera G, O'Toole Á, et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England:
1019 Insights from linking epidemiological and genetic data. Available from:
1020 http://dx.doi.org/10.1101/2020.12.30.20249034

1021 Wang Z, Schmidt F, Weisblum Y, Muecksch F, Barnes CO, Finkin S, Schaefer-Babajew D,
1022 Cipolla M, Gaebler C, Lieberman JA, et al. 2021. mRNA vaccine-elicited antibodies to
1023 SARS-CoV-2 and circulating variants. *Nature* [Internet]. Available from:

1024 http://dx.doi.org/10.1038/s41586-021-03324-6

1025 Washington NL, Gangavarapu K, Zeller M, Bolze A, Cirulli ET, Schiabor Barrett KM, Larsen
1026 BB, Anderson C, White S, Cassens T, et al. 2021. Emergence and rapid transmission of
1027 SARS-CoV-2 B.1.1.7 in the United States. *Cell* 184:2587–2594.e7.

1028 Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, Graham BS, McLellan JS.
1029 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*
1030 367:1260–1263.

1031 Xia H, Cao Z, Xie X, Zhang X, Chen JY-C, Wang H, Menachery VD, Rajsbaum R, Shi P-Y.
1032 2020. Evasion of Type I Interferon by SARS-CoV-2. *Cell Rep.* 33:108234.

1033 Yan L, Zhang Y, Ge J, Zheng L, Gao Y, Wang T, Jia Z, Wang H, Huang Y, Li M, et al. 2020.
1034 Architecture of a SARS-CoV-2 mini replication and transcription complex. *Nat. Commun.*
1035 11:5874.

1036 Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. 2020. Structural basis for the recognition of
1037 SARS-CoV-2 by full-length human ACE2. *Science* 367:1444–1448.

1038 Zhang B-Z, Hu Y-F, Chen L-L, Yau T, Tong Y-G, Hu J-C, Cai J-P, Chan K-H, Dou Y, Deng J,
1039 et al. 2020. Mining of epitopes on spike protein of SARS-CoV-2 from COVID-19 patients.
1040 *Cell Res.* 30:702–704.

1041 Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, Rangarajan ES, Pan A,
1042 Vanderheiden A, Suthar MS, et al. 2020. SARS-CoV-2 spike-protein D614G mutation
1043 increases virion spike density and infectivity. *Nat. Commun.* 11:6013.

1044

1045 **Supplementary Material**

1046

1047 **The emergence of highly fit SARS-CoV-2 variants accelerated by recombination**

1048 Michael R. Garvin[1,2+*], Erica T. Prates[1,2+], Jonathon Romero[3], Ashley Cliff [3], Joao Gabriel Felipe

1049 Machado Gazolla[1,2], Monica Pickholz[4,5], Mirko Pavicic[1,2], Daniel Jacobson[1,2,*]

1050

1051 **Affiliations:**

1052 [1]Oak Ridge National Laboratory, Computational Systems Biology, Biosciences, Oak Ridge, TN; [2]National Virtual

1053 Biotechnology Laboratory, US Department of Energy; [3]The Bredesen Center for Interdisciplinary Research and

1054 Graduate Education, University of Tennessee Knoxville, Knoxville, TN; [4]Departamento de Física, Facultad de

1055 Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina; [5] Instituto de Física de

1056 Buenos Aires (IFIBA), CONICET-Universidad de Buenos Aires, Buenos Aires, Argentina.

1057

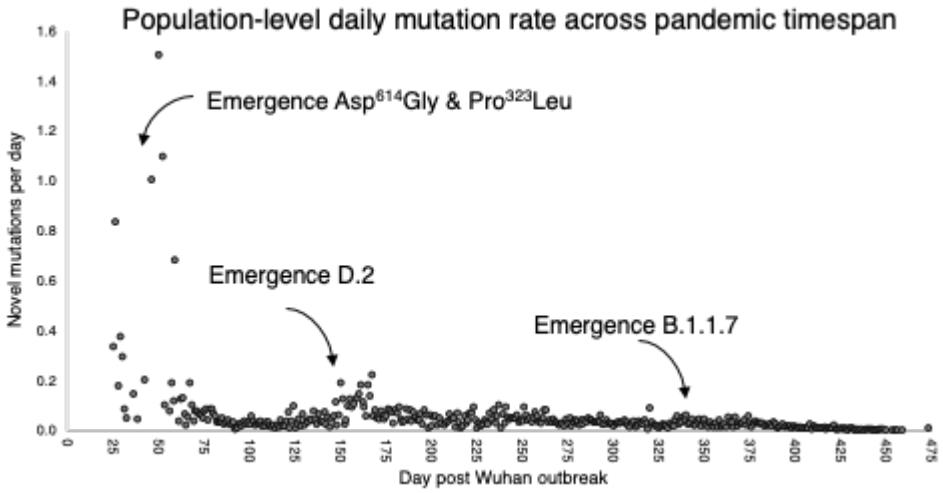
1058 ***Correspondence:** garvinmr@ornl.gov, jacobsonda@ornl.gov

1059 [+]**Contributed equally**

1060

1061 **1. Supplementary Figures**

1062



1063

1064 **Fig. S1.**

1065 **Population level mutation rate over the course of the pandemic.** Number of novel mutations sampled across the
1066 globe for each day are plotted against time (days from the Wuhan outbreak). Emergence of major VOC are
1067 provided for context and show small increases in the number of new mutations but there is an overall decrease
1068 across time, even accounting for multiple mutations at a site to different nucleotide states and deletions.

1069

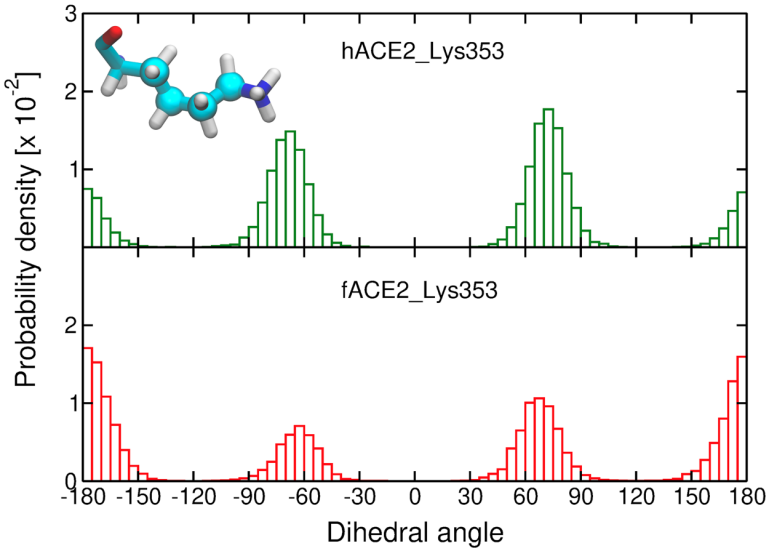


Fig. S2.

The probability density of the conformations of Lys^{353} in human and ferret ACE2 in the simulations. Histograms of the distribution of a dihedral angle of the Lys^{353} side chain carbon atoms in human ACE2 (hACE2, upper figure) and ferret ACE2 (fACE2, lower figure) in complex with the SARS-CoV-2 S receptor-binding domain. The atoms forming the selected dihedral are depicted as spheres in the molecular representation of Lys^{353}. Three independent simulations are considered for the calculation of the histograms. Dihedral angles near ±180º correspond to a more stretched conformation (i.e., *trans*).

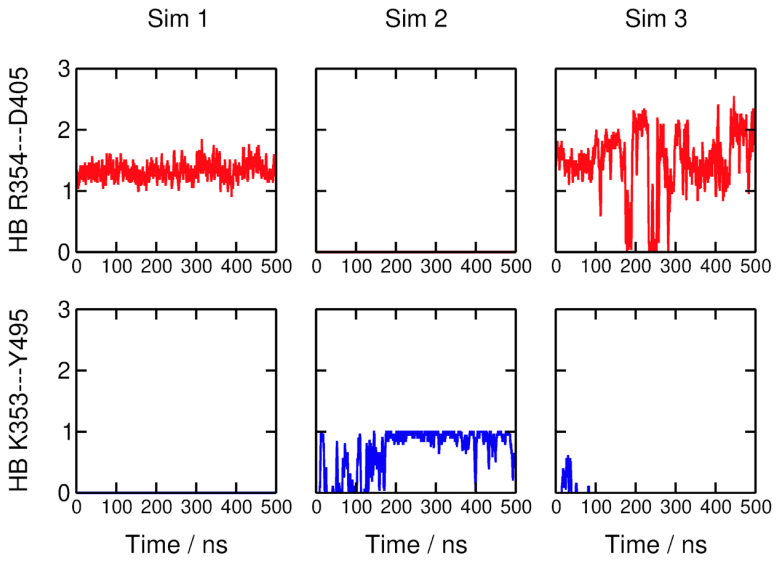


Fig. S3.

Competing hydrogen bond interactions formed between positively charged amino acid residues in ferret ACE2 (fACE2) and the SARS-CoV-2 S receptor-binding domain. Time evolution of the number of hydrogen bonds (HB) that fACE2 Arg3^{54} and Lys^{353} form with Asp^{405} and Tyr^{495} from the SARS-CoV-2 S receptor-binding domain. The columns correspond to the three simulation replicas. The geometric criteria adopted for hydrogen bonds are a cutoff of 3.0 Å for donor-acceptor distance and 20° for acceptor-donor-H angle.

1093 **2. Supplementary Tables**

1094

1095 **Table S4.**

1096 **Average number of contacts formed between Asn^{501} in the receptor-binding domains of SARS-CoV-2 S and**
1097 **residues in ACE2 from human (hACE2) and ferret (fACE2).** A distance of 4 Å between any atom pairs was
1098 defined as the cut-off for contact statistics.

1099

ACE2 residue	hACE2|	fACE2
Tyr^{41}	0.96 ± 0.02	0.80 ± 0.03
Lys^{353}	0.99 ± 0.01	0.90 ± 0.01
Asp^{353}	0.98 ± 0.01	0.70 ± 0.04

1100

1101

1102