Evolutionary and phenotypic characterization of spike mutations in a new SARS-CoV-2 Lineage reveals two Variants of Interest

Paula Ruiz-Rodriguez[1#], Clara Francés-Gómez[1#], Álvaro Chiner-Oms[2], Mariana G. López[2], Santiago Jiménez-Serrano[2], Irving Cancino-Muñoz[2], Paula Ruiz-Hueso[3], Manuela Torres-Puente[2], Maria Alma Bracho[1,4], Giuseppe D'Auria[1,3,4], Llúcia Martinez-Priego[3], Manuel Guerreiro[5,6], Marta Montero-Alonso[7], María Dolores Gómez[8], José Luis Piñana[9], SeqCOVID-SPAIN consortium, Fernando González-Candelas[1,4], Iñaki Comas[2,4], Alberto Marina[2,10], Ron Geller[1]*, Mireia Coscolla[1]*

1. I^{2}SysBio, University of Valencia-CSIC, FISABIO Joint Research Unit Infection and Public Health, Valencia, Spain.

2. Instituto de Biomedicina de Valencia (IBV-CSIC), Valencia, Spain.

3. Sequencing and Bioinformatics Service of FISABIO (Valencian Region Foundation for the Promotion of Health and Biomedical Research), Valencia, Spain.

4. CIBER in Epidemiology and Public Health, Spain.

5. Hematology Department, Hospital Universitari i Politècnic la Fe, Valencia, Spain

6. Instituto de Investigación Sanitaria La Fe (IIS-La Fe), Valencia, Spain.

7. Infectious Diseases Unit, Hospital Universitari i Politècnic la Fe, Valencia, Spain.

8. Microbiology Department, Hospital Universitari i Politècnic La Fe, Valencia, Spain

9. Hematology Service, Hospital Clínico Universitario, Institute for Research INCLIVA, Valencia, Spain.

10. CIBER in Rare Diseases (CIBERER), Spain.

#equal contributions

*corresponding author

Keywords: SARS-CoV-2, Spike, HR2, variants, homoplasy, and antibody escape.

Abstract

Molecular epidemiology of SARS-CoV-2 aims to monitor the appearance of new variants with the potential to change the virulence or transmissibility of the virus. During the first year of SARS-CoV-2 evolution, numerous variants with possible public health impact have emerged. We have detected two mutations in the Spike protein at amino acid positions 1163 and 1167 that have appeared independently multiple times in different genetic backgrounds, indicating they may increase viral fitness. Interestingly, the majority of these sequences appear in transmission clusters, with the genotype encoding mutations at both positions increasing in frequency more than single-site mutants. This genetic outcome that we denote as Lineage B.1.177.637, belongs to clade 20E and includes 12 additional single nucleotide polymorphisms but no deletions with respect to the reference genome (first sequence in Wuhan). B.1.177.637 appeared after the first wave of the epidemic in Spain, and subsequently spread to eight additional countries, increasing in frequency among sequences in public databases. Positions 1163 and 1167 of the Spike protein are situated in the HR2 domain, which is implicated in the fusion of the host and viral membranes. To better understand the effect of these mutations on the virus, we examined whether B.1.177.637 altered infectivity, thermal stability, or antibody sensitivity. Unexpectedly, we observed reduced infectivity of this variant relative to the ancestral 20E variant *in vitro* while the levels of viral RNA in nasopharyngeal swabs did not vary significantly. In addition, we found the mutations do not impact thermal stability or antibody susceptibility in vaccinated individuals but display a moderate reduction in sensitivity to neutralization by convalescent sera from early

stages of the pandemic. Altogether, this lineage could be considered a Variant of Interest (VOI), we denote VOI1163.7. Finally, we detected a sub-cluster of sequences within VOI1163.7 that have acquired two additional changes previously associated with antibody escape and it could be identified as VOI1163.7.V2. Overall, we have detected the spread of a new Spike variant that may be advantageous to the virus and whose continuous transmission poses risks by the acquisition of additional mutations that could affect pre-existing immunity.

Introduction

Genomic surveillance of viral mutations is the first step in detecting viral changes that could impact public health by interfering with diagnostics, modifying pathogenicity, or altering susceptibility to existing immunity or treatments. In many countries, the challenge of detecting new mutations of interest in SARS-CoV-2 is approached by sequencing representative genomes from circulating viruses, sharing sequence information on public databases (e.g. GISAID[1]), and analysing them in real-time using platforms, such as Nextstrain[2]. While mutations appear randomly, their fate in the population depends on a combination of the conferred fitness advantage as well as stochastic and demographic processes. A first step in assessing the potential public health impact of mutations is to decipher if their increase in frequency is due to chance or adaptation. If found to be adaptive, it is important to evaluate whether their adaptation is linked to an improved ability to replicate, colonize, transmit, or evade antiviral hosts defences[3]. An important challenge in the field is to decipher which of all the variants that appear should be monitored to implement measures to mitigate their risk to public health. Genotypes that are phenotypically different from a reference isolate or have mutations that lead to changes associated with either established or suspected phenotypes could be considered Variants of Interest (VOI) if they also fit one of the following criteria: i) cause community transmission/multiple COVID-19 clusters or ii) have been detected in multiple countries[4]. Among VOI, only those genotypes that are associated with higher transmissibility, with detrimental

changes in COVID-19 epidemiology, with increased virulence, with changes to clinical presentation, or with decrease effectiveness of public health measures, diagnostics, vaccines or therapeutics are further categorized as variants of concern[4].

Mutations in SARS-CoV-2 have been reported since the early stages of the epidemic[5-7]. While no signs of recombination have been detected so far among SARS-CoV-2 variants[8], the most common mutations described are single nucleotide polymorphisms (SNPs) and small deletions[9-11]. Genomic surveillance of mutations has been mostly focused on the Spike (S) protein because of its key roles in viral entry and immunity[12], as well as the fact that this protein constitutes the basis of numerous SARS-CoV-2 vaccines[13]. S is a homotrimeric protein, whose heavily glycosylated ectodomain protrudes from the viral membrane, showing a bat-like shape with a N-terminal globular head portion connected to the membrane by an elongated stalk[14]. The S protein is proteolytically processed by the cellular furin protease into the S1 and S2 subunits[15,16]. Additional proteolytic cleavage occurs following S protein binding to the host receptors, facilitating S1 subunit release. The C-terminal S2 subunit remains trimeric in the viral membrane but undergoes conformational changes that promote viral membrane fusion with the host cell[17]. A key role in these conformational changes is played by two heptad repeat motifs, HR1 and HR2 that, starting from the head and stalk regions in the pre-fusion state of S protein, form a HR1-HR2 six-helix bundle in the post-fusion state that is critical for viral entry[18].

The first mutation that was identified as of potential concern was an aspartic acid to a glycine mutation in the S1 subunit of the S protein at position 614 (D614G). D614G emerged early in the epidemic, became predominant in most countries within 2 months, and completely dominated the epidemic by August 2020[19]. As with any mutant, the initial spread of this mutation could have resulted from stochastic events, the dynamics of epidemic, or an intrinsically higher viral fitness. More than six months after the initial report of this mutation, several studies have found evidence in favour of higher transmission efficacy in animal models and human populations[5,20-

[22]. This variant replicates better in some cell culture and animal models[20,21,23], and is associated with higher viral loads in infected individuals[19]; importantly, however, it does not impact diagnostics or vaccine efficacy.

Following the first wave of the pandemic, additional variants have been reported from many countries. Among the first of these was the A222V mutation at the N-terminal domain (NTD) of the S1 subunit, which occurred in the background of the D614G S protein mutation. This variant, termed 20E, was first sequenced in Spain and expanded throughout Europe[6]. Other variants have been reported since, such as the so-called "cluster 5", which harbours a combination of 3 SNPs and single deletion related to mink farms in Denmark[24]. One of the SNPs is in the S protein of this variant, Y453F, occurs in the receptor binding domain (RBD) and may increase binding to cell receptors in mink[25]. Transmission of this variant between humans and minks has been reported[7], highlighting a possible risk of expansion in the human population, which resulted in proposals for large scale culling of mink populations in Denmark. Studies to assess the biological impact of this mutant have not been reported but there is no evidence for its wide spread over the course of >6 months since its description[26]. By December 2020, three variants of concern (VOCs) were described, all of which share the N501Y amino acid replacement in the RBD of the S protein: 20I/501Y.V1 (also called Lineage B.1.1.7) was originally described in the UK[11], 20H/501Y.V2 (B.1.351) in South Africa, and 20J/501Y.V3 (P1) in Brazil. These variants are of particular concern because they are more transmissible[27-29] and, although data on antigenicity and disease severity is not conclusive, could result in reduced susceptibility to neutralization by existing immunity and affect vaccine efficacy[30-33]. Importantly, all of these variants have spread outside of the country where they were initially identified and are estimated to spread faster than other co-circulating genotypes[27,34,35].

In this work, we have performed a detailed phylogenomic analysis of the appearance, spread and evolution of two mutations involving amino acid positions 1163 and 1167 of the HR2

functional motif of S protein. Our results provide evidence of repeated, independent emergence, suggesting these mutations contribute to increased viral fitness. In addition, we have evaluated the biological relevance of these mutations to viral infectivity, virion stability, and neutralization by sera from convalescent and vaccinated individuals.

Results

Multiple and independent mutations in amino acid positions 1163 and 1167 of the Spike protein.

SARS-CoV-2 genetic variation has been monitored by the Spanish sequencing consortium SeqCOVID to follow the expansion of mutations that could potentially result in a change of the biological properties of the virus. We focused on mutations in the S protein because of its relevance for infection and immunity[12]. We detected two mutations in the *spike* gene: G25049T (D1163Y) and G25062T (G1167V), which appeared in Spain as early as March and April 2020, respectively (Supplementary Fig.1). These mutations continued arising independently of each other and, by the end of June, when the predominant circulating genotypes from the first wave in Spain had already been replaced by other variants[36], were also observed together (Supplementary Fig.1). Both positions have mutated multiple times independently and to different amino acids at a lower frequency. On the one hand, D1163 appears mutated at least 99 times (D1163Y: 84, D1163V: 4, D1163G: 3, D1163A: 2, D1163E: 2, D1163H: 2, D1163N: 1, and D1163H/Y: 1) in 47 lineages according to the PANGO scheme[37]. On the other hand, G1167 appears mutated at least 54 times (G1167V: 39, G1167D: 4, G1167C: 3, G1167R: 3, G1167S: 3, G1167F: 1, and G1167A: 1) in 20 PANGO lineages including B.1 (Supplementary Fig.2e) and its derivatives B.26, B.40 (Supplementary Fig.2c), and D.2 (Supplementary Fig.2f).

Clusters of transmission with amino acid changes in positions 1163 and 1167 of Spike

Positions 1163 and 1167 in the S protein have mutated independently multiple times in SARS-CoV-2. The majority of mutated sequences (94.43%) were found in transmission clusters (see methods for definition of clusters; Figure 1a,b), with a small minority not belonging to a transmission cluster due to either incomplete sampling or failure to spread. While different amino acids changes have been detected at both positions, only one change at each position appeared in most clusters: D1163Y in 83.33% and G1167V in 69.23% clusters. D1163Y appeared in 22 transmission clusters (Figure 1a) and G1167V in 8 clusters (Figure 1b). Interestingly, the biggest cluster included both the D1163Y and G1167V mutations together. We denote this cluster as B.1.177.637, which was detected in 65 sequences from Spain until December 2020, representing 1.17% of the Spanish sequences. Globally, this cluster, which is within lineage 20E (also described as 20A.EU1[6], and B.1.177[37]), includes 1,627 sequences (Figure 1c). B.1.177.637 is characterized by nine nonsynonymous and six synonymous mutations with respect to the reference sequence from Wuhan (Supplementary table 2), but no deletions were shared among B.1.177.637 sequences. Amino acid changes were found in A222V, D614G, D1163Y, and G1167V in the S protein, A220V and P365S in the N protein, V30L in ORF10, L67F in ORF14, and P4715L in ORF1ab (Supplementary Fig.3 and Supplementary table 2). Synonymous mutations were also observed in the *orf1ab*, *n* and *m* genes (Supplementary Fig.3 and Supplementary table 2). All evidence support we consider this cluster a new Lineage, and we requested to be under the name B.1.177.637 (https://github.com/cov-lineages/pango-designation/issues/22).

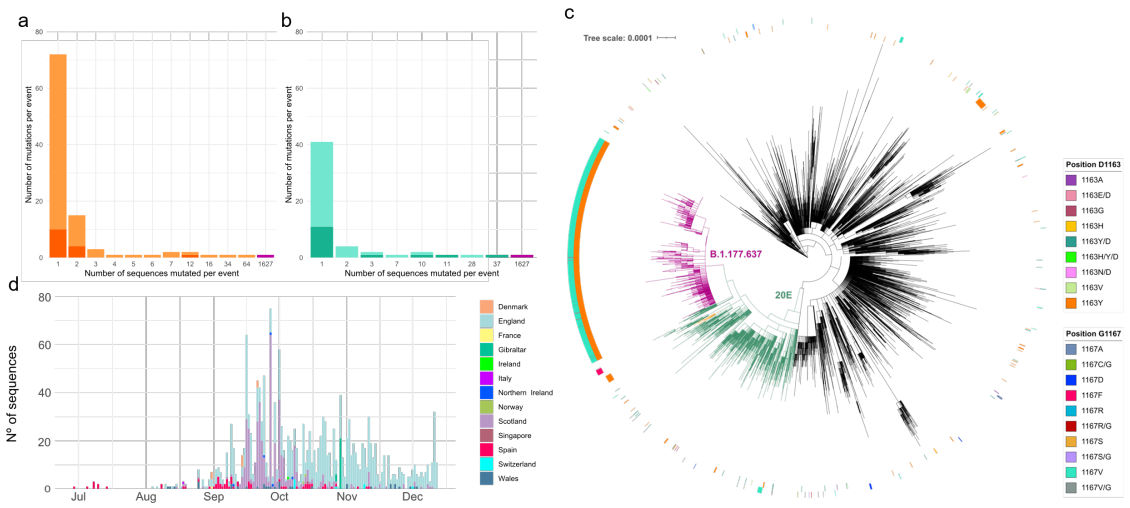


Figure 1. Sequences mutated at positions 1163 and 1167 of the S protein. a. The number of mutation events for amino acid replacement D1163Y (light orange) or another D1163 amino acid replacements (dark orange). **b**. The number of mutation events for amino acid replacement G1167V (light turquoise) or another G1167 amino acid replacements (dark turquoise). Bars coloured in magenta indicate the appearance of both the D1163Y and G1167V amino acid replacements in the same sequences. **c.** Maximum-likelihood phylogeny of 10,450 SARS-CoV-2 genomes. The inner circle of the rings represents sequences with amino acid changes in position D1163 of the S protein. The external circle represents sequences with amino acid changes in position G1167 of the S protein. Branches are coloured in magenta for B.1.177.637, green for clade 20E, and orange for cluster 1163.654. The scale bar indicates the number of nucleotide substitutions per site. **d.** Temporal distribution and frequency of sequences with variant B.1.177.637 coloured by geographical origin.

Within 20E, the second largest cluster including any of these mutations was observed in 34 sequences with E654Q and D1163Y in S protein plus 7 nonsynonymous and 6 synonymous mutations (Supplementary table 2 and Supplementary Fig.3). We denote this second cluster, which is also embedded within lineage 20E, cluster 1163.654 (Figures 1c and S3). Cluster 1163.654 appeared first in Ireland on 2020-07-23 and subsequently appeared in Spain and

England. However, after three months, cluster 1163.654 is no longer being detected. A large cluster within Lineage 20E is formed by 37 sequences with the mutation G1167F. Sequences for this cluster were obtained in Wales between the end of October and the beginning of November 2020 (dark pink in the external circle in Supplementary Fig.2g).

We found additional clusters involving mutations in position 1163 of the S protein; the largest cluster is composed of 64 sequences within Lineage B.1. The majority of which are from Denmark but also includes sequences from England and Sweden (orange in external ring in Supplementary Fig.2e). Among other smaller clusters, a cluster of 10 sequences with G1167V (within Lineage B.1) appeared in Spain in March (indicated in cyan within the external circle of Supplementary Fig.2e). The cluster was detected in two regions in Spain (Valencia, and Galicia), but was controlled with lockdown measures imposed in Spain from March to May and was not detected after May 2020. The same mutation was found in a cluster of 28 sequences within 20E from England and Wales between October and November 2020 (indicated in cyan within the external circle of Supplementary Fig.2g). Similarly, another cluster of 11 sequences in Lineage B.1.1 and with the mutation G1167A was detected in early January 2021, encompassing sequences from Ecuador, Colombia, and Peru (Supplementary Fig.2f). Within Lineage B.53, we found a cluster of 12 sequences from Lithuania with mutation D1163Y. Finally, we found small clusters with mutations in 1163 or 1167 within Lineage B.40 and Lineage A (Supplementary Fig.2b, c).

Because of the risk posed by VOC[11,26,35], we examined whether mutations in 1163 and 1167 of the S protein were observed in the three VOC described to date. Indeed, mutations in these two positions were observed in two VOC, 20I/501Y.V1 and 20H/501Y.V2. Mutations involving amino acid positions 1163 and 1167 appeared independently in the background of 20I/501Y.V1 multiple times. Specifically, mutations in D1163 have occurred at least 13 times in 24 sequences,

including four transmission clusters and nine unique sequences (Supplementary Fig.4), while mutations at G1167 were observed in at least five independent sequences. Interestingly, D1163Y and G1167V were observed together in only one individual within 20I/501Y.V1, although they were not fixed (relative frequency of 27% and 17% of the reads with D1163Y and G1167V, respectively; Supplementary table 3). Finally, only two sequences that harbour the amino acid replacement G1167V in the S protein were observed in the genomic background 20H/501Y.V2 (Supplementary table 3).

Evolution of B.1.177.637

We explored the emergence and evolution of B.1.177.637, the largest and most successful cluster involving amino acid changes in positions 1163 and 1167 of the S protein. Lineage B.1.177.637 appeared in Spain in June 2020 in sequences from the Basque Country (Figure 1d, and Supplementary Video 1) and subsequently appeared in individuals from other countries, comprising a total of 1,627 sequences in GISAID (0.60% of 270,869 analysed sequences by 23rd of December 2020) (Figure 1d and Supplementary Video 1). The majority of the B.1.177.637 sequences were obtained from the United Kingdom, including England (1,058), Scotland (419), Wales (34) and Northern Ireland (5), but were also observed in Gibraltar (24 sequences), indicating successful migration and transmission (Supplementary Video 1). Although B.1.177.637 is not well represented in sequences from other countries, it has been found in multiple sequences from Denmark (9), Switzerland (8), Norway (2), and single sequences from Italy, France, Singapore, and Ireland. By the end of 2020, B.1.177.637 was still circulating in Europe (Figure 1d and Supplementary Video 1), and by the end of February 2021 it was represented by 1,923 sequences in GISAID (0.33% of submitted sequences).

Within B.1.177.637, we detected additional SNPs in individual sequences or small groups of sequences. One of these changes is E484K in the RDB of the S protein, a mutation present in three VOCs (20I/501Y.V1, 20H/501Y.V2 and 20J/501Y.V3) that is implicated in increased ACE2

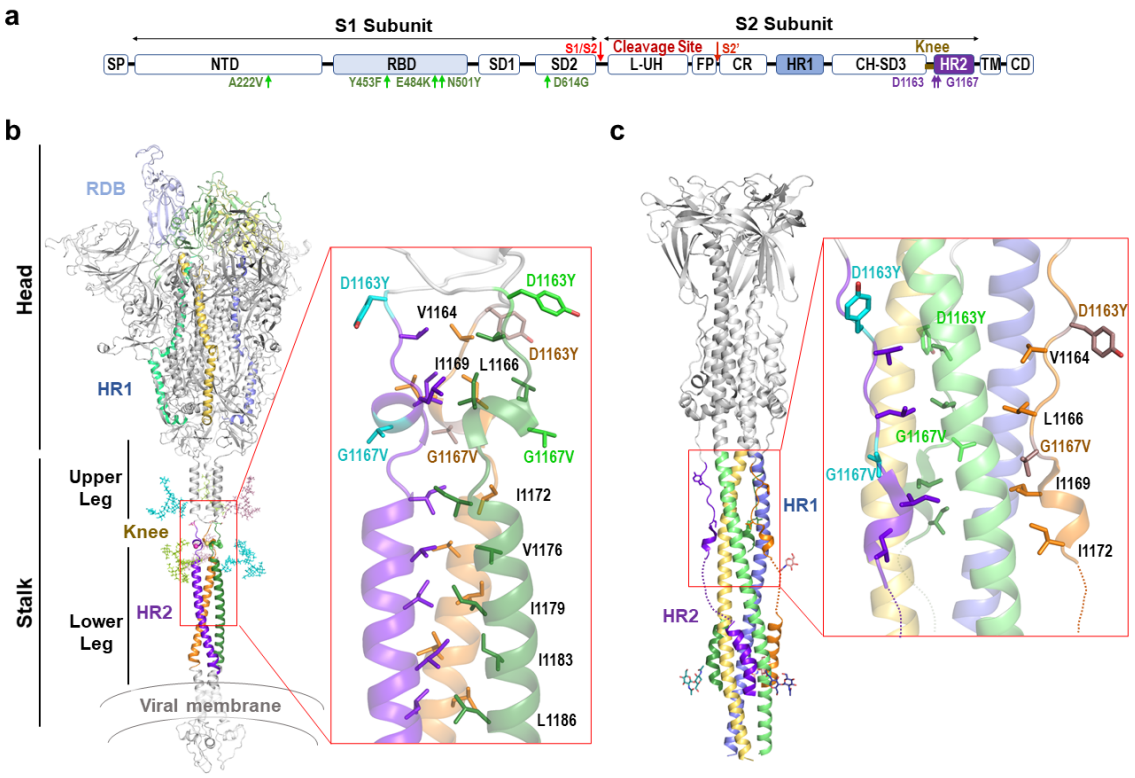
binding[38] and reduced neutralization by antibodies[39]. In addition, we found another change associated with evasion of antibody immunity: a deletion of positions 141-144 in the S protein, which partially overlaps with a smaller deletion at 144 reported in VOC 20I/501Y.V1[40]. This sub-cluster included five sequences along January 2021 from England and Wales (Supplementary table 2). The five sequences formed a monophyletic group embedded in B.1.177.637 (Supplementary Fig.5), identified as cluster B.1.177.637.V2, which displays other nonsynonymous and synonymous mutations (Supplementary table 2) and only two sites are polymorphic within B.1.177.637.V2.

Positions 1163 and 1167 of the S protein are located in the heptad repeat 2 motif

The S protein mediates both the binding to cellular receptors and entry into the host cells[12]. For the former, the RBD motif in the S1 subunit interacts with the cellular receptor in the pre-fusion state. In the post-fusion state, two heptads repeat sequences (HR1 and HR2) in the S2 subunit must form a six-helix bundle in order to bring the viral and cellular membrane into close proximity[41,42] (Figure 2). S protein positions 1163 and 1167 are both located within the HR2 domain. Specifically, 1167 is present at the beginning of the HR2 motif and 1163 in its upstream linker region (Figure 2a). Interestingly, this motif is highly invariable, showing 100% conservation across 14 related sarbecoviruses to which SARS-CoV-2 belongs to (Supplementary table 1)[43,44]. Structural characterization of full-length ectodomain of S protein has shown that the stalk portion encompassing positions 1163 and 1167 presents intrinsic flexibility in the pre-fusion state[18], precluding its atomic visualization. This has been recently confirmed by high-resolution cryo-electron tomographic reconstitution of SARS-CoV-2[14], where this region was observed to constitute a flexible hinge that acts as a “knee”, connecting two helical coiled-coil regions of the stalk (upper and lower legs; Figure 2b). Within this structure, the conformational freedom provided by the glycine residue at position 1163 should play a key role in the flexibility of the knee. In contrast, in the post-fusion state, this region shows high rigidity due to a strong

structural rearrangement of the HR2 motif, which adopts an extended conformation and tightly packs along the central 3-helix bundle stem formed by the HR1 motif (Figure 2c). The resulting HR1-HR2 bundle plays a key role in the mechanism of viral-host membrane fusion[18,45] and mutations in this region could have significant impact on the function of the S protein. In addition, the HR2 region is highly glycosylated, which is observed to be regularly spaced in both the pre- and post-fusion states and to mostly align to the side of the helix bundle[14,18,45]. Of note, two of these branched sugars are placed at positions N1158 and N1173, hiding positions 1163 and 1167 (Figure 2b). Therefore, changes in stalk flexibility might have relevance in immunity by influencing both the intrinsic degree of exposure of this region and its sugar shielding.

Using the available structural information of the S protein in the pre- and post-fusion conformations[18], we examined the possible implications of these mutations to viral infectivity. Based on these structures, G1167V mutation is predicted to confer significant rigidity to the structure in two ways. First, the introduction of a side chain strongly reduces the conformational freedom provided by the glycine residue. Second, the presence of the new aliphatic side chain provided by the valine residue strongly increases hydrophobicity, likely promoting the burial of this side chain in the HR1 helix 3-bundle stem in the post-fusion state or favouring its integration in the neighbour helical coiled-coil in the pre-fusion state (Figure 2b,c). Unlike position 1163, position 1167 is fully exposed to the solvent in both the pre- and post-fusion states (Figure 2b,c). Hence, the effect of D1163Y is likely to stem a change in nature of the side chain, switching from a charged aspartic acid residue at physiological pH to a polar group with hydrophobic properties in the tyrosine.



277

278 **Figure 2**. **The structure of 1163 and 1167 in the pre and post-fusion states of S protein. a.**
279 Schematic representation of the S protein. SP. Signal peptide; NTD, N-terminal domain; RBD,
280 receptor-binding domain; SD1-2, subdomains 1 and 2; L-UH, Linker-Upstream helix; FP, fusion
281 peptide; CR, connecting region; HR1, heptad repeat 1; CH-SD3, central helix subdomain 3; BH,
282 β-hairpin; HR2, heptad repeat 2; TM, transmembrane; CD, cytoplasmic domain. Mutations
283 D1163Y and G1167V are indicated in purple and other mutations described in the text in green.
284 **b.** Cartoon representation (*left*) of a structural model of pre-fusion membrane-bound trimeric S
285 protein[46]. In each subunit the RBD, HR1 and HR2 domains are coloured in different tones (light
286 to dark) of blue, yellow, and green. The N-glycosylation of N1155 and N1176 are shown in stick
287 representation and coloured as the corresponding subunit. Functional and structural regions are
288 marked. A close-view (*right*) of the N-terminal portion of HR2 where D1163Y and G1167V
289 mutations are found. The side-chains of mutated and hydrophobic residues in the HR2 region
290 are shown in stick representation and coloured as corresponding subunit (mutated residues in
291 lighter tone). **c.** Cartoon representation (left) of S2 subunit in post-fusion conformation with HR1

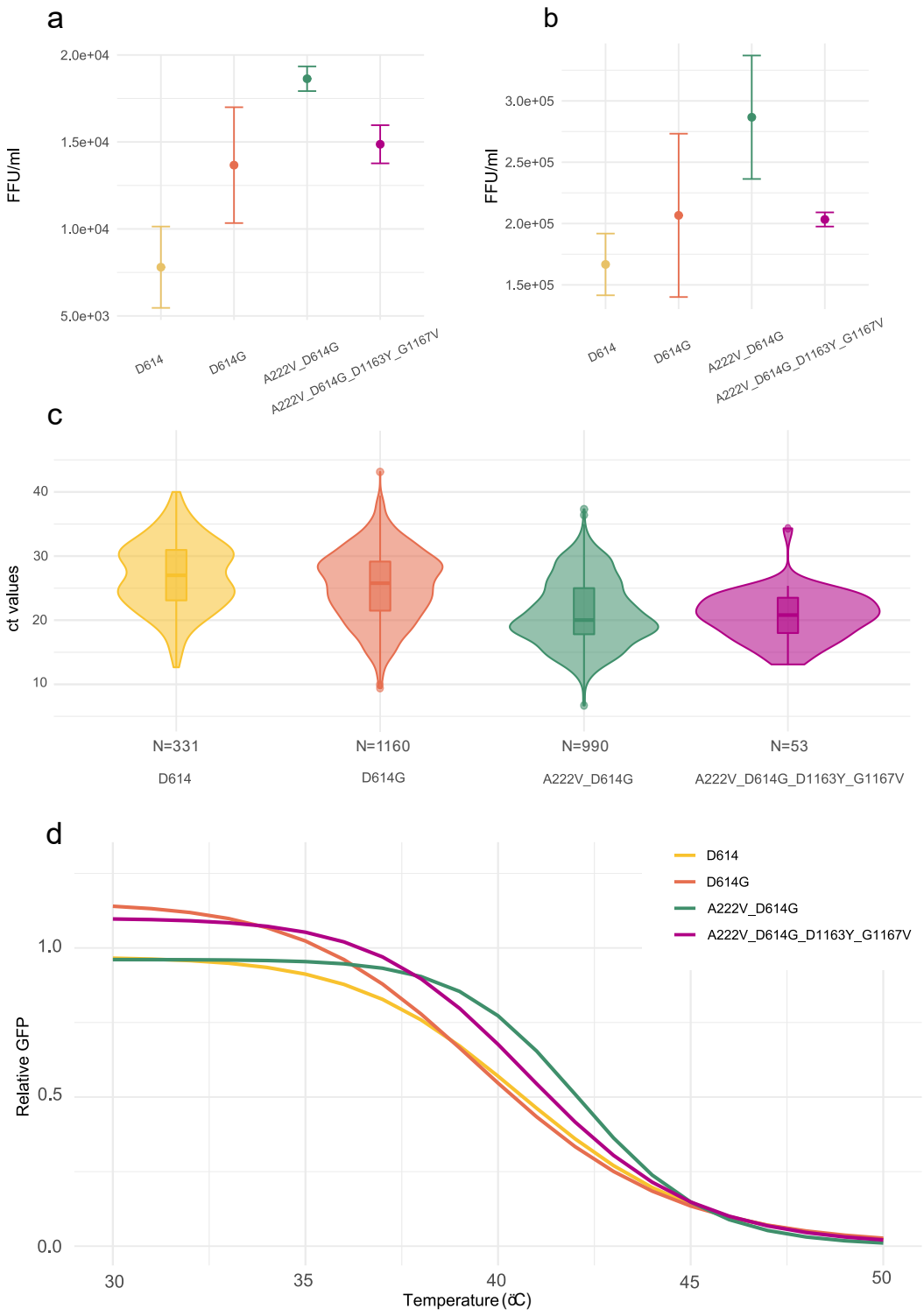
292 and HR2 regions coloured as in b and N-glycosylation around mutation position shown as sticks.
293 A close-view of the region encompassing the mutations (right), showing in stick representation
294 the mutated and hydrophobic residues from the HR2 region shown in panel b. Dotted lines
295 highlight HR2 disordered regions in the Cryo-EM structure.

296 **Spike aminoacid changes D1163Y and G1167V do not increase viral infectivity**

297 Previous reports have indicated that mutations in the S protein can increase infectivity[5,21,47-49].
298 Because the highest transmission success for mutations in S positions 1163 and 1167
299 corresponds to the double mutant D1163Y and G1167V (characteristic of B.1.177.637), we
300 explored whether these mutations in combination have an influence on infectivity. For this, we
301 pseudotyped vesicular stomatitis virus lacking its glycoprotein and encoding GFP[50] (VSVΔG-GFP)
302 with different S genotypes: Wuhan (D614), D614G, 20E (A222V and D614G), or B.1.177.637
303 (A222V, D614G, D1163Y and G1167V). Infectious virus production was then assessed by limiting
304 dilution and counting of GFP-positive cells in both Vero cells and A549 human alveolar basal
305 epithelial cells expressing human ACE2 and TMPRSS2 (A549-hACE2-TMPRSS2). As previously
306 reported[19,21,51], the 20E S genotype enhanced infectivity relative to the Wuhan S genotype by
307 70% in both Vero (p-value = 0.005 by unpaired t-test; Figure 3a) and A549-hACE2-TMPRSS2 cells
308 (p-value = 0.016 by unpaired t-test; Figure 3b). The 20E S genotype also showed a trend towards
309 increased infectivity versus the D614G mutation alone (35% increase in both cell lines), as has
310 been previously reported[49], yet the difference was not statistically significant (p-value > 0.05 by
311 unpaired t-test; Figure 3a,b). In contrast, B.1.177.637 S genotype significantly diminished virus
312 infectivity versus the 20E genotype, reducing virus titers by 20% in Vero cells (p-value = 0.009 by
313 unpaired t-test; Figure 3a) and 29% in A549-hACE2-TMPRSS2 (p-value = 0.03 by unpaired t-test;
314 Figure 3b). This is in agreement with a potential stabilization of the HR2 helix (Figure 2), which
315 should limit the ability of the S protein to sample different structural conformations that could

316 be required for binding host-cell receptors. Hence, B.1.177.637 S genotype does not increase

317 infectivity *in vitro*.



318

Figure 3. Comparison of the infectivity and stability of different S genotypes. a, b. The infectivity of VSV particles pseudotyped with each S protein genotype in either Vero (**a**) or human A549 cells expressing ACE2 and TMPRSS2 (**b**). The mean and standard deviation of three replicates is plotted. **c.** Comparison of cycle threshold (Ct) values for the *n* gene from patients infected with viruses encoding different S protein variants. Data is derived from 2,534 sequences from SeqCOVID consortium. The number of observations (N) analysed for each genotype is indicated. **d.** The thermal sensitivity of VSV pseudotyped with different S genotypes following incubation at 15 minutes. Data are standardized to the surviving fraction following incubation at 30°C, and the three-parameter log-logistic equation is plotted. FFU: focus forming units.

To examine whether reduced infectivity could also be observed *in vivo*, we tested if individuals infected with B.1.177.637 have different viral loads. For this, we used the cycle threshold (Ct) of real-time PCR used for diagnosis as a surrogate. As previously reported[19], we detected higher Ct values for D614 wild-type variant (Ct mean = 27.00) compared to genotypes encoding the D614G S protein mutation (Ct mean = 25.32; p-value < 0.01 by unpaired Wilcoxon test, Figure 3c). However, we did not find significant differences in viral loads between individuals infected with B.1.177.637 genotype and other genotypes within 20E (Ct mean = 21.14 vs Ct mean = 20.63, p-value = 0.72 by unpaired Wilcoxon test, Figure 3c). Interestingly, higher viral loads were observed in individuals infected with B.1.177.637 and others 20E (D614G and A222V) compared to the D614G alone (D614G Ct mean = 25.32, 20E Ct mean = 21.14, B.1.177.637 Ct mean = 20.63, p-value < 0.01 for both comparisons by unpaired Wilcoxon test, Figure 3c). This data suggests that, unlike the results of the in vitro studies, B.1.177.637 replicates as efficiently as other 20E genotypes *in vivo*, although mutations outside of the S protein could contribute to this result.

Amino acid changes D1163Y and G1167V do not alter S protein stability

As increased S protein stability could impact transmissibility by maintaining virion infectivity during the intra-host transmission period, we assessed the temperature sensitivity of the

different S variants. For this, we subjected VSV particles pseudotyped with different S genotypes to a range of temperatures for 15 minutes, after which we evaluated the surviving fraction. Overall, no major differences in the degree to which the different S proteins lost infectivity upon heat exposure were observed, with all S proteins showing a 50% reduction in infectivity at a similar temperature range (39.8-42.2°C; p-value > 0.05 for all except Wuhan S genotype (D614) versus 20E S genotype (A222V and D614G), where p-value = 0.01; Figure 3d). Hence, the D1163Y and G1167V mutations do not seem to have a major impact in the thermal stability of the S protein.

Spike D1163Y and G1167V modestly reduce sensitivity to neutralization by existing antibody immunity

Positions 1163 and 1167 of the S protein have been reported to occur in both T and B cell SARS-CoV-2 epitopes[52-54]. Moreover, numerous studies have shown that mutations in the S protein can affect antibody neutralization[30,31]. We therefore examined if the presence of D1163Y and G1167V alters the neutralization capacity of convalescent sera using VSV pseudotyped with either the 20E or B.1.177.637 S genotypes. In order to capture the potential influence of different infecting variants on antibody neutralization, we tested the sensitivity of these pseudotyped viruses to neutralization by sera from early (April 2020; First wave in Spain) or later (October 2020; Second wave in Spain) in the pandemic, when newer variants were dominant[6,36]. Overall, B.1.177.637 genotype conferred a modest but statistically significant reduction in sensitivity to neutralization by six serum samples tested from the early stage of the pandemic, as measured by the titers required to inhibit viral entry by 80% (ID_{80}; mean = 6.75, range: 1.30-17.68; p-value = 0.008 by paired t-test; Figure 4a). A statistically significant but smaller effect was observed when the titers required to inhibit viral entry by 50% were examined (ID_{50}; mean = 2.27, range: 1.61-3.54; p-value < 0.001 by paired t-test; Supplementary Fig.6). In contrast, both 20E and B.1.177.637 were equally susceptible to sera from patients infected during the second

wave (ID_{80}; mean = 1.03, range: 0.87-1.23; p-value = 0.83 by paired t-test; Figure 4b). These results indicate that the D1163Y and G1167V mutations can provide some degree of escape from pre-existing antibody-based immunity relative to the 20E S genotype depending on the genomic background of the infecting genotype. As a modest reduction in titers was observed with sera from early in the pandemic (Figure 4a), which is more closely related to the current S genotype present in approved vaccines[55,56], we examined if B.1.177.637 S genotype resulted in reduced neutralization by sera from donors vaccinated with the BNT162b2 vaccine. No significant differences in susceptibility to antibody neutralization from vaccinated donors were observed between the two genotypes, indicating that VOI1163.7 is unlikely to alter the efficacy of vaccines based on the Wuhan S genotype (Figure 4c).

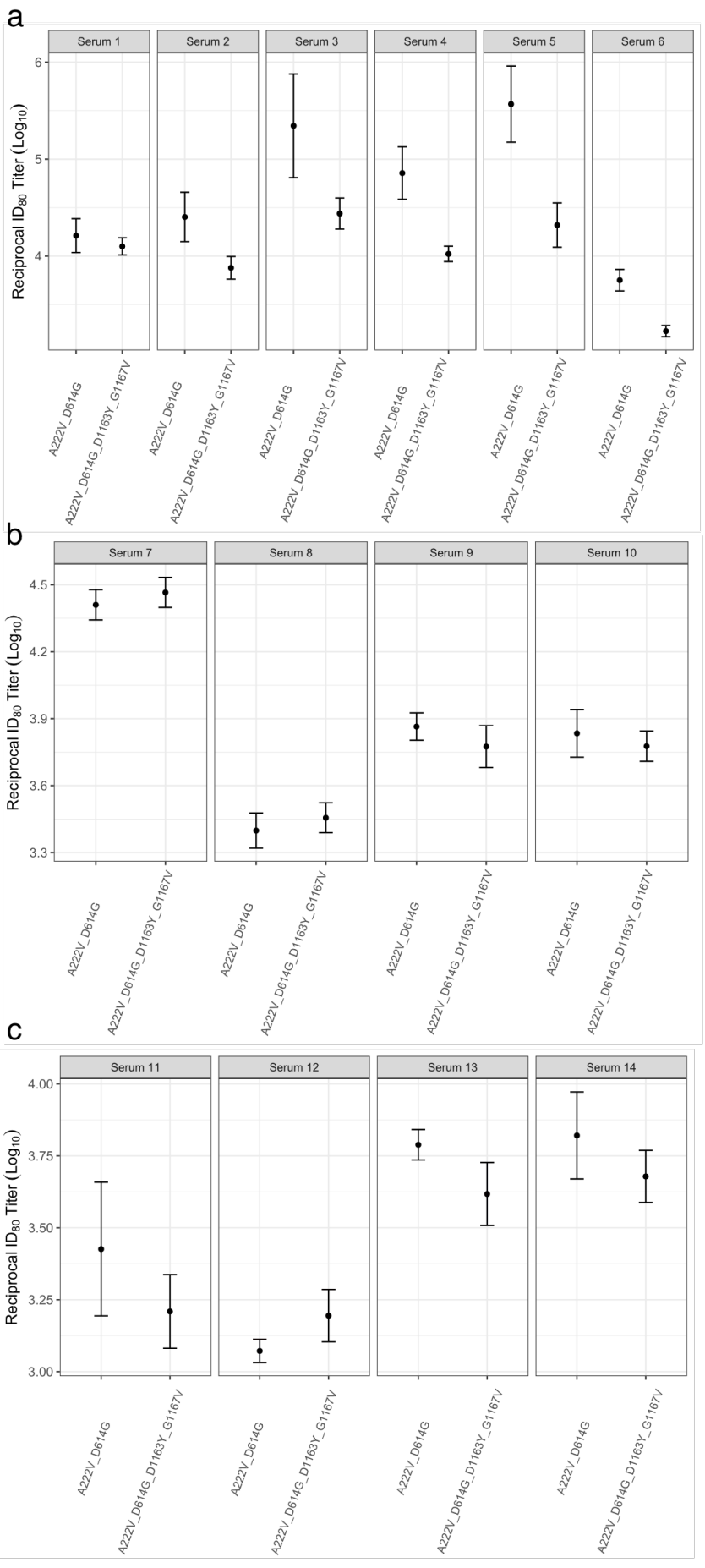

a
Serum 1
Serum 2
Serum 3
Serum 4
Serum 5
Serum 6
Reciprocal ID80 Titer (Log10)
6
5
4
A222V_D614G
A222V_D614G_D1163Y_G1167V
b
Serum 7
Serum 8
Serum 9
Serum 10
4.5
4.2
3.9
3.6
3.3
c
Serum 11
Serum 12
Serum 13
Serum 14
4.00
3.75
3.50
3.25
3.00

Figure 5. Antibody neutralization of 20E and B.1.177.637 variants. The reciprocal titer at which infection with the 20E S genotype (A222V and D614G) or B.1.177.637 S genotype (20E plus D1163Y and G1167V) is reduced by 80% (ID_{80}) by sera from individuals infected during the early stage of the pandemic in (**a**) or during a later stage of the pandemic (**b**) or from donors vaccinated with the BNT162b2 vaccine (**c**). The mean and standard error of three replicates is plotted.

Discussion

SARS-CoV-2 success is linked to its ability to infect and be transmitted. Mutations that emerge independently several times and increase in frequency are likely to confer enhanced viral infectivity, transmission, or immune evasion. The identification of such mutants is of great importance, as they can significantly impact public health. In this work, we have identified two mutations in the S protein that are likely to be beneficial for the virus based on several lines of evidence. First, these mutations are highly variable within SARS-CoV-2 but conserved across the closely related coronaviruses. Second, the vast majority of sequences harbouring these mutations appear in clusters (Figure 1a and 1b). Third, the largest cluster, and therefore the most successful in terms of transmission, includes both mutations together (Figure 1a and 1b). Additionally, both positions have been reported as positively selected multiple times throughout the SARS-CoV-2 phylogeny indicating a fitness advantage[57]. Although either mutation in isolation could be advantageous, their co-occurrence in a large cluster that has been sustained for more than six months across Europe is suggestive of increased fitness when both mutations are present together.

Positions 1163 and 1167 are found in the HR2 domain of the S protein, adjacent to the transmembrane domain. Examination of available structural data suggests that G1167V might alter the flexibility of the S protein stalk by both restricting the conformational freedom normally conferred by the glycine residue and by introducing a hydrophobic side chain that will favour

burial in the HR2 coiled-coil leucine zipper of the pre-fusion state (Figure 2). This extensive flexibility of S prefusion stalk seems to be unique to the SARS-CoV-2 S protein and has not been reported for other class I fusion proteins[18]. The stalk flexibility has been suggested to increase avidity for the host receptors by allowing the engagement of multiple S proteins[18]. Therefore, stalk stabilization is likely to result in a reduced ability of S to bind receptors in the target cell. Indeed, we find the B.1.177.637 genotype to have reduced infectivity compared to the 20E genotype in both Vero and A549-hACE2-TMPRSS2 cells (Figure 3a, b). In contrast, viral load in individuals infected with different S genotypes indicated that 20E and B.1.177.637 reach similar degrees of viral replication in vivo (Figure 3c). This apparent discrepancy in the effect of the two mutations on infectivity may stem from the differences of SARS-CoV-2 infection *in vivo* and in the *in vitro* assay. A recent publication has suggested that the tyrosine-protein kinase receptor UFO (AXL) is an important mediator of SARS-CoV-2 entry and may be of higher relevance for infection of the lung than ACE2[58]. The effect of entry via AXL in the two cell lines used may be overshadowed by high levels of ACE2 expression. Alternatively, additional factors could underlie this difference, including the presence of additional mutations outside of the S protein or differences of viral loads across sampling times i*n vivo* during the infection.

Increased temperature stability can potentially confer a fitness advantage to the virus by reducing losses to infectivity during environmental transition between hosts. Hence, we also examined whether these mutations altered the temperature stability of the virions. Overall, no major difference in stability was observed between VSV pseudoparticles bearing the D614, D614G, 20E, or B.1.177.637 S protein genotypes (Figure 3d). Hence, changes in protein stability are unlikely to underlie the increased transmission of these variants.

Finally, as the S protein is a major target of the immune response[12], immune evasion represents one possible consequence of mutations in this protein. Both S positions 1163 and 1167 are embedded in experimentally confirmed T cell and B cell epitopes. For T cell epitopes, a predicted

HLA-II epitopes including position 1163 and 1167 has been experimentally verified to bind to HLA DRB1*01:01, the prototype molecule for the DR supertype (epitope identifier in Immune Epitope Data Base: 9006[59]). Additionally, D1163 is included in a SARS-CoV-2 T cell epitope eliciting T-cell responses in convalescent COVID-19 cases[60] as well as in SARS-CoV-2-naïve individuals[53], indicating cross-reactivity in epitopes involving these regions. B cell linear epitopes that span D1163 and G1167 have also been reported[52], with D1163 belonging to a dominant linear B cell epitope recognized by more than 40% COVID-19 patients used in the assay[54]. D1163 is fully solvent exposed in available structure[18,45], making its side-chain easily accessible to antibodies, providing a potential mechanism for altering antibody binding. To directly examine whether the mutated S positions 1163 and 1167 influence susceptibility to pre-existing humoral immunity, we examined the neutralization capacity of convalescent sera against the 20E variant or B.1.177.637. For this, we used sera from both the first (April 2020) and second (October 2020) waves of the infection in Spain, because an almost complete replacement of SARS-CoV-2 S variants occurred between these two times of the pandemic in Spain[36]. When utilizing sera from donors infected during the first wave of the pandemic in Spain, we found a modest but statistically significant reduction in susceptibility to neutralization of the B.1.177.637 S genotype compared to the 20E S genotype of approximately 6-fold (Figure 4a). However, no difference in neutralization was observed between the two variants when sera from patients infected during the second wave was used (Figure 4b), highlighting variant-specific differences in antibody responses. Overall, the magnitude of the observed reduction in neutralization susceptibility to sera from individuals infected during the first wave was much less pronounced than that observed for other genotypes implicated in immune evasion[31]. Nevertheless, the degree of reduced neutralization required to confer a biologically relevant fitness advantage *in vivo* has not been established, and even relatively small reductions in susceptibility to antibody neutralizations could potentially confer a significant advantage to replication. Indeed, this has been suggested to be the case in an immunosuppressed individual treated with convalescent

serum, where mutations selected during the course of treatment conferred a similar reduction to that observed in the current study[61]. Finally, no evidence was found for reduced neutralization by sera from donors immunized with the BNT162b2 vaccine (Figure 4c), which is based on the Wuhan S genotype, indicating antibody immunity elicited by Pfizer-BioNTech COVID-19 vaccine (BNT162b2; February 2021) would not be affected by amino acid replacements in 1163 and 1167 sites in S protein.

Although further experiments are needed to decipher the mechanisms by which the two S mutations identified could confer a selective advantage to the virus, the evidence presented supports B.1.177.637 as a VOI (VOI1163.7.V1) according to recently published criteria[4]. First, amino acid replacements in S protein: D1163Y and G1167V lead to changes associated with suspected phenotypic change because of the rigidity it poses to the S protein. Second, the genotype showed moderate but significantly lower antibody susceptibility compared to 20E S genotype. And third, it has been identified to cause community transmission, appearing in multiple COVID-19 cases, and detected in multiple countries. Additionally, a subgroup within B.1.177.637 (denoted as VOI1163.7.V2) includes two additional mutations leading to amino acid changes in S with established phenotypic impact on humoral immunity: E484K[62] and 141-144Del[40,63].

Whether VOI1163.7.V1 and VOI1163.7.V2 will continue to increase in frequency and accumulate additional mutations that could improve its fitness and/or present challenges to vaccines or diagnostics remains to be seen. However, their characterization as VOI would help to discover if enough evidence holds to consider them VOC and therefore required monitoring.

Methods

Whole-genome sequencing and genome assembly of SeqCOVID consortium sequences

A total of 5,017 clinical samples were received, sequenced, and analysed by the SeqCOVID consortium from all autonomous communities of Spain. These samples were confirmed as SARS-CoV-2 positive by RT-PCR carried out by Clinical Microbiology Services from each hospital. Sequencing of the samples has been approved by the ethics committee: Comité Ético de Investigación de Salud Pública y Centro Superior de Investigación en Salud Pública (CEI DGSP-CSISP) Nº 20200414/05. All sequences are available at GISAID under the accession numbers detailed in Supplementary table 4.

For sequencing, RNA samples were retro-transcribed into cDNA. SARS-CoV-2 complete genome amplification was performed in two multiplex PCR, according to the protocol developed by the ARTIC network[64], using the V3 multiplex primers scheme[65]. From this step, two amplicon pools were prepared, combined, and used for library preparation. The genomic libraries were constructed with the Nextera DNA Flex Sample Preparation kit (Illumina Inc., San Diego, CA) according to the manufacturer's protocol, with 5 cycles for indexing PCR. Whole-genome sequencing was performed in the MiSeq platform (2×200 cycles paired-end run; Illumina).

Reads obtained were processed through a bioinformatic pipeline based on iVar[66], available at https://gitlab.com/fisabio-ngs/sars-cov2-mapping. The first step in the pipeline removed human reads with Kraken[67]; then fastq files were filtered using fastp[68] v 0.20.1 (arguments employed: --cut tail, --cut-window-size, --cut-mean-quality, -max_len1, -max_len2). Finally, mapping and variant calling were performed with iVar v 1.2, and quality control assessment was carried out with MultiQC[69].

Analysis of the *spike* gene of sarbecoviruses related to SARS-CoV-2

14 sequences including SARS-COV-2 belonging to sarbecoviruses, sequences were annotated with annotation files available at NCBI database in order to locate the *spike* gene coordinates (accession numbers are available at supplemental table 1). For each sequence, nucleotide coordinates belonging to this gene were extracted with EMBOSS[70]. The 14 sequences harbouring the *spike* gene were concatenated and aligned with MEGA-X[71] using amino acids with ClustalW algorithm with default options (alignment is available at https://github.com/PathoGenOmics/B.1.177.637_SARS-CoV-2)

Sampling SARS-CoV-2 from non-Spanish consortium sequences

To build the global alignment, sequences were downloaded from GISAID[1] including all the pandemic periods since the first known case sequenced (from 24 December 2019) until the last sample on 22 December 2020. We used two filters to select the dataset: sequences with more than 29,000 bp, and sequences with known dates of sampling. Sequences downloaded from GISAID were aligned against the SARS-CoV-2 reference genome[72] using MAFFT[73], omitting all insertions and getting an alignment length of 29,903 bp. The final alignment constructed included 270,869 sequences, all sequences with GISAID ID used for this study are available in Supplementary Table 5.

Frequency and detection of mutated positions

Single nucleotide variants were detected using the global dataset alignment, generating a VCF file with SNP-sites[74] v 2.5.1 (argument employed: -v), using the reference genome as the reference bases for detecting mutations. This VCF file was processed with a Python script (available at https://github.com/PathoGenOmics/B.1.177.637_SARS-CoV-2) to assess all mutated samples by position, calculating the frequencies of the global dataset and annotating sequences with the detected mutations. After that, the mutated positions were annotated with

snpEff[75] v 5.0 using SARS-CoV-2 reference (Wuhan first sequenced) database annotation (arguments employed: -c, -noStats, -no-downstream, -no-upstream, NC_045512.2).

Genotypes detected that involved mutations in 1163 and 1167 such as B.1.177.637 and cluster 163.654 were represented in a circos plot with the R package circlize[76] v 0.4.12.1004. Nucleotide coordinates of SARS-CoV-2 are plotted in a non-closed circle, circle is annotated and coloured with genes of the virus, and mutated nucleotide positions of each genotype are connected between them through lines.

Alignments

For the phylogenetic analysis, a reduced dataset was selected from the 270,869 sequences. Duplicated sequences were removed with seqkit v 0.13.2 (arguments employed: rmdup -s). 8,397 sequences were selected at random with the same temporal distribution by month as the initial dataset by Python scripting (available at https://github.com/PathoGenOmics/B.1.177.637_SARS-CoV-2). The 8,397 sequences were concatenated with 2,053 sequences selected as indicate above because harboured amino acid replacements in D1163 and G1167 of the S protein and resulted in an alignment of 10,450 sequences (Supplementary Table 6).

The dataset to represent 20I/501Y.V1 phylogenetic relationships include 3,067 randomly selected samples identified by Pangolin typing system (https://github.com/cov-lineages/pangolin) as lineage B.1.1.7 plus the 33 sequences with amino acid replacements in D1163 and/or G1167 (Supplementary Table 7).

For all the alignments, problematic positions reported by Lanfear, R.[77] were masked for the phylogenetic reconstruction using masked_alignment.sh script.

Phylogenetic analysis

Maximum-likelihood phylogenies in Figure 1 and supplementary Supplementary Fig.2, S4 and S5 were reconstructed from the masked alignment using IQ-TREE[78] v 1.6.12 with GTR model and collapsing near-zero branches (arguments employed: -czb, -m GTR). The phylogenies were rooted in the reference sequence from Wuhan[72] on 2019-12-24. The phylogenies were annotated and visualized with iTOL[79] v 4.

The phylogeny in Supplementary Video S1, composed by 10,450 sequences, was build up with Nextstrain pipeline (available at https://github.com/nextstrain/augur) in order to monitor and visualize temporal and geographical transmission of B.1.177.637. This dataset file is available at https://github.com/PathoGenOmics/B.1.177.637_SARS-CoV-2.

Clusters of transmission involving 1163 and 1167 S amino acid replacements.

We used the phylogeny of 10,450 sequences enriched with all sequences mutated in 1163 and 1167 to quantify the minimum number of mutational events involving positions 1163 and 1167 in S protein. We first defined which mutations characterize internal nodes using R packages: tidytree v 0.3.3 and treeio v 1.14.3[80]. We then depicted monophyletic clusters sharing at least one of the two mutations. Transmission clusters were defined as all sequences that: i) are derived from an internal node characterized by the same nucleotide mutation involving 1163 or 1167 amino acid replacements, ii) include more than one sequence, and iii) at least 95% of sequences share the nucleotide mutation. Additionally, redundant nodes were eliminated, keeping the ancestral node of the cluster (harbouring the largest number of leaves). Sequences with at least one mutation but not in clusters were counted as single events of mutation in the phylogeny.

Structural analysis of 1163 and 1167 S amino acid replacements.

The atomic coordinates for S protein in pre-fusion state was retrieved from the CHARMM-GUI COVID-19 Archive (http://www.charmm-gui.org/docs/archive/covid19). The atomic coordinates for S protein in post-fusion sate were retrieved from Protein Data Bank (PDB: 6XRA[18] and PDB: 6LXT[81]). Mutations were introduced using single mutation tool embedded in COOT[82]and figures were generated with PyMOL (www.pymol.org).

SARS-CoV-2 pseudotyped vesicular stomatitis virus production, titration, and thermal stability evaluation

Mutations were introduced into a plasmid encoding a codon-optimized S protein[16] by site directed mutagenesis (see Supplementary table8 for primers). All mutations were verified by Sanger sequencing (see Supplementary table 9 for primers). To evaluate the efficiency of virus production, three transfections in HEK293 cells (CRL-1573 from ATCC) were performed for each plasmid to generate pseudotyped VSV harbouring the indicated S protein[83]. The titers of the virus produced were then assayed by serial dilution, followed by infection of either Vero cells (CCL-81 from ATCC) or A549 cells expressing ACE2 and TMPRSS2 (InvivoGen catalog code a549-hace2tpsa), and counting of GFP positive cells (focus forming units; FFU) at 16 hours post infection. Statistical comparisons were performed by unpaired t-test (R package: stats v 3.6.1) with normalized logarithmic data. For assessing thermal stability, 1000 FFU (as measured on Vero cells) were incubated for 15 minutes at 30.4, 31.4, 33, 35.2, 38.2, 44.8, 47, 48.6 or 49.6°C before addition to Vero cells previously seeded in a 96 well plate (10,000 cells/well). GFP signal in each well was determined 16 hours post-infection using an Incucyte S3 (Essen Biosciences). The mean GFP signal observed in several mock-infected wells was subtracted from all infected wells, followed by standardization of the GFP signal to the mean GFP signal from wells incubated at 30.4°C. Finally, a three parameter log-logistic function was fitted to the data using the drc package v 3.0-1 in R (LL.3 function) and the temperature resulting in 50% inhibition calculated using the drc ED function. Statistical differences in the temperature resulting in 50% reduction

of infection was evaluated using the drc EDcomp function. The data and scripts for analysing the temperature resistance of the different mutants is available at https://github.com/PathoGenOmics/B.1.177.637_SARS-CoV-2.

Evaluation of neutralization by convalescent sera and efficacy of virus particle production.

Pseudotyped VSV bearing 20E or B.1.177.637 S variants were evaluated for sensitivity to neutralization by convalescent sera as previously described[83] with slight modifications. Briefly, 16-hours post-infection, GFP signal in each well was determined using an Incucyte S3 (Essen Biosciences). The mean GFP signal observed in several mock-infected wells was subtracted from all infected wells, followed by standardization of the GFP signal in each well infected with antibody-treated virus to that of the mean GFP signal from wells infected with mock-treated virus. Any negative values resulting from background subtraction were arbitrarily assigned a low, non-zero value (10^{-5}). The serum dilutions were then converted to their reciprocal, their logarithm (Log_{10}) was taken, and the dose resulting in 50% (ID50) or 80% (ID80) reduction in GFP signal was calculated in R using the drc package v 3.0-1. A two-parameter log-logistic regression (LL2 function) was used for all samples except when a three-parameter logistic regression provided a significant improvement to fit, as judged by the ANOVA function in the drc package (e.g. $p < 0.05$ following multiple testing correction using the Bonferroni method). The script for calculating the ID50 and ID80 as well as the standardized GFP signal for each condition is available at https://github.com/PathoGenOmics/B.1.177.637_SARS-CoV-2. For the first wave, serum samples and data from patients included in this study were provided by the Consorcio Hospital General de Valencia Biobank, integrated in the Valencian Biobanking Network, and they were processed following standard operating procedures with the appropriate approval of the Ethics and Scientific Committees. All first wave samples were obtained from donors that were admitted to the intensive care unit and were collected during April 2020. For the second wave donors, sera were obtained (October 2020) from severe COVID-19 patients requiring inpatient

617 treatment at Hospital Universitario y Politécnico La Fe de Valencia. Similarly, samples from
618 immunized donors were collected at Hospital Universitario y Politécnico La Fe de Valencia from
619 hospital health-workers, with no previous history of SARS-COV-2 infection, and after receiving a
620 second dose of Pfizer-BioNTech COVID-19 vaccine (BNT162b2; February 2021). All samples from
621 Hospital Universitario y Politécnico La Fe de Valencia were collected after informed written
622 consent and the project has been approved by the ethical committee and institutional review
623 board (registration number 2020-123-1).

624 ## Acknowledgments

625 We want to particularly acknowledge the patients and the Consorcio Hospital General de
626 Valencia Biobank integrated in the Valencian Biobanking Network for their collaboration, as well
627 as the patients and hospital staff at Hospital Universitario y Politécnico La Fe de Valencia. In
628 addition, the authors would like to thank Gert Zimmer (Institute of Virology and Immunology,
629 Mittelhäusern/Switzerland), Stefan Pohlmann, and Markus Hoffmann (German Primate Center,
630 Infection Biology Unit, Goettingen/Germany) for providing the reagents required for the
631 generation of VSV pseudotyped viruses and the codon-optimized S plasmid.

632 Also, we want to acknowledge all the efforts from different laboratories and authorities
633 submitting all possible sequences of SARS-CoV-2 worldwide and making them available on the
634 GISAID platform.

635 ## Funding

636 M.C. and R.G. are supported by Ramón y Cajal program from Ministerio de Ciencia. This work
637 was funded by the Instituto de Salud Carlos III project COV20/00140 and COV20/00437, Spanish
638 National Research Council project CSIC-COV19-021 and CSIC-COVID19-082, and the Generalitat
639 Valenciana (SEJI/2019/011 and Covid_19-SCI).

Action co-financed by the European Union through the Operational Program of the European Regional Development Fund (ERDF) of the Valencian Community 2014-2020.

References

1. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46 (2017).
2. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).
3. Lauring, A.S. & Hodcroft, E.B. Genetic Variants of SARS-CoV-2-What Do They Mean? *Jama* **325**, 529-531 (2021).
4. WHO. COVID-19 Weekly Epidemiological Update 2021-02-25 . (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210225_weekly_epi_update_voc-special-edition.pdf, 2021).
5. Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75.e11 (2020).
6. Hodcroft, E.B. *et al.* Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv*, 2020.10.25.20219063. Preprint at https://www.medrxiv.org/content/10.1101/2020.10.25.20219063v2 (2020).
7. Oude Munnink, B.B. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371**, 172 (2021).
8. Richard, D., Owen, C.J., van Dorp, L. & Balloux, F. No detectable signal for ongoing genetic recombination in SARS-CoV-2. *bioRxiv*, 2020.12.15.422866. Preprint at https://www.biorxiv.org/content/10.1101/2020.12.15.422866v1 (2020).
9. Welkers, M.R.A., Han, A.X., Reusken, C. & Eggink, D. Possible host-adaptation of SARS-CoV-2 due to improved ACE2 receptor binding in mink. *Virus Evol* **7**, veaa094 (2021).

10. Young, B.E. *et al.* Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *Lancet* **396**, 603-611 (2020).
11. Chand, M. *et al.* Investigation of novel SARS-COV-2 variant Variant of Concern 202012/01. (Public Health England, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959438/Technical_Briefing_VOC_SH_NJL2_SH2.pdf, 2020).
12. Walls, A.C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).
13. Salvatori, G. *et al.* SARS-CoV-2 SPIKE PROTEIN: an optimal immunological target for vaccines. *Journal of Translational Medicine* **18**, 222 (2020).
14. Turoňová, B. *et al.* In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science* **370**, 203 (2020).
15. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol Cell* **78**, 779-784.e5 (2020).
16. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
17. Kielian, M. Mechanisms of Virus Membrane Fusion Proteins. *Annu Rev Virol* **1**, 171-89 (2014).
18. Cai, Y. *et al.* Distinct conformational states of SARS-CoV-2 spike protein. *Science* **369**, 1586 (2020).
19. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
20. Hou, Y.J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464-1468 (2020).

21. Plante, J.A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* (2020).
22. Zhou, B. *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* (2021).
23. Daniloski, Z. *et al.* The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *Elife* **10**(2021).
24. Larsen, H.D. *et al.* Preliminary report of an outbreak of SARS-CoV-2 in mink and mink farmers associated with community spread, Denmark, June to November 2020. *Euro Surveill* **26**(2021).
25. Rodrigues, J. *et al.* Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *PLoS Comput Biol* **16**, e1008449 (2020).
26. Hodcroft, E.B. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. (https://covariants.org/variants/, 2021).
27. Leung, K., Shum, M.H., Leung, G.M., Lam, T.T. & Wu, J.T. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Euro Surveill* **26**(2021).
28. Galloway, S. *et al.* Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States, December 29, 2020–January 12, 2021. *MMWR. Morbidity and Mortality Weekly Report* **70**(2021).
29. Davies, N.G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, eabg3055 (2021).
30. Wang, Z. *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature* (2021).
31. Wibmer, C.K. *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *bioRxiv*, 2021.01.18.427166. Preprint at https://www.biorxiv.org/content/10.1101/2021.01.18.427166v2 (2021).

32. Muik, A. *et al.* Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine-elicited human sera. *bioRxiv*, 2021.01.18.426984. Preprint at https://www.biorxiv.org/content/10.1101/2021.01.18.426984v1 (2021).

33. Madhi, S.A. *et al.* Safety and efficacy of the ChAdOx1 nCoV-19 (AZD1222) Covid-19 vaccine against the B.1.351 variant in South Africa. *medRxiv*, 2021.02.10.21251247. Preprint at https://www.medrxiv.org/content/10.1101/2021.02.10.21251247v1 (2021).

34. Washington, N.L. *et al.* Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *medRxiv*, 2021.02.06.21251159. Preprint at https://www.medrxiv.org/content/10.1101/2021.02.06.21251159v1 (2021).

35. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*, 2020.12.21.20248640. Preprint at https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1 (2020).

36. López, M.G. *et al.* The first wave of the Spanish COVID-19 epidemic was associated with early introductions and fast spread of a dominating genetic variant. *medRxiv*, 2020.12.21.20248328. Preprint at https://www.medrxiv.org/content/10.1101/2020.12.21.20248328v1 (2020) (2020).

37. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology* **5**, 1403-1407 (2020).

38. Nelson, G. *et al.* Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. *bioRxiv*, 2021.01.13.426558. Preprint at https://www.biorxiv.org/content/10.1101/2021.01.13.426558v1 (2021).

741 39. Liu, Z. *et al.* Landscape analysis of escape variants identifies SARS-CoV-2 spike
742 mutations that attenuate monoclonal and serum antibody neutralization. *bioRxiv*,
743 2020.11.06.372037. Preprint at
744 https://www.biorxiv.org/content/10.1101/2020.11.06.372037v2 (2021).
745 40. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of
746 vulnerability for SARS-CoV-2. *bioRxiv*, 2021.01.14.426475. Preprint at
747 https://www.biorxiv.org/content/10.1101/2021.01.14.426475v1 (2021).
748 41. Huang, Y., Yang, C., Xu, X.-f., Xu, W. & Liu, S.-w. Structural and functional properties of
749 SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta*
750 *Pharmacologica Sinica* **41**, 1141-1149 (2020).
751 42. Xia, S. *et al.* Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1
752 domain in spike protein. *Cellular & Molecular Immunology* **17**, 765-767 (2020).
753 43. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus:
754 implications for virus origins and receptor binding. *The Lancet* **395**, 565-574 (2020).
755 44. Xia, S. *et al.* Peptide-Based Membrane Fusion Inhibitors Targeting HCoV-229E Spike
756 Protein HR1 and HR2 Domains. *Int J Mol Sci* **19**(2018).
757 45. Fan, X., Cao, D., Kong, L. & Zhang, X. Cryo-EM analysis of the post-fusion structure of
758 the SARS-CoV spike glycoprotein. *Nature Communications* **11**, 3618 (2020).
759 46. Woo, H. *et al.* Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein
760 Model in a Viral Membrane. *The Journal of Physical Chemistry B* **124**, 7128-7137
761 (2020).
762 47. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike
763 density and infectivity. *Nature Communications* **11**, 6013 (2020).
764 48. Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2
765 Spike Protein Variant. *Cell* **183**, 739-751.e8 (2020).

766 49. Wang, Y. *et al.* The Infectivity and Antigenicity of Epidemic SARS-CoV-2 Variants in the
767 United Kingdom. *Research Square* (2021).
768 50. Berger Rentsch, M. & Zimmer, G. A vesicular stomatitis virus replicon-based bioassay
769 for the rapid and sensitive determination of multi-species type I interferon. *PLoS One*
770 **6**, e25858 (2011).
771 51. Weissman, D. *et al.* D614G Spike Mutation Increases SARS CoV-2 Susceptibility to
772 Neutralization. *Cell Host & Microbe* **29**, 23-31.e4 (2021).
773 52. Li, Y. *et al.* Linear epitopes of SARS-CoV-2 spike protein elicit neutralizing antibodies in
774 COVID-19 patients. *Cellular & Molecular Immunology* **17**, 1095-1097 (2020).
775 53. Mateus, J. *et al.* Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed
776 humans. *Science* **370**, 89-94 (2020).
777 54. Yi, Z. *et al.* Functional mapping of B-cell linear epitopes of SARS-CoV-2 in COVID-19
778 convalescent population. *Emerg Microbes Infect* **9**, 1988-1996 (2020).
779 55. Zhao, J. *et al.* COVID-19: Coronavirus Vaccine Development Updates. *Frontiers in*
780 *Immunology* **11**, 3435 (2020).
781 56. Izda, V., Jeffries, M.A. & Sawalha, A.H. COVID-19: A review of therapeutic strategies
782 and vaccine candidates. *Clinical Immunology* **222**, 108634 (2021).
783 57. Pond, S. Natural selection analysis of global SARS-CoV-2/COVID-19 enabled by data
784 from GISAID. (https://observablehq.com/@spond/revised-sars-cov-2-analytics-
785 page?collection=@spond/sars-cov-2).
786 58. Wang, S. *et al.* AXL is a candidate receptor for SARS-CoV-2 that promotes infection of
787 pulmonary and bronchial epithelial cells. *Cell Research* **31**, 126-140 (2021).
788 59. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Research* **43**,
789 D405-D412 (2015).

60. Tarke, A. *et al.* Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *Cell Reports Medicine* **2**, 100204 (2021).

61. Kemp, S.A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* (2021).

62. Collier, D.A. *et al.* SARS-CoV-2 B.1.1.7 sensitivity to mRNA vaccine-elicited, convalescent and monoclonal antibodies. *medRxiv*, 2021.01.19.21249840. Preprint at https://www.medrxiv.org/content/10.1101/2021.01.19.21249840v4 (2021).

63. Wang, P. *et al.* Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7. *bioRxiv*, 2021.01.25.428137. Preprint at https://www.biorxiv.org/content/10.1101/2021.01.25.428137v2 (2021).

64. Quick, J. nCoV-2019 sequencing protocol. (https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w.pdf, 2020).

65. Artic-network/artic-ncov2019. (Artic-network, https://github.com/artic-network/artic-ncov2019, 2020).

66. Grubaugh, N.D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 8 (2019).

67. Wood, D.E. & Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46 (2014).

68. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).

69. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-8 (2016).

70. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends in Genetics* **16**, 276-277 (2000).

71. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549 (2018).
72. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).
73. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-66 (2002).
74. Page, A.J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056 (2016).
75. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92 (2012).
76. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).
77. Lanfear, R. A global phylogeny of SARS-CoV-2 sequences from GISAID. (Zenodo, 2020).
78. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-74 (2015).
79. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**, W256-W259 (2019).
80. Wang, L.G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol Biol Evol* **37**, 599-603 (2020).
81. Xia, S. *et al.* Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Research* **30**, 343-355 (2020).

82. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).

83. Gozalbo-Rovira, R. *et al.* SARS-CoV-2 antibodies, serum inflammatory biomarkers and clinical severity of hospitalized COVID-19 patients. *J Clin Virol* **131**, 104611 (2020).

Author contributions

PRR: Conceptualization, Methodology, Formal Analysis, Investigation, Visualization, Writing Original Draft.

CFG: Conceptualization, Methodology, Formal Analysis, Writing Original Draft.

ACO: Formal Analysis, Review and edit draft.

MGL: Formal Analysis, Review and edit draft.

SJS: Software, Validation, Review and edit draft.

ICM: Methodology, Resources, Review and edit draft.

PRH: Investigation, Review and edit draft.

MTP: Methodology, Resources, Review and edit draft.

MAB: Investigation, Methodology, Review and edit draft.

GD: Methodology, Software, Data Curation, Review and edit draft.

LLM: Methodology, Project Administration, Review and edit draft.

MG: Resources, Review and edit draft.

MMA: Resources, Review and edit draft.

MDG: Resources, Review and edit draft.

JLP: Resources, Review and edit draft.

FG: Funding, Project Administration, Supervision, Review and edit draft.

IC: Funding, Project Administration, Supervision, Review and edit draft.

AM: Formal analysis, Writing Original Draft.

RG: Conceptualization, Methodology, Formal Analysis, Writing Original Draft, Supervision, Funding.

MC: Conceptualization, Methodology, Investigation, Formal Analysis, Writing Original Draft, Supervision, Funding.

List of SeqCOVID consortium members

Iñaki Comas (icomas@ibv.csic.es), Fernando González-Candelas (fernando.gonzalez@uv.es), Galo A. Goig-Serrano (ggoig@ibv.csic.es), Álvaro Chiner-Oms (achiner@ibv.csic.es), Irving Cancino-Muñoz (icancino@ibv.csic.es), Mariana Gabriela López (mglopez@ibv.csic.es), Manoli Torres-Puente (mtorres@ibv.csic.es), Inmaculada Gómez (igomez@ibv.csic.es), Santiago Jiménez-Serrano (sjimenez@ibv.csic.es), Lidia Ruiz-Roldán (lidiarroldan@gmail.com), María Alma Bracho (bracho_alm@gva.es), Neris García-González (neris@uv.es), Llúcia Martínez Priego (martinez_lucpri@gva.es), Inmaculada Galán-Vendrell (galan_inm@gva.es), Paula Ruiz-Hueso (ruiz_pau@gva.es), Griselda De Marco (demarco_gri@gva.es), Mª Loreto Ferrús Abad (ferrus_mlo@gva.es), Sandra Carbó-Ramírez (carbo_sanram@gva.es), Mireia Coscollá (mireia.coscolla@uv.es), Paula Ruiz Rodríguez (ruizro5@alumni.uv.es), Giuseppe D'Auria (dauria_giu@gva.es), Francisco Javier Roig Sena (roig_fco@gva.es), Hermelinda Vanaclocha Luna (vanaclocha_her@gva.es), Isabel San Martín Bastida (isanmartin@rjb.csic.es), Daniel García Souto (danielgarciasouto@gmail.com), Ana Pequeño Valtierra (ana.pequeno@usc.es), Jose M. C. Tubio (jmctubio@gmail.com), Fco. Javier Temes Rodríguez (fjtemes@gmail.com), Jorge Rodríguez-Castro (jorge.rodriguez@usc.es), Martín Santamarina García (martin.santamarina.garcia@usc.es), Nuria Rabella Garcia (nrabella@santpau.cat), Ferrán

Navarro Risueño (fnavarror@santpau.cat), Elisenda Miró Cardona (emiro@santpau.cat), Manuel Rodríguez-Iglesias (manuel.rodrigueziglesias@uca.es), Fátima Galán-Sanchez (fatima.galan@uca.es), Salud Rodríguez-Pallares (salud361@gmail.com), María de Toro (mthernando@riojasalud.es), María Pilar Bea-Escudero (mpbea@riojasalud.es), José Manuel Azcona-Gutiérrez (jmazcona@riojasalud.es), Miriam Blasco-Alberdi (mblasco@riojasalud.es), Alfredo Mayor (alfredo.mayor@isglobal.org), Alberto L. GarcÍa-Basteiro (alberto.garcia-basteiro@isglobal.org), Gemma Moncunill (gemma.moncunill@isglobal.org), Carlota Dobaño (carlota.dobano@isglobal.org), Pau Cisteró (pau.cistero@isglobal.org), Oriol Mitjà (omitja@flsida.org), Camila González-Beiras (cgonzalez@flsida.org), Martí Vall-Mayans (mvall@flsida.org), Marc Corbacho-Monné (mcorbacho@flsida.org), Andrea Alemany (ealemany@gmail.com), Darío García de Viedma (dgviedma2@gmail.com), Laura Pérez-Lago (lperezg00@gmail.com), Marta Herranz (m_herranz01@hotmail.com), Jon Sicilia (jsiciliamambrilla@gmail.com), Pilar Catalán (pcatalan.hgugm@salud.madrid.org), Julia Suárez (julia.suarez@iisgm.com), Patricia Muñoz (pmunoz@hggm.es), Cristina Muñoz-Cuevas (cristina.munozc@salud-juntaex.es), Guadalupe Rodríguez Rodríguez (guadalupe.rodriguez@salud-juntaex.es), Juan Alberola Enguídanos (juan.alberola@uv.es), Jose Miguel Nogueira Coito (Jose.M.Nogueira@uv.es), Juan José Camarena Miñana (juan.camarena@uv.es), Antonio Rezusta López (arezusta@unizar.es), Alexander Tristancho Baró (aitristancho@salud.aragon.es), Ana Milagro Beamonte (amilagro@salud.aragon.es), Nieves Martínez Cameo (nmcameo@gmail.com), Yolanda Gracia Grataloup (ygrataloup@yahoo.es), Elisa Martró (emartro@igtp.cat), Antoni E. Bordoy (aescalas@igtp.cat), Anna Not (anot@igtp.cat), Adrián Antuori (adrian.antuori@gmail.com), Anabel Fernández (afernandezn.ics@gencat.cat), Nona Romaní (nonaromani@gmail.com), Rafael Benito Ruesca (rbenito@unizar.es), Sonia Algarate Cajo (sonialgarate@gmail.com), Jessica Bueno Sancho (jbuenosan@salud.aragon.es), Jose Luis del Pozo (jdelpozo@unav.es), Jose Antonio Boga Riveiro (joseantonio.boga@sespa.es), Cristián Castelló Abietar (crcaab@hotmail.com), Susana Rojo

Alba (ssnrj4@gmail.com), Marta Elena Álvarez Argüelles (martaealvarez@gmail.com), Santiago Melón García (santiago.melon@sespa.es), Maitane Aranzamendi Zaldumbide (maitane.aranzamendizaldumbide@osakidetza.eus), Óscar Martínez Expósito ("OSCAR.MARTINEZEXPOSITO@osakidetza.eus"), Mikel Gallego Rodrigo ("MIKEL.GALLEGORODRIGO@osakidetza.eus"), Maialen Larrea Ayo ("MAIALEN.LARREAAYO@osakidetza.eus"), Nerea Antona Urieta ("NEREA.ANTONAURIETA@osakidetza.eus"), Andrea Vergara Gómez (vergara@clinic.cat), Miguel J Martínez Yoldi (myoldi@clinic.cat), Jordi Vila Estapé (jvila@clinic.cat), Elisa Rubio García (elrubio@clinic.cat), Aida Peiró-Mestres (aida.peiro@isglobal.org), Jessica Navero-Castillejos (jessica.navero@isglobal.org), David Posada (dposada@uvigo.es), Diana Valverde (dianaval@uvigo.es), Nuria Estévez (nuestevez@uvigo.es), Iria Fernández-Silva (irfernandez@uvigo.es), Loretta de Chiara (Ldechiara@uvigo.es), Pilar Gallego (mgallego@alumnos.uvigo.es), Nair Varela (nvarela@alumnos.uvigo.es), Rosario Moreno Muñoz (moreno_rosmuny@gva.es), Mª Dolores Tirado Balaguer (tirado_dolbal@gva.es), Ulises Gómez-Pinedo (ulisesalfonso.gomez@salud.madrid.org), Mónica Gozalo Margüello (monica.gozalo@scsalud.es), Mª Eliecer Cano García (meliecer.cano@scsalud.es), José Manuel Méndez Legaza (josemanuel.mendez@scsalud.es), Jesús Rodríguez Lozano (jesus.rodriguez@scsalud.es), María Siller Ruiz (maria.siller@scsalud.es), Daniel Pablo Marcos (daniel.pablo@scsalud.es), Antonio Oliver (antonio.oliver@ssib.es), Jordi Reina (jorge.reina@ssib.es), Carla López-Causapé (carla.lopez@ssib.es), Andrés Canut Blasco (andres.canutblasco@osakidetza.eus), Silvia Hernáez Crespo (silvia.hernaezcrespo@osakidetza.eus), Mª Luz Cordón Rodríguez (marialuzalbina.cordonrodriguez@osakidetza.eus), Mª Concepción Lecaroz Agara (mariaconcepcion.lecarozagara@osakidetza.eus), Carmen Gómez González (carmen.gomezgonzalez@osakidetza.eus), Amaia Aguirre Quiñonero (amaia.aguirrequiñonero@osakidetza.eus), José Israel López Mirones

(joseisrael.lopezmirones@osakidetza.eus), Marina Fernández Torres (marina.fernandeztorres@osakidetza.eus), Mª Rosario Almela Ferrer (mariadelrosario.almelaferrer@osakidetza.eus), José Antonio Lepe Jiménez (josea.lepe.sspa@juntadeandalucia.es), Verónica González Galán (veronica.gonzalez.galan.sspa@juntadeandalucia.es), Ángel Rodríguez Villodres (angel.rodriguez.villodres.sspa@juntadeandalucia.es), Nieves Gonzalo Jiménez (gonzalo_nie@gva.es), Mª Montserrat Ruiz García (ruiz_mongar@gva.es), Antonio Galiana Cabrera (antoniogaliana1@gmail.com), Judith Sánchez-Almendro (judhsa@gmail.com), Gustavo Cilla Eguiluz (CARLOSGUSTAVOSANTIAGO.CILLAEGUILUZ@osakidetza.eus), Milagrosa Montes Ros (MARIAMILAGROSA.MONTESROS@osakidetza.eus), Luis Piñeiro Vázquez (LUISDARIO.PINEIROVAZQUEZ@osakidetza.eus), Ane Sorrarain (ane.sorarrain@biodonostia.org), José María Marimón Ortiz de Zarate (JOSEMARIA.MARIMONORTIZDEZ@osakidetza.eus), Mª Dolores Gómez Ruiz, (gomez.mdo@gmail.com), Eva González Barberá (gonzalez_evabar@gva.es), José Luis López Hontangas (lopez_jlu@gva.es), José María Navarro-Marí (josem.navarro.sspa@juntadeandalucia.es), Irene Pedrosa Corral (irene.pedrosa.sspa@juntadeandalucia.es), Sara Sanbonmatsu Gámez (saral.sanbonmatsu.sspa@juntadeandalucia.es), M. Carmen Perez Gonzalez (mcpergon@gobiernodecanarias.org), Francisco Javier Chamizo López (fchalop@gobiernodecanarias.org), Ana Bordes Benítez (aborben@gobiernodecanarias.org), David Navarro Ortega (david.navarro@uv.es), Eliseo Albert Vicent (eliseo.al.vi@gmail.com), Ignacio Torres (nachotfink@gmail.com), Mª Isabel Gascón Ros (gascon_isa@gva.es), Cristina Torregrosa Hetland (cjtorregrosahetland@hotmail.com), Eva Pastor Boix (pastor_eva@gva.es), Paloma Cascales Ramos (cascales_pal@gva.es), Begoña Fuster Escrivá (begona.fuster@gmail.com), Concepción Gimeno Cardona (concepcion.gimeno@uv.es), María Dolores Ocete Mochón (ocete_mar@gva.es), Rafael Medina González

(rafa.medina.gonzalez@gmail.com), Julia González Cantó (juliagonzalez1992@hotmail.com), Olalla Martínez Macias (martinez_ola@gva.es), Begoña Palop Borrás (bpalop@hotmail.com), Inmaculada de Toro Peinado (inmadetoro@yahoo.es), Mª Concepción Mediavilla Gradolph (gradolphilla@hotmail.com), Mercedes Pérez Ruiz (mercedes.perez.ruiz.sspa@juntadeandalucia.es), Oscar González-Recio (gonzalez.oscar@inia.es), Mónica Gutiérrez-Rivas (mgrivas9@gmail.com), Encarnación Simarro Córdoba (mesimarro@sescam.jccm.es), Julia Lozano Serra (jlozanos@sescam.jccm.es), Lorena Robles Fonseca (lrobles@sescam.jccm.es), Adolfo de Salazar (adolsalazar@gmail.com), Laura Viñuela (lauravinuelagon@gmail.com), Natalia Chueca (naisses@yahoo.es), Federico García (fegarcia@ugr.es), Cristina Gomez-Camarasa (gomezcamarasa@gmail.com), Ana Carvajal (ana.carvajal@unileon.es), Vicente Martín (vicente.martin@unileon.es), Juan Fregeneda (juan.fregeneda@unileon.es), Antonio J. Molina (ajmolt@unileon.es), Héctor Arguello (hector.arguello@unileon.es), Tania Fernandez-Villa (tferv@unileon.es), Amparo Farga Martí (farga_amp@gva.es), Rocío Falcón (falcon_roc@gva.es), Victoria Domínguez Márquez (m.victoria.dominguez@uv.es), José Javier Costa Alcalde (jose.javier.costa.alcalde@sergas.es), Rocío Trastoy Pena (rocio.trastoy.pena@sergas.es), Gema Barbeito Castiñeiras (gema.barbeito.castineiras@sergas.es), Amparo Coira Nieto (amparo.coira.nieto@sergas.es), María Luisa Pérez del Molino Bernal (maria.luisa.perez.del.molino.bernal@sergas.es), Antonio Aguilera (antonio.aguilera.guirao@sergas.es), Anna M. Planas (anna.planas@iibb.csic.es), Álex Soriano (asoriano@clinic.cat), Israel Fernández-Cádenas (israelcadenas@yahoo.es), Jordi Pérez-Tur (jpereztur@ibv.csic.es), Mª Ángeles Marcos Maeso (mmarcos@clinic.cat), Carmen Ezpeleta Baquedano (cezpeleb@navarra.es), Ana Navascués Ortega (ana.navascues.ortega@navarra.es), Ana Miqueleiz Zapatero (ana.miqueleiz.zapatero@navarra.es), Manuel Segovia Hernández (msegovia@um.es), Antonio Moreno Docón (a.moreno@um.es), Esther Viedma Moreno (ester.viedma@salud.madrid.org), Jesús Mingorance (jesus.mingorance@idipaz.es), Juan Carlos Galán Montemayor (juancarlos.galan@salud.madrid.org), Iván Sanz Muñoz

(isanzm@saludcastillayleon.es), Diana Pérez San José (dianaperezsj@gmail.com), Maria Gil Fortuño (gil_marfor@gva.es), Juan B. Bellido Blasco (bellido_jua@gva.es), Alberto Yagüe Muñoz (yague_alb@gva.es), Noelia Henández Pérez (hernandez_noeper@gva.es), Helena Buj Jordá (helenitabuj@hotmail.com), Óscar Pérez Olaso (perez_oscola@gva.es), Alejandro González Praetorius (agonzalezp@sescam.jccm.es), Aida Esperanza Ramírez Marinero (aramirezm@lrc.cat), Eduardo Padilla León (epadillal@lrc.cat), Alba Vilas Basil (avilasb@lrc.cat), Mireia Canal Aranda (mcanala@lrc.cat), Albert Bernet Sánchez (abernet.lleida.ics@gencat.cat), Alba Bellés Bellés (abelles.lleida.ics@gencat.cat), Eric López González (elopezg.lleida.ics@gencat.cat), Iván Prats Sánchez (iprats@gss.cat), Mercè García González (mgarciag.lleida.ics@gencat.cat), Miguel Martínez Lirola (miguelj.martinez.lirola.sspa@juntadeandalucia.es), Maripaz Ventero Martín (ventero_marmar@gva.es), Carmen Molina Pardines (molina_carpar@gva.es), Nieves Orta Mira (orta_nie@gva.es), María Navarro Cots (navarro_dia@gva.es), Inmaculada Vidal Catalá (vidal_inm@gva.es), Isabel García Nava (garcianava@hotmail.com), Soledad Illescas Fernández-Bermejo (msillescasf@sescam.jccm.es), José Martínez-Alarcón (jmalarcon@sescam.jccm.es), Marta Torres-Narbona (mtorresn@sescam.jccm.es), Cristina Colmenarejo (ccolmenarejo@sescam.jccm.es), Lidia García-Agudo (lgagudo@sescam.jccm.es), Jorge Alfredo Pérez García (jorgep@sescam.jccm.es), Martín Yago López (yago_mar@gva.es), María Ángeles Goberna Bravo (mariangoberna@gmail.com), Carolina Pla Cortes (caly.cortes@gmail.com), Noelia Lozano Rodríguez (lozano_noe@gva.es), Nieves Aparici Valero (aparici_nie@gva.es), Sandra Moreno Marro (moreno_sanmar@gva.es), Agustín Irazo Tatay (iranzo_agu@gva.es), Isabel Mariscal Pieper (mariscal_isa@gva.es), Mª Pilar Ramos (ramos_marrei@gva.es), Mónica Parra Grande (parra_mongra@gva.es), Bárbara Gómez Alonso (gomez_baralo@gva.es), Francisco José Arjona Zaragozí (arjona_fra@gva.es), Amparo Broseta Tamarit (amparo2400@gmail.com), Juan José Badiola Díez (badiola@unizar.es), Alicia Otero García (aliciaogar@unizar.es), Eloísa Sevilla Romeo (esevillr@unizar.es), Belén Marín González

(belenm@unizar.es), Mirta García Martínez (gmmirta@hotmail.com), Marina Betancor Caro (mbetancorcaro@gmail.com), Diego Sola Fraca (diegosola95@gmail.com), Sonia Pérez Lázaro (soniperez97@gmail.com), Eva Monleón Moscardó (emonleon@unizar.es), Marta Monzón Garcés (mmonzon@unizar.es), Cristina Acín Tresaco (crisacin@unizar.es), Rosa Bolea Bailo (rbolea@unizar.es), Bernardino Moreno Burgos (bmoreno@unizar.es), Amparo Broseta Tamarit (abroseta@hospitalmanises.es), Carlos Gulin Blanco (cgulin@hospitalmanises.es)

Supplementary Material

Supplementary video 1. Geographical transmission of B.1.177.637 visualized with Nextstrain build.

Supplementary Tables

Supplementary table 1. Accession numbers for analysed sarbecoviruses.

Supplementary table 2. Defining SNPs for B.1.177.637, B.1.177.637.V2, and VOI1163.654.

Supplementary table 3. Accession numbers of sequences of interest for mutated amino acids 484, 501, 1163 and 1167 of the S protein.

Supplementary table 4. Accession numbers of 5,017 Spanish sequences from SeqCOVID-SPAIN consortium.

Supplementary table 5. Accession numbers of 270,869 analysed sequences.

Supplementary table 6. Accession numbers of 10,450 analysed sequences.

Supplementary table 7. Accession numbers of 3,067 analysed sequences belonging to 20I/501Y.V1.

Supplementary table 8. Primers for site directed mutagenesis of plasmid encoding codon-optimized S protein.

Supplementary table 9. Primers for sanger sequencing to detect mutations of interest.

Supplementary Figures

Supplementary Figure 1. Temporal distribution of mutated samples coloured by region. a. Distribution of S protein amino acid replacement D1163Y over the pandemic (N=1874). **b.** Distribution of S protein amino acid replacement G1167V over time (N=1708).

Supplementary Figure 2. Maximum-likelihood phylogenies for different PANGO lineages. a. Complete phylogeny coloured by the PANGO lineages. **b.** Phylogeny of B.53 lineage. The circle represents sequences with D1163 amino acid replacement. **c.** Phylogeny of B.53 lineage. The inner circle represents sequences with D1163 amino acid replacements, and the external circle represents sequences with G1167 amino acid replacements. **d.** Phylogeny of A lineage. The circle represents sequences with D1163 amino acid replacements. **e.** Phylogeny of B.1 and derived lineages. The inner circle represents sequences with D1163 amino acid replacements, and the external circle represents sequences with G1167 amino acid replacements. **f.** Phylogeny of B.1.1 and derivative D.1 lineages. The inner circle represents sequences with D1163 amino acid replacements, and the external circle represents sequences with G1167 amino acid replacements. **g.** Phylogeny of B.1.177 and B.1.177.637 lineages. The inner circle represents sequences with D1163 amino acid replacements, and the external circle represents sequences with G1167 amino acid replacements. The scale bar of each indicates the number of nucleotide substitutions per site. The legend of positions 1163 and 1167 is common for all represented panels.

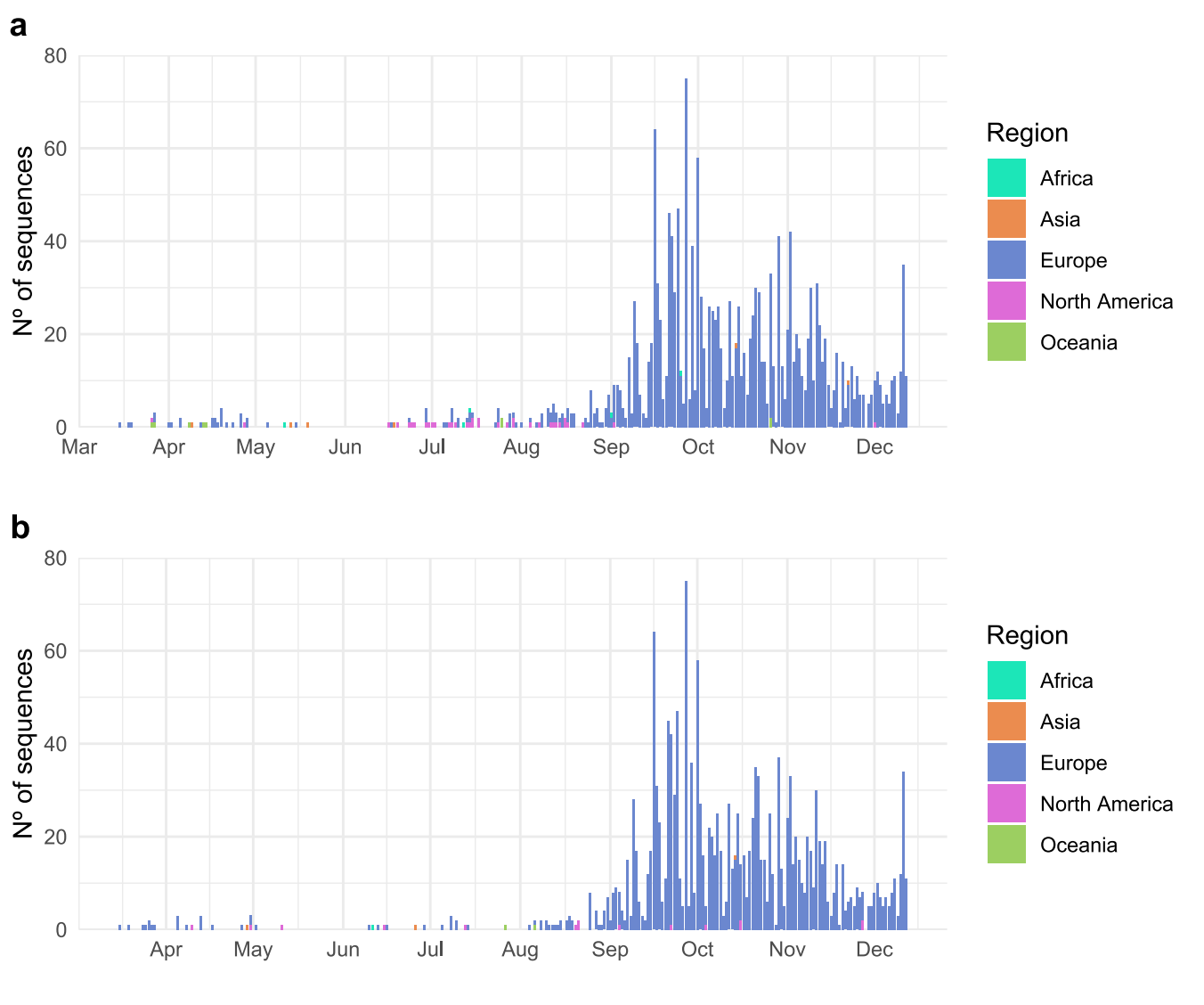
Supplementary Figure 3. **Whole genome mutated positions in genotypes with changes 1163Y and 1167V of the S protein.** B.1.177.637 is coloured in magenta and cluster 1163.654 is coloured

in orange. Other less frequent genotypes (found in at least 20 sequences) that include changes in S position 1163 and/or 1167 are coloured in turquoise. Cluster 1163.654 is coloured in navy blue. Line width is proportional to the frequency of the genotype. Sites 1163 and 1167 in the S protein are indicated by magenta star symbols, and position 484 in S, whose mutations are associated with antigenicity changes, is indicated by a navy blue star.

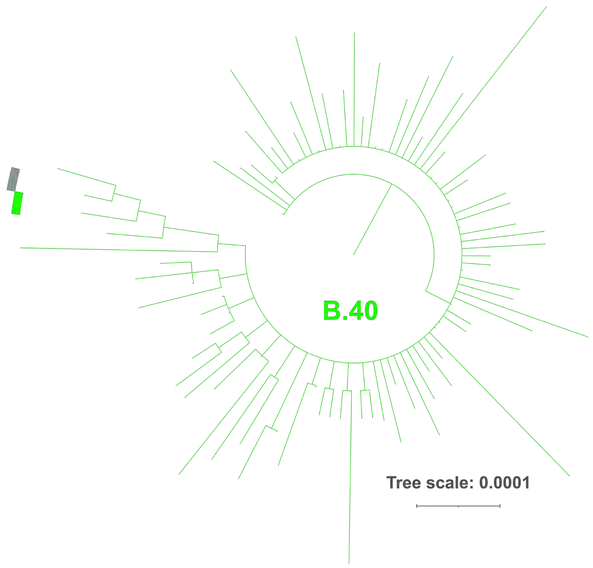
Supplementary Figure 4. Maximum-likelihood phylogeny of 3,067 genomes belonging to 20I/501Y.V1, rooted with Wuhan reference sequence. Mutated sequences with 1163 and/or 1167 amino acids of the S protein are coloured in the circle. The scale bar indicates the number of nucleotide substitutions per site. The biggest clades are collapsed with isosceles triangles.

Supplementary Figure 5. Maximum-likelihood phylogeny of 3,266 SARS-CoV-2 genomes representing 20E clade rooted with Wuhan reference sequence. Sequences from B.1.177.637 are coloured in magenta, sequences not identified as B.1.177.637 are coloured in green, and cluster B.1.177.637.V2 (S protein amino acid replacements: A222V, D614G, E484K, D1163Y, and 141-144Del) is coloured in blue. The scale bar indicates the number of nucleotide substitutions per site. The biggest clades are collapsed, represented as isosceles triangles.

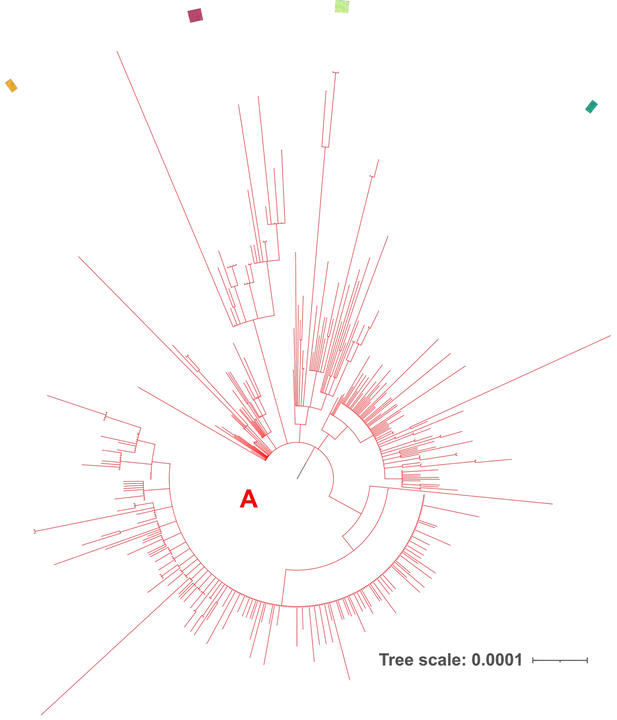
Supplementary Figure 6. Neutralization of the different mutated S protein variants by convalescent sera from six individuals infected during the first epidemic wave. The reciprocal titer at which each of the different convalescent sera neutralizes the different variants by 50% is indicated. Plotted are the mean and standard error of (n=3).


a
b
Nº of sequences
0
20
40
60
80
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Region
Africa
Asia
Europe
North America
Oceania

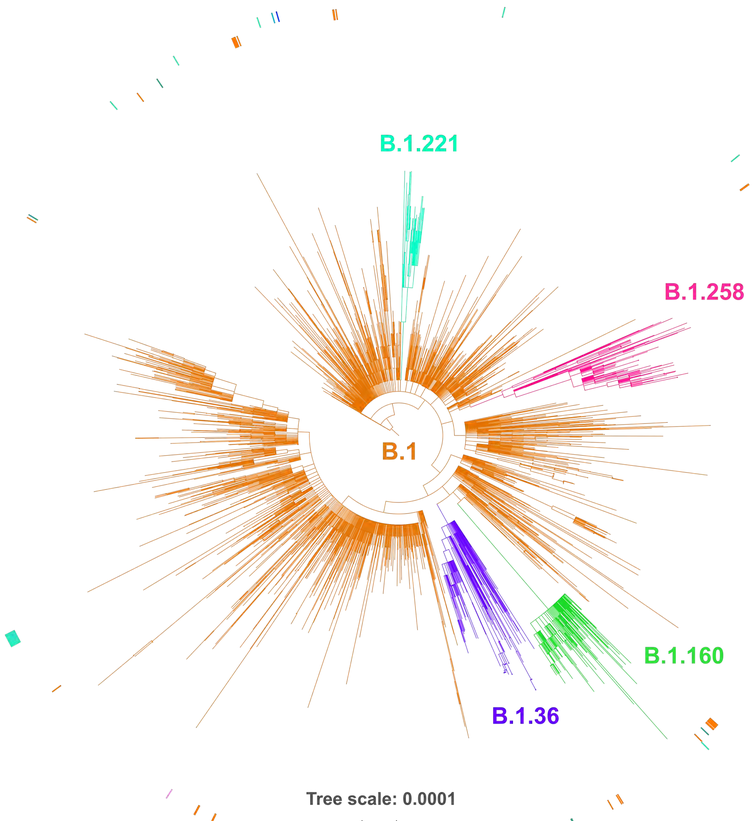
c


B.40
Tree scale: 0.0001

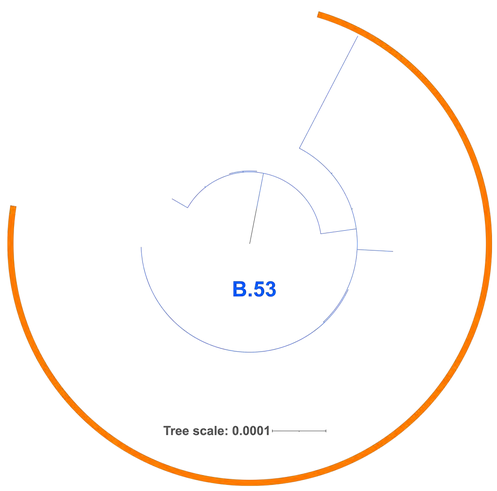
d


A
Tree scale: 0.0001

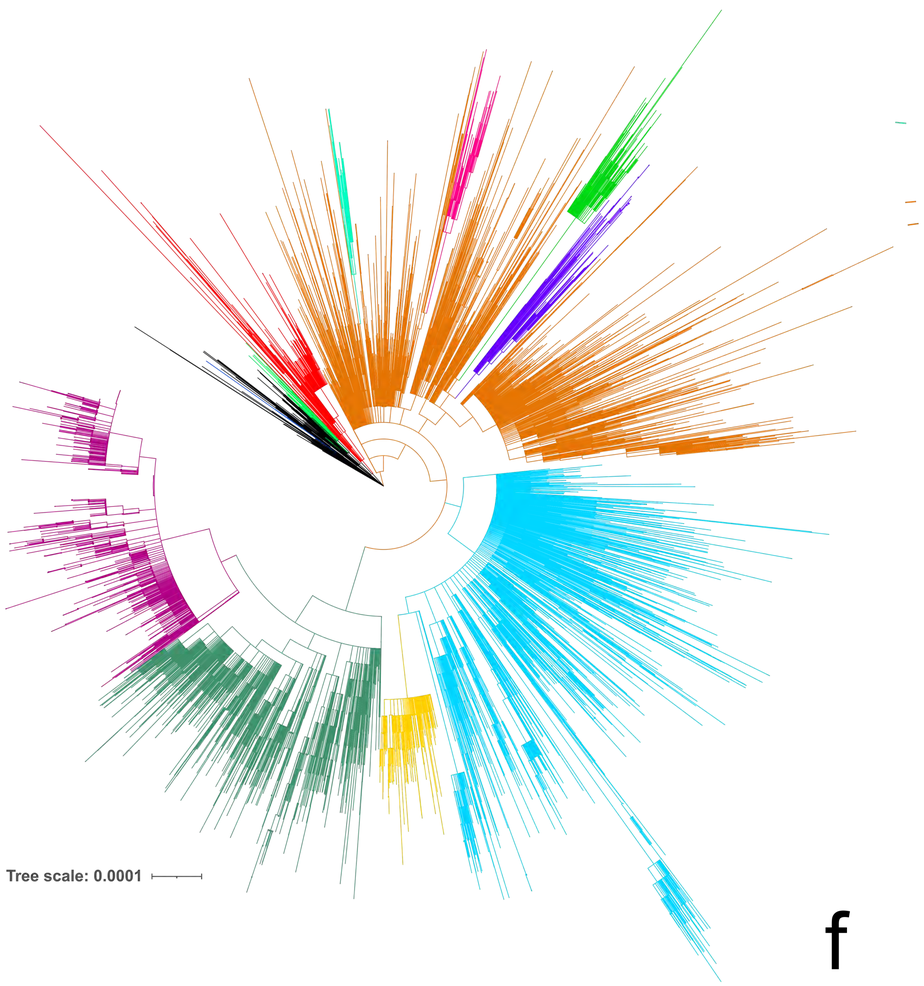
e


B.1.221
B.1.258
B.1
B.1.160
B.1.36
Tree scale: 0.0001

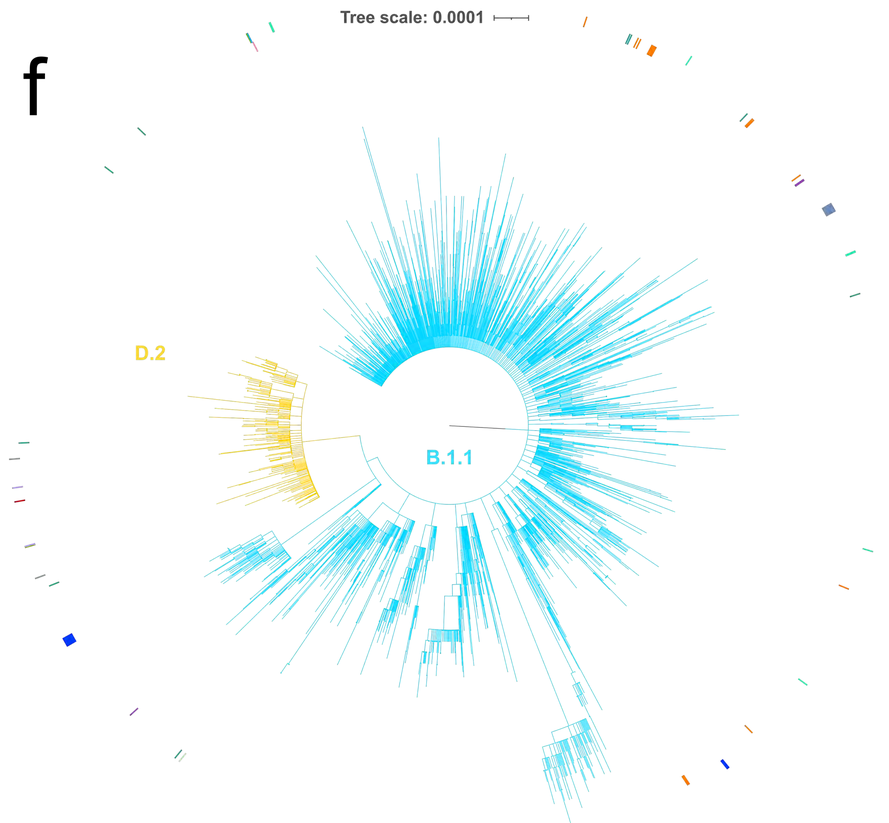
b


B.53
Tree scale: 0.0001

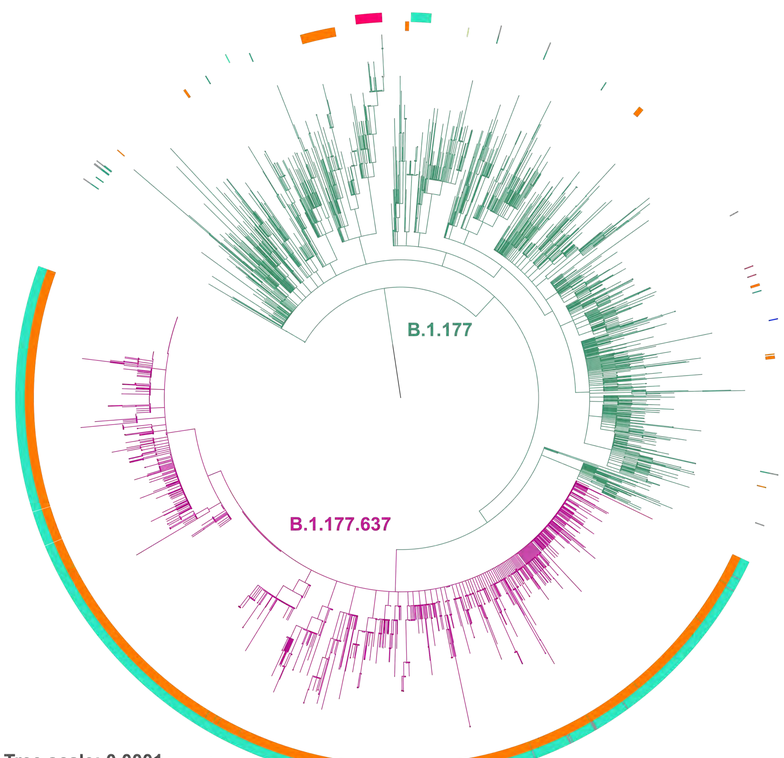
a


Tree scale: 0.0001

f

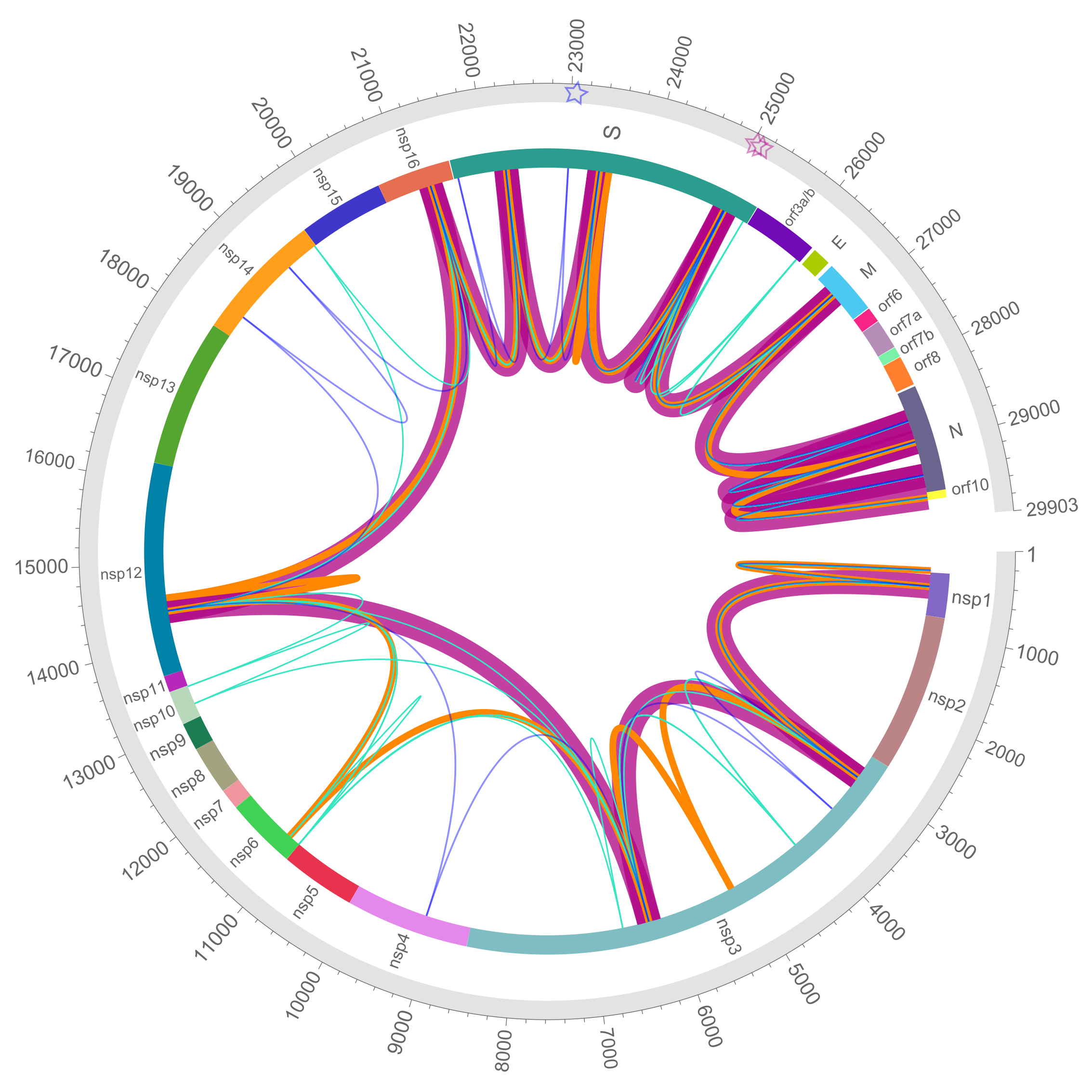

D.2
B.1.1
Tree scale: 0.0001

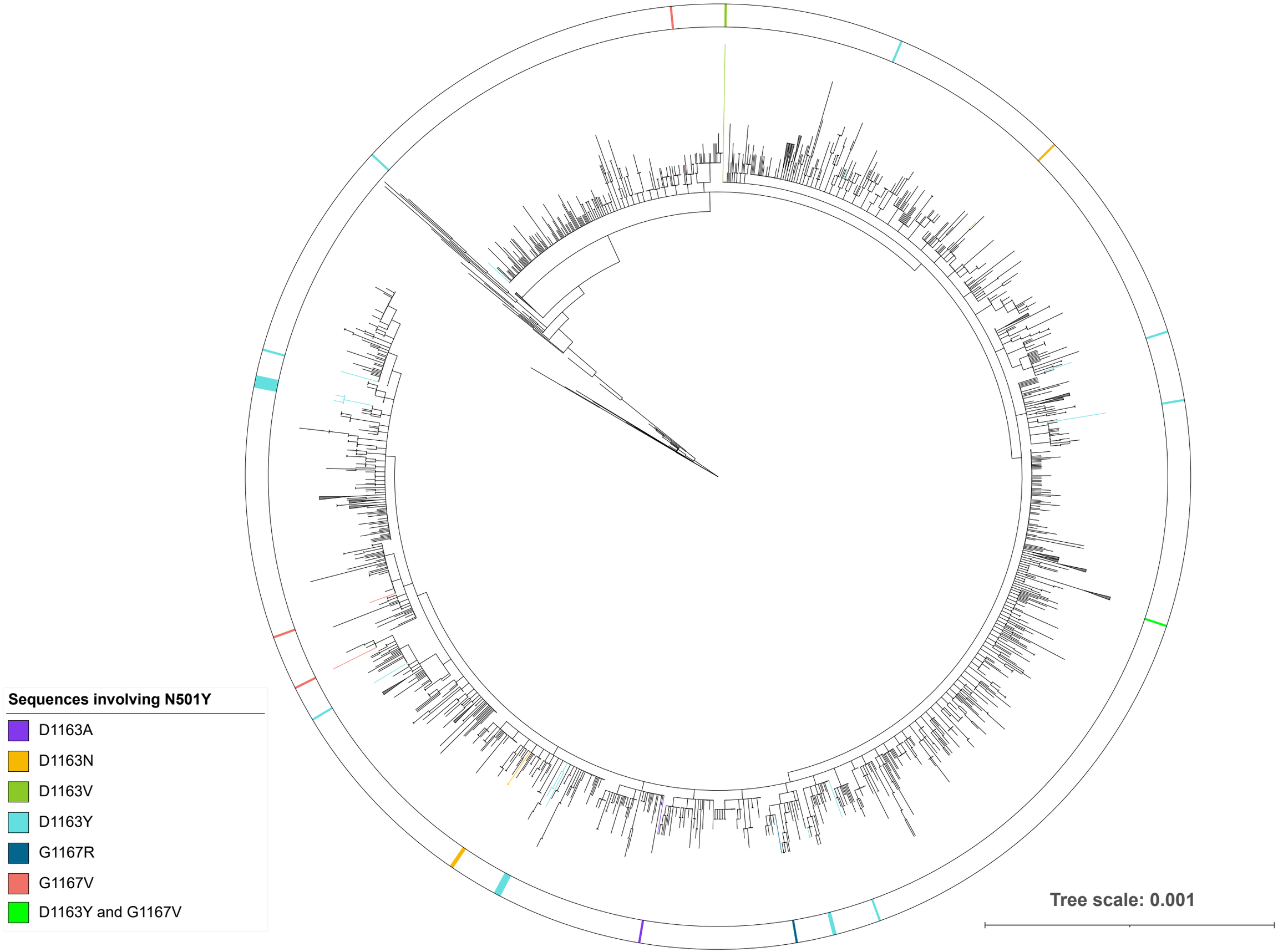
g

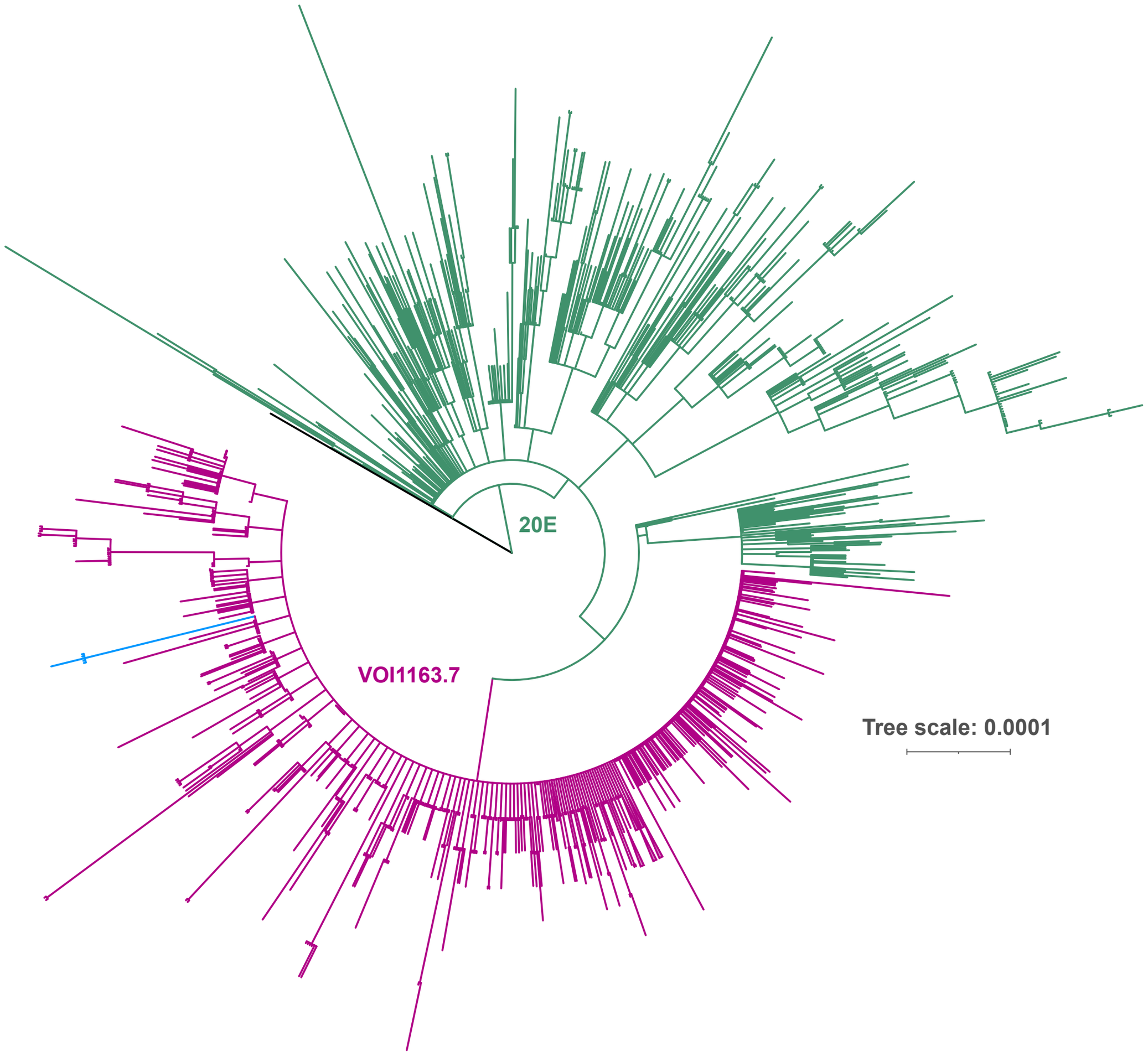

B.1.177
B.1.177.637
Tree scale: 0.0001

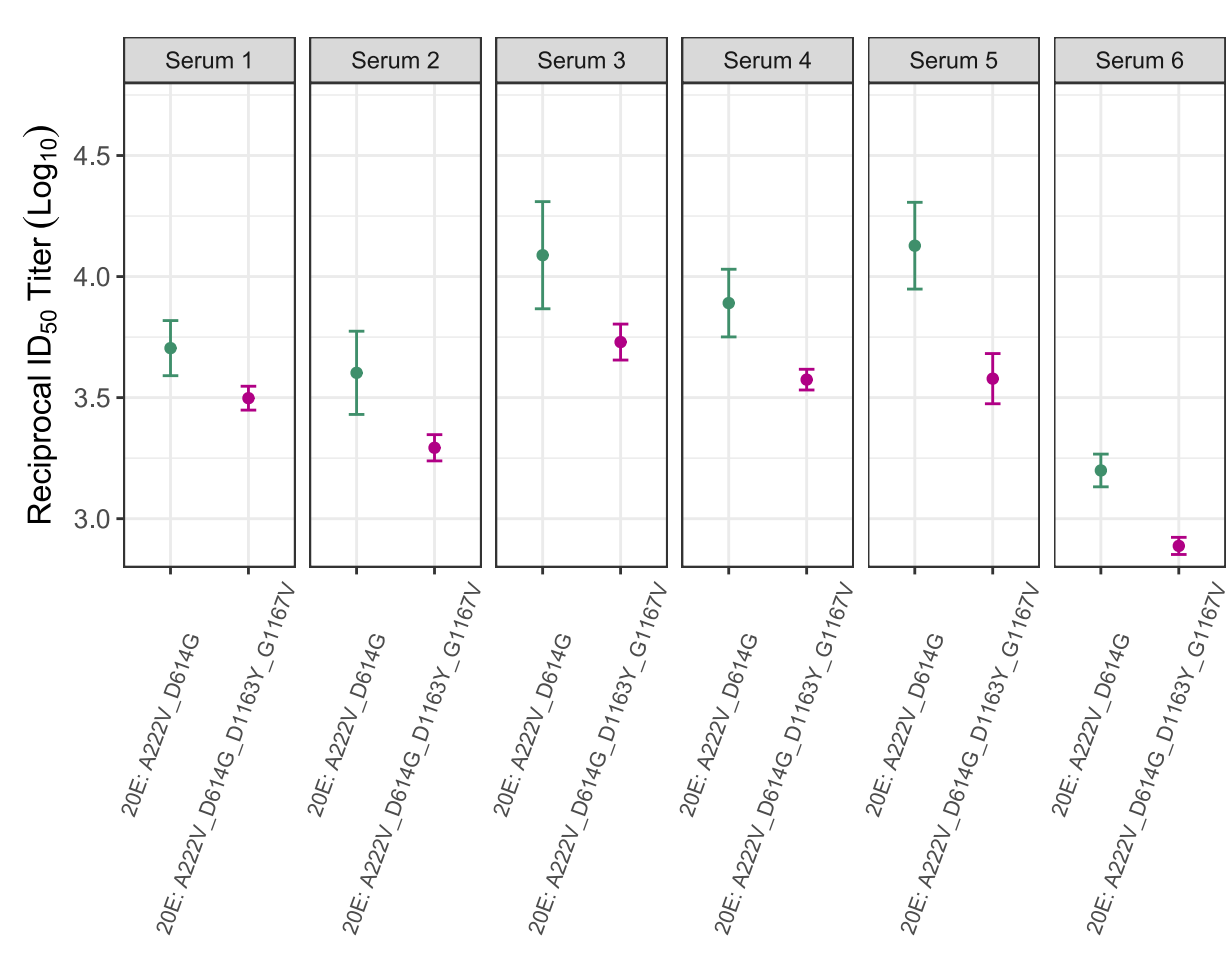

Position D1163
1163A
1163E/D
1163G
1163H
1163Y/D
1163H/Y/D
1163N/D
1163V
1163Y

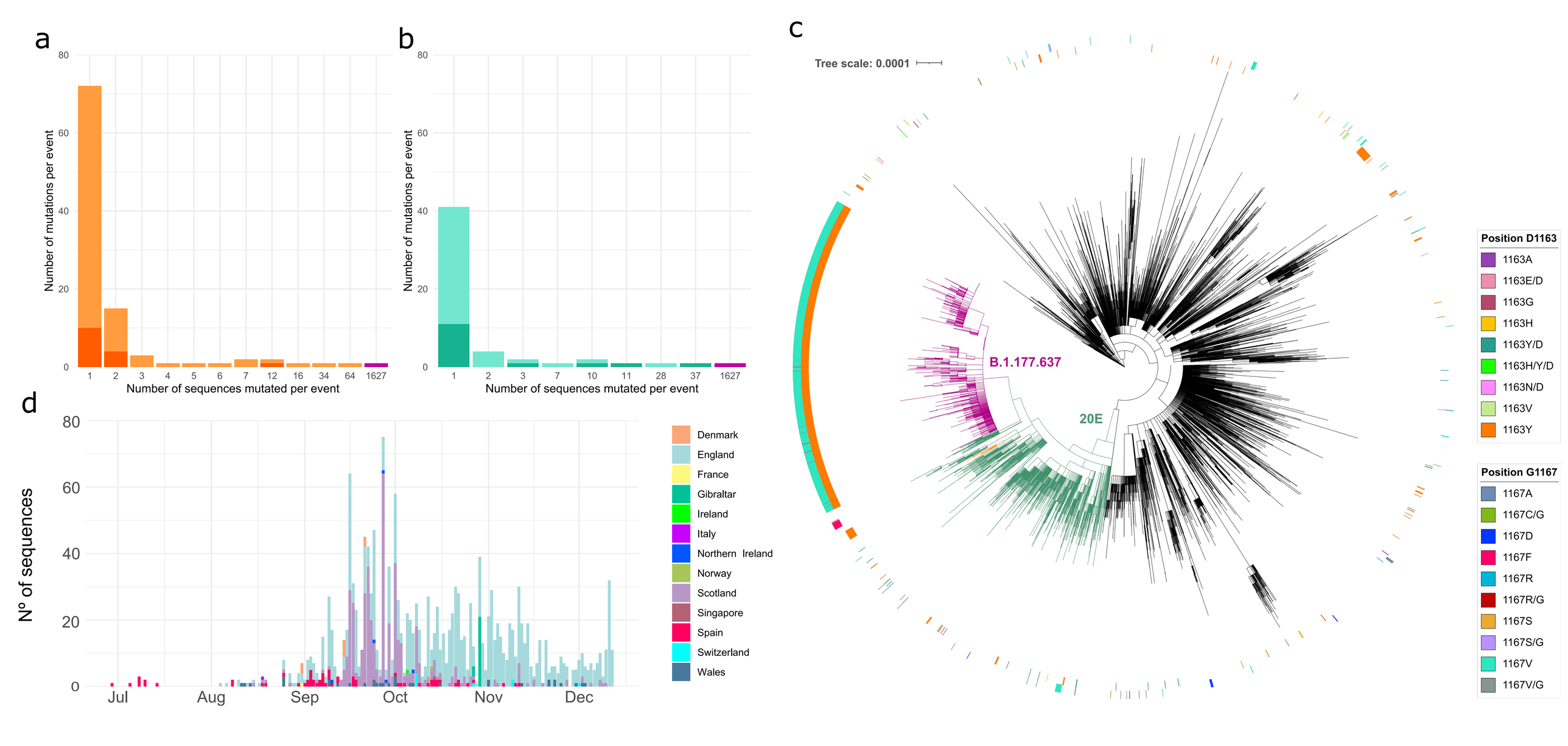
Position G1167
1167A
1167C/G
1167D
1167F
1167R
1167R/G
1167S
1167S/G
1167V
1167V/G

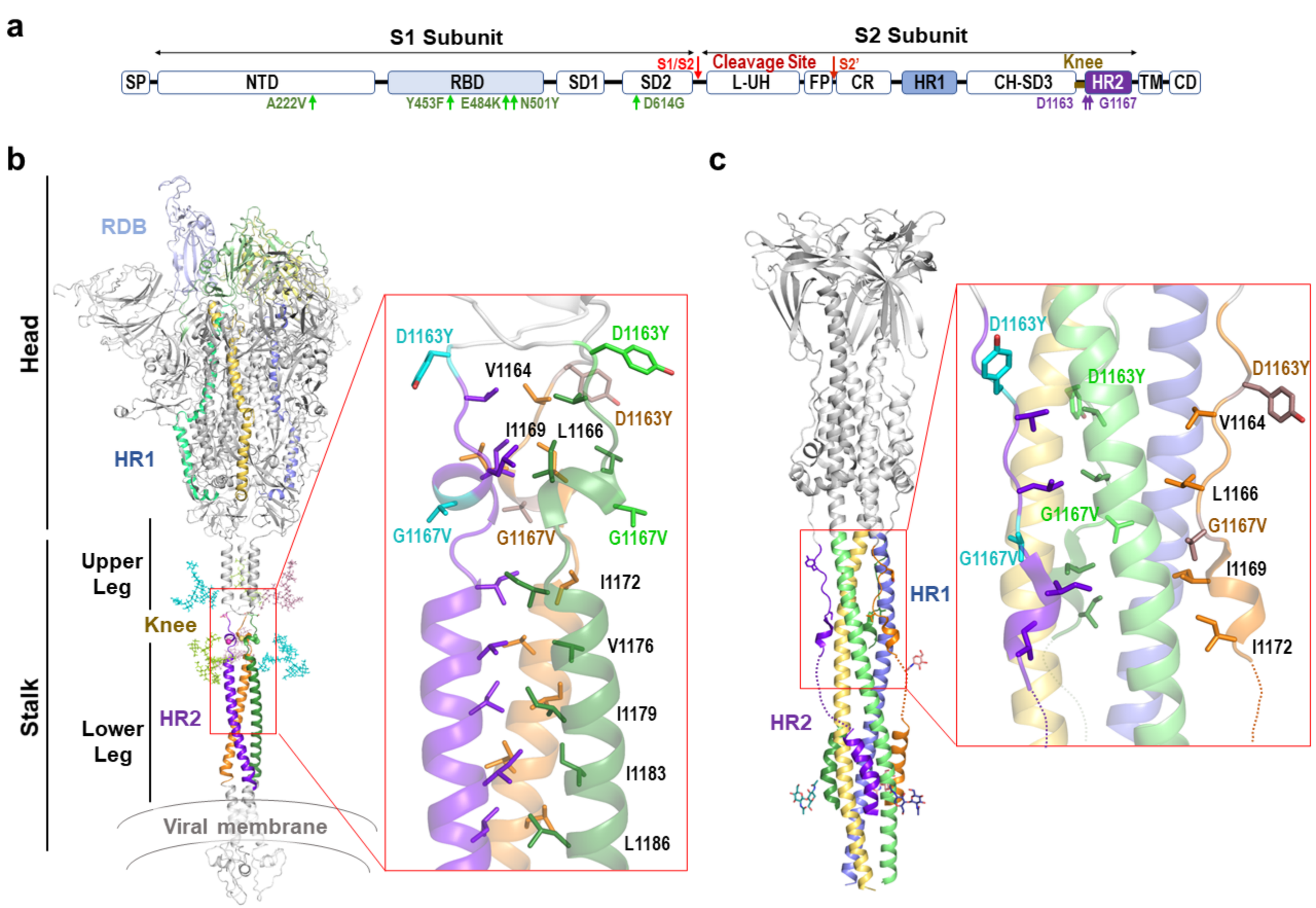

1
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
24000
25000
26000
27000
28000
29000
29903
nsp1
nsp2
nsp3
nsp4
nsp5
nsp6
nsp7
nsp8
nsp9
nsp10
nsp11
nsp12
nsp13
nsp14
nsp15
nsp16
S
orf3a/b
E
M
orf6
orf7a
orf7b
orf8
N
orf10

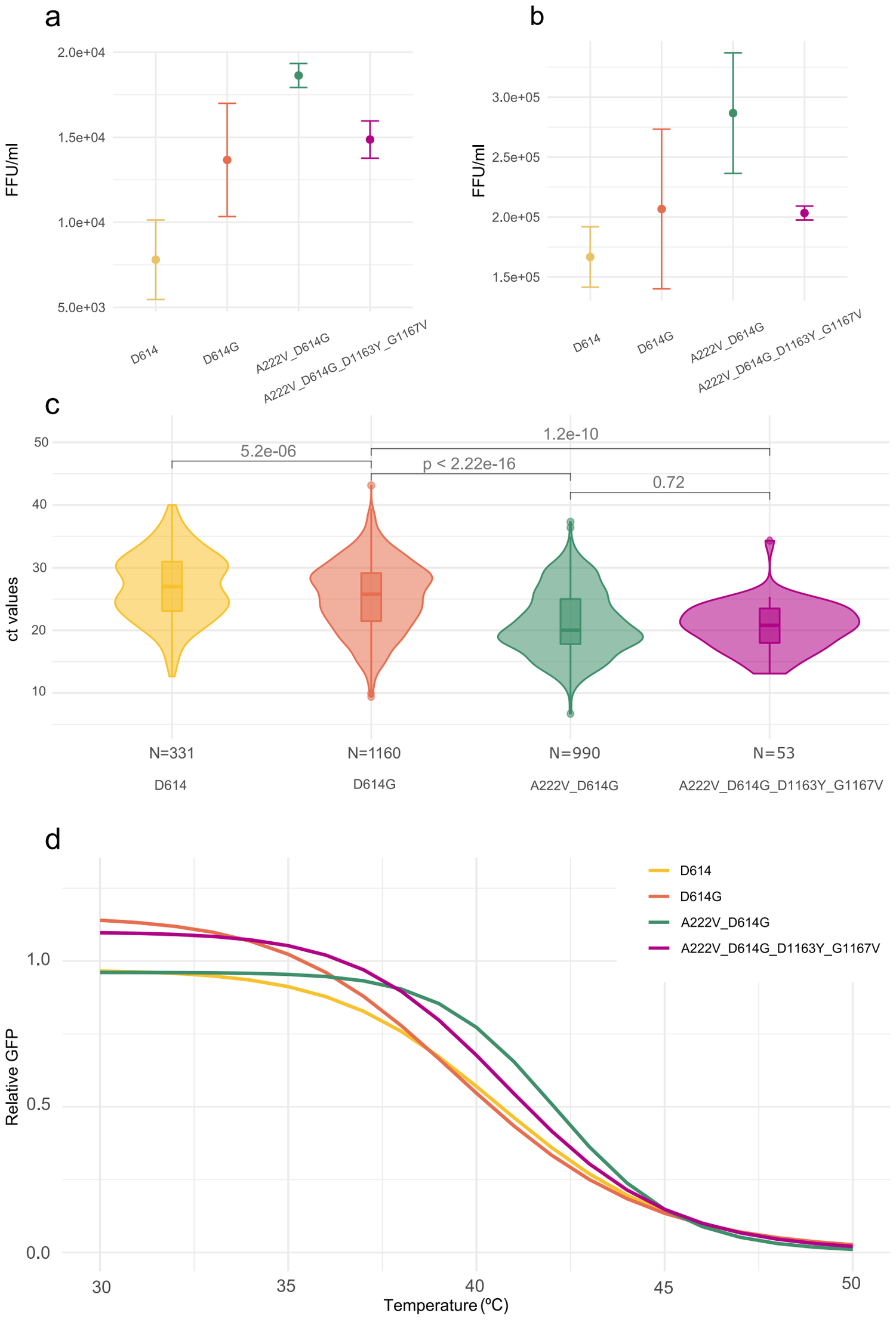

Sequences involving N501Y
D1163A
D1163N
D1163V
D1163Y
G1167R
G1167V
D1163Y and G1167V
Tree scale: 0.001


20E
VOI1163.7
Tree scale: 0.0001

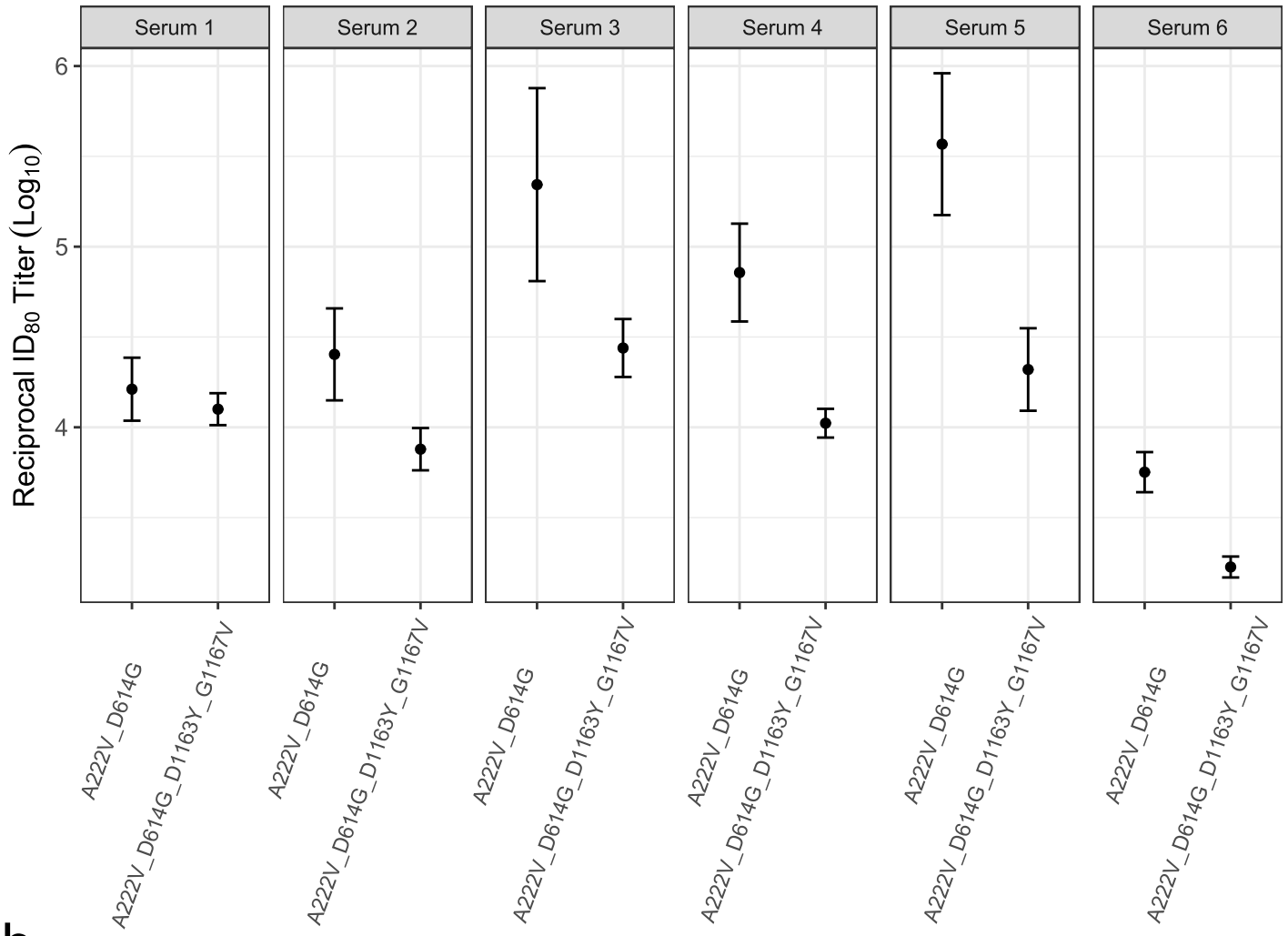

Serum 1
Serum 2
Serum 3
Serum 4
Serum 5
Serum 6
Reciprocal ID50 Titer (Log10)
4.5
4.0
3.5
3.0
20E: A222V_D614G
20E: A222V_D614G_D1163Y_G1167V


a
b
c
d
Number of mutations per event
Number of sequences mutated per event
1
2
3
4
5
6
7
12
16
34
64
1627
10
11
28
37
Tree scale: 0.0001
B.1.177.637
20E
Position D1163
1163A
1163E/D
1163G
1163H
1163Y/D
1163H/Y/D
1163N/D
1163V
1163Y
Position G1167
1167A
1167C/G
1167D
1167F
1167R
1167R/G
1167S
1167S/G
1167V
1167V/G
Nº of sequences
Jul
Aug
Sep
Oct
Nov
Dec
Denmark
England
France
Gibraltar
Ireland
Italy
Northern Ireland
Norway
Scotland
Singapore
Spain
Switzerland
Wales

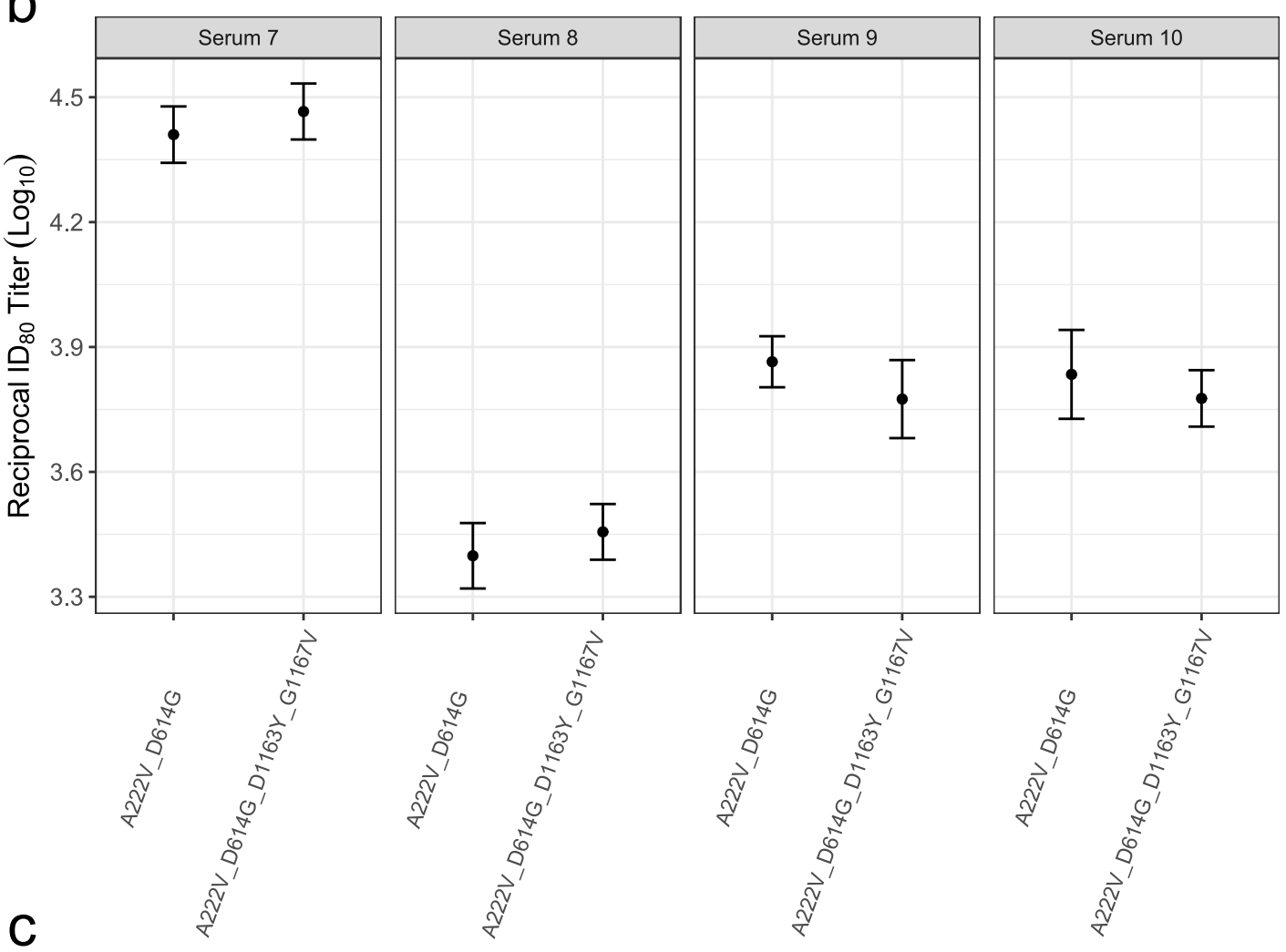

a
S1 Subunit
S2 Subunit
S1/S2
Cleavage Site
S2'
Knee
SP
NTD
RBD
SD1
SD2
L-UH
FP
CR
HR1
CH-SD3
HR2
TM
CD
A222V
Y453F
E484K
N501Y
D614G
D1163
G1167
b
Head
Stalk
RDB
HR1
Upper Leg
Knee
Lower Leg
HR2
Viral membrane
D1163Y
D1163Y
V1164
D1163Y
I1169
L1166
G1167V
G1167V
G1167V
I1172
V1176
I1179
I1183
L1186
c
HR1
HR2
D1163Y
D1163Y
D1163Y
V1164
L1166
G1167V
G1167V
G1167V
I1169
I1172


a
FFU/ml
2.0e+04
1.5e+04
1.0e+04
5.0e+03
D614
D614G
A222V_D614G
A222V_D614G_D1163Y_G1167V
b
FFU/ml
3.0e+05
2.5e+05
2.0e+05
1.5e+05
D614
D614G
A222V_D614G
A222V_D614G_D1163Y_G1167V
c
ct values
50
40
30
20
10
5.2e-06
1.2e-10
p < 2.22e-16
0.72
N=331
D614
N=1160
D614G
N=990
A222V_D614G
N=53
A222V_D614G_D1163Y_G1167V
d
Relative GFP
1.0
0.5
0.0
30
35
40
45
50
Temperature (°C)
D614
D614G
A222V_D614G
A222V_D614G_D1163Y_G1167V

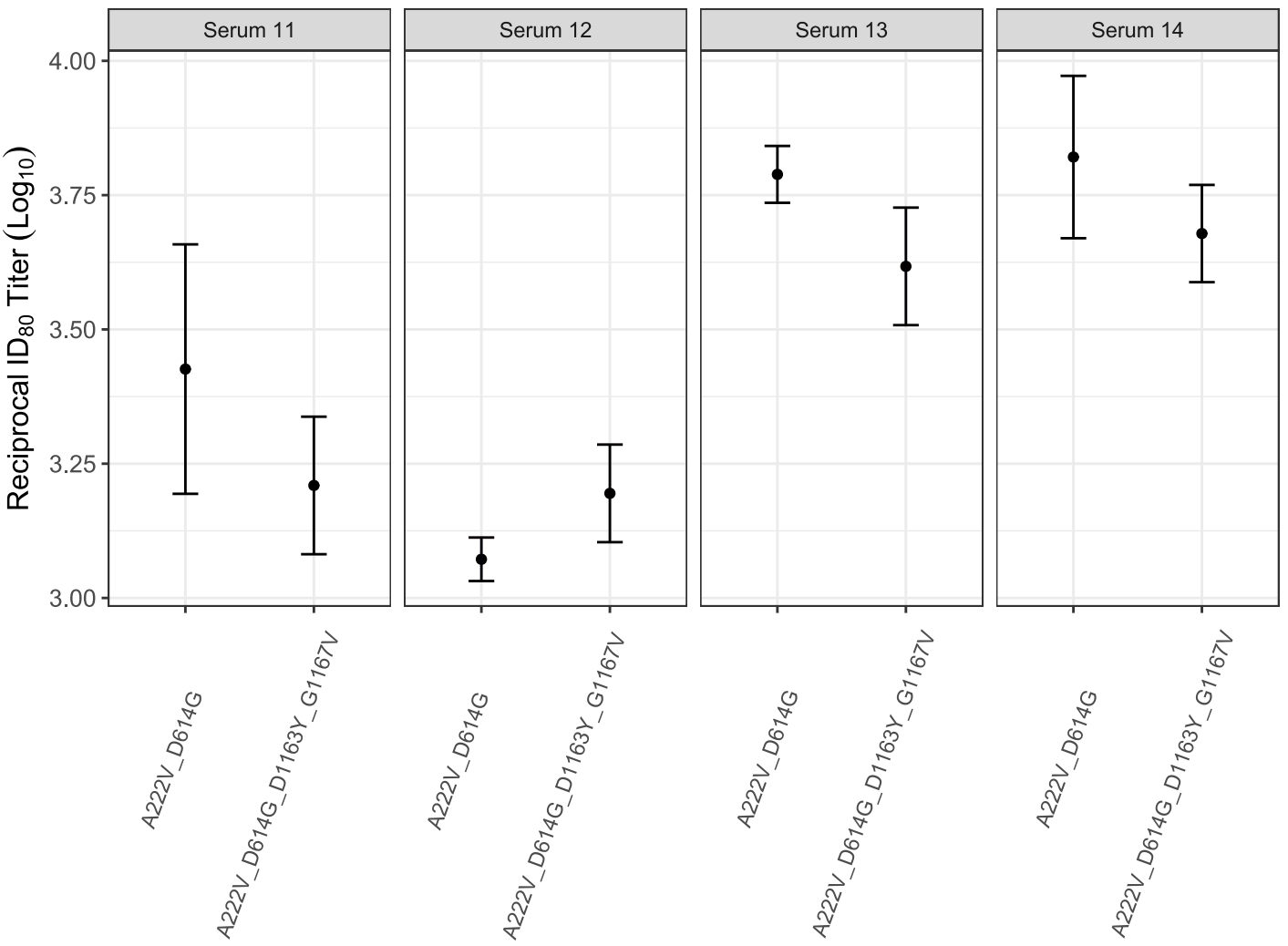
a


Serum 1
Serum 2
Serum 3
Serum 4
Serum 5
Serum 6
Reciprocal ID80 Titer (Log10)
6
5
4
A222V_D614G
A222V_D614G_D1163Y_G1167V

b


Serum 7
Serum 8
Serum 9
Serum 10
Reciprocal ID80 Titer (Log10)
4.5
4.2
3.9
3.6
3.3
A222V_D614G
A222V_D614G_D1163Y_G1167V

c


Serum 11
Serum 12
Serum 13
Serum 14
Reciprocal ID80 Titer (Log10)
4.00
3.75
3.50
3.25
3.00
A222V_D614G
A222V_D614G_D1163Y_G1167V