

1 Genomic Perspectives on the Emerging SARS-CoV-2 Omicron 2 Variant

3 Wentai Ma^{1,2,#}, Jing Yang^{1,2,#}, Haoyi Fu^{1,2}, Chao Su³, Caixia Yu⁴, Qihui Wang³, Ana
4 Tereza Ribeiro de Vasconcelos⁵, Georgii A. Bazykin^{6,7}, Yiming Bao^{2,4}, Mingkun
5 Li^{1,2,8,*}

6

7 ¹ CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of
8 Genomics, Chinese Academy of Sciences, and China National Center for
9 Bioinformation, Beijing 100101, China

10 ²University of Chinese Academy of Sciences, Beijing 100049, China

11 ³CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of
12 Microbiology, Chinese Academy of Sciences, Beijing 100101, China

13 ⁴National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of
14 Sciences / China National Center for Bioinformation Beijing 100101, China

15 ⁵Laboratório de Bioinformática, Laboratório Nacional de Computação Científica,
16 Petrópolis 25651-075, Brazil

17 ⁶Skolkovo Institute of Science and Technology, Moscow 121205, Russia

18 ⁷Kharkevich Institute for Information Transmission Problems of the Russian Academy
19 of Sciences, Moscow 127051, Russia

20 ⁸Center for Excellence in Animal Evolution and Genetics, Chinese Academy of
21 Sciences, Kunming 650201, China

22 [#] Equal contribution.

23 ^{*} Corresponding author.

24 E-mail: limk@big.ac.cn (Li M)

25

26 Running title: Ma W et al / Genomic Perspective on Omicron

27

28 The total number of letters in article title: 56

- 29 The total number of letters in running title: 35
- 30 The total number of words in abstract: 249
- 31 The total number of keywords: 5
- 32 The total number of word (from “Introduction” to “ORCID”): 3225
- 33 The total number of reference count: 38
- 34 The total number of figures: 4
- 35 The total number of tables: 0
- 36 The total number of supplementary figures: 2
- 37 The total number of supplementary tables: 1
- 38

39 Abstract

40 A new variant of concern for SARS-CoV-2, Omicron (B.1.1.529), was designated by
 41 the World Health Organization on November 26, 2021. This study analyzed the viral
 42 genome sequencing data of 108 samples collected from patients infected with
 43 Omicron. First, we found that the enrichment efficiency of viral nucleic acids was
 44 reduced due to mutations in the region where the primers anneal to. Second, the
 45 Omicron variant possesses an excessive number of mutations compared to other
 46 variants circulating at the same time (62 vs. 45), especially in the *Spike* gene.
 47 Mutations in the *Spike* gene confer alterations in 32 amino acid residues, which was
 48 more than those observed in other SARS-CoV-2 variants. Moreover, a large number
 49 of nonsynonymous mutations occur in the codons for the amino acid residues located
 50 on the surface of the Spike protein, which could potentially affect the replication,
 51 infectivity, and antigenicity of SARS-CoV-2. Third, there are 53 mutations between
 52 the Omicron variant and its closest sequences available in public databases. Many of
 53 those mutations were rarely observed in the public database and had a low mutation
 54 rate. In addition, the linkage disequilibrium between these mutations were low, with a
 55 limited number of mutations (6) concurrently observed in the same genome,
 56 suggesting that the Omicron variant would be in a different evolutionary branch from
 57 the currently prevalent variants. To improve our ability to detect and track the source
 58 of new variants rapidly, it is imperative to further strengthen genomic surveillance and
 59 data sharing globally in a timely manner.

60 **Keywords:** Omicron; Genomics; Mutation; Variant of concern; SARS-CoV-2

61 Introduction

62 On November 22, 2021, the first genome sequence of a new variant of concern (VOC),
 63 Omicron (also known as B.1.1.529), was released in GISAID (Global initiative on
 64 sharing all influenza data) (EPI_ISL_6590782) [1]. The sample was obtained from a
 65 patient who arrived in Hong Kong on November 11 from South Africa via Doha in
 66 Qatar
 67 ([https://news.sky.com/story/covid-19-how-the-spread-of-omicron-went-from-patient-](https://news.sky.com/story/covid-19-how-the-spread-of-omicron-went-from-patient-zero-to-all-around-the-globe-12482183)
 68 [zero-to-all-around-the-globe-12482183](https://news.sky.com/story/covid-19-how-the-spread-of-omicron-went-from-patient-zero-to-all-around-the-globe-12482183)). To date, the first known Omicron variant
 69 sample was collected on November 5, 2021 in South Africa (EPI_ISL_7456440).
 70 Until December 12, 2021, there were over 2000 Omicron sequences submitted to the
 71 GISAID from South Africa, Botswana, Ghana, the United Kingdom, and many other
 72 countries. The emergence of this variant has attracted much attention due to the sheer
 73 number of mutations in the *Spike* gene, which may affect the viral transmissibility,
 74 replication, and binding of antibodies, and its dramatic increase in South Africa [2].
 75 Preliminary studies showed that the new variant could substantially evade immunity
 76 from prior infection and vaccination [3,4]. Meanwhile, a preprint report proposed that
 77 the emergence of the Omicron variant was associated with an increased risk of
 78 SARS-CoV-2 reinfection [5]. However, it is still unclear where the new variant came.

79 In this study, we characterized the genomic features of the Omicron variant using
 80 data from 108 patients infected with the Omicron variant, which were generated by
 81 the Network for Genomic Surveillance in South Africa (NGS-SA) [2,6], and we
 82 speculate that the new variant is unlikely derived from recently discovered variants
 83 through either mutation or recombination.

84 Results

85 Reduced enrichment efficiency of the PCR-tiling amplicon protocols on the 86 Omicron variant

87 Of the 207 Omicron samples sequenced, 158 samples had more than 90% of the viral
 88 genome covered by at least 5-fold, which were used in the subsequent analysis.
 89 Notably, two sequencing protocols were implemented. The first was to enrich the viral
 90 genome with the Midnight V6 primer sets followed by sequencing on the GridION
 91 platform (hereinafter referred to as Midnight,
 92 [dx.doi.org/10.17504/protocols.io.bwyppfvn](https://doi.org/10.17504/protocols.io.bwyppfvn)). The second protocol involved

93 enrichment by the Artic V4 primer set, and the amplicons were sequenced on the
94 Illumina MiSeq platform (hereinafter referred to as Artic,
95 [dx.doi.org/10.17504/protocols.io.bdp7i5rn](https://doi.org/10.17504/protocols.io.bdp7i5rn)). Fifty samples were sequenced using both
96 protocols, and we found a high consistency in the major allele frequency between the
97 two protocols (Figure S1). Artic data were preferred due to higher sequencing depth
98 (median: 191 vs. 250, $P < 0.01$, Mann–Whitney U test). Finally, 49 samples
99 sequenced by the Midnight protocol and 59 samples sequenced by the Artic protocol
100 were included in the study.

101 Both protocols enabled efficient enrichment of viral nucleic acids from total
102 RNA, the fraction of SARS-CoV-2 reads in the sequencing data were 84% and 94%,
103 respectively for the Midnight and the Artic protocol. Although the Artic protocol had a
104 relatively higher in-target percentage ($P < 0.001$, Mann–Whitney U test), the evenness
105 of the sequencing depth of the SARS-CoV-2 was higher for the Midnight protocol
106 (variance of the sequencing depth, 0.121 vs. 0.159, $P < 0.001$, Mann–Whitney U test).
107 The sequencing depth profile of the SARS-CoV-2 genome was similar among
108 samples sequenced by the same protocol but differed markedly between the two
109 protocols (**Figure 1A**). The sequencing depth varied among different genomic regions,
110 reflecting the differential enrichment efficiency of the primers. Moreover, we found
111 that the large number of mutations possessed by the Omicron variant had a significant
112 impact on the efficiency of the primers. In particular, seven primers in the Artic
113 protocol were affected by at least one mutation, and three primers in the Midnight
114 protocol were affected (Figure 1A). The worst coverage of the three regions for
115 Primers 76, 79, and 90 using the Artic protocol were all associated with the presence
116 of mutations in the region where these primers annealed to, whose sequencing depths
117 were reduced by 2586, 246, and 234-fold, respectively, compared to the expected
118 depth (Figure 1B). Strikingly, five mutations were located at the 5' end of the least
119 efficient Primer 76. The enrichment efficiency of other four primers (Primers 10, 27,
120 88, 89) was less affected by the mutations, which showed 1.3, 1.4, 3.4, and 1.9-fold
121 reductions, respectively. Thus, the results suggest that the Omicron mutations can
122 decrease the enrichment efficiency by PCR amplification, and there is an urgent need

123 to update the Arctic V4 primers. We noted that the developer of the Artic protocol had
124 already proposed a solution on this, and all seven affected primers had been updated
125 (<https://community.artic.network/t/sars-cov-2-v4-1-update-for-omicron-variant/342>).
126 In contrast, the efficiency of Midnight primers was less influenced by mutations in the
127 Omicron variant. The three affected primers, Primers 10, 24, 28, showed no reduction,
128 2-fold, and 28-fold reduction respectively in sequencing depth compared to the
129 expected depth.

130 **An extraordinary number of mutations in the *Spike* gene of the Omicron variant**

131 The number of mutations (with major allele frequency $\geq 70\%$) of the Omicron variant
132 varied from 61 to 64, and 61 of them were identified in more than 90% of the samples,
133 which included 54 SNPs, six deletions, and one insertion. All these mutations were
134 fixed at the individual level (**Figure 2A**). The total number of mutations was
135 significantly higher than that of other variants detected in South Africa in November
136 (median 62 vs. 45, $P < 0.001$, Mann–Whitney U test). Strikingly, over half of these
137 mutations (34, 55.7%) were located in the *Spike* gene, whose length was 12.8% of the
138 whole genome. Moreover, 32 of these mutations were nonsynonymous mutations. The
139 proportion was significantly higher than that observed in the same region in other
140 variants (94% vs. 67%, $P < 0.001$, Fisher’s exact test, $Ka/Ks [7] = 8.65$), suggesting
141 positive selection on this gene.

142 The Omicron variant showed a greater number of mutations than other VOCs
143 (Figure 2B). The difference was more marked in the *Spike* gene, where the Omicron
144 variant possessed 2-15 times more amino acid changes than other VOCs collected
145 simultaneously (Figures 2C, D). Strikingly, the divergence in the amino acid sequence
146 between the Omicron variant and the early SARS-CoV-2 sequence (Wuhan-Hu-1) in
147 the *Spike* and RBD regions was greater or equivalent to that between SARS-like
148 coronavirus (Pangolin MP789, Bat BANAL-20-52, and Bat RaTG13) and
149 Wuhan-Hu-1 [8–12]. The dramatic changes in the *Spike* and RBD regions may
150 substantially change the antigenicity and susceptibility to pre-existing antibodies.

151 **Potential risks associated with Omicron mutations**

152 Most mutations occurred on the surface of the trimeric spike protein, especially in the
153 RBD region (Figure S2). Eight of the 16 mutations in the RBD (K417N, G446S,

154 E484A, Q493R, G496S, Q498R, N501Y, and Y505H) were located at positions that
155 were proposed to be critical for viral binding to the host receptor
156 angiotensin-converting enzyme 2 (ACE2) [13]. Among them, the K417N and N501Y
157 mutations, which were also identified in the Beta variant, were reported to influence
158 binding to human ACE2 [14]; N501Y confers a higher affinity of the viral Spike
159 protein to ACE2 [15]. How the other mutations affect the affinity to ACE2 of humans
160 and other animal hosts is still unknown.

161 Moreover, some other mutations in the *Spike* gene are known to be associated
162 with changes in replication and infectivity of the virus. For example, $\Delta 69-70$ could
163 enhance infectivity associated with increased cleaved Spike incorporation [16];
164 P681H could potentially confer replication advantage through increased cleavage
165 efficacy by furin and adaptation to resist innate immunity [3,17]; H655Y was
166 suspected to be an adaptive mutation that could increase the infectivity of the virus in
167 both human and animal models [16]. In addition, mutations in other genes, such as
168 R203K and G204R in the *Nucleoprotein* gene, could also potentially increase the
169 infectivity, fitness, and virulence of the virus [18]. Of note, the function of these
170 mutations was investigated because they were present in other VOCs. The effect of
171 other less frequent mutations and the combination of the aforementioned mutations on
172 the biology of the virus warrants further investigation.

173 Mutations in the RBD region, which is the target of many antibodies, may
174 compromise the neutralization of existing antibodies induced by vaccination or
175 natural infection [19]. Recent studies have shown severely reduced neutralization of
176 the Omicron variant by monoclonal antibodies and vaccine sera [4,20,21]. Meanwhile,
177 preliminary studies suggested that the Omicron variant caused three times more
178 reinfection than previous strains, further supporting the speculation that the new
179 variant can evade immunity from prior infection and vaccination [5]. However, the
180 escape was incomplete, and a vaccine booster shot is likely to provide a high level of
181 protection against the Omicron variant [4]. Here, we analyzed the epitope regions of
182 182 protein complex structures of antibodies that bound to SARS-CoV-2 Spike, the
183 RBD, or NTD from the Protein Data Bank. We found that mutations in the Omicron
184 variant were enriched in the epitope region of the Spike protein (**Figure 3A**). The
185 median number of antibodies bound to the Omicron mutation sites was 53, which was
186 significantly higher than those bound to other positions (median 3, $P < 0.001$,

187 Mann–Whitney U test). Moreover, we found that these mutations could potentially
188 impact the binding of different classes of antibodies (by analyzing the deep mutational
189 scanning data [22], Figure 3B), which was classified by the location and conformation
190 of antibody binding [23], suggesting that the therapeutic strategy of antibody cocktails
191 may also be affected.

192 **Obscure evolutionary trajectory of the Omicron variant**

193 In addition to the 61 shared mutations, some private mutations were identified in
194 different individuals, ranging from one to three, indicating relatively low population
195 diversity at the time of sampling (**Figure 4A**). Meanwhile, no obvious clusters were
196 found in the phylogenetic tree, suggesting that the Omicron variant was still in the
197 early transmission stage during sampling. The time to the most recent common
198 ancestor (TMRCA) was estimated to be in the middle of October 2021 (95% highest
199 density interval: October 7 to 20).

200 To screen for the possible predecessors of the Omicron variant, the 108 Omicron
201 sequences were used as queries to look for the closest sequences in the public
202 database, which included more than 5 million sequences released before November 1,
203 2021. We found three closest sequences to the queries, which differed by 53-56
204 nucleotides from the Omicron genomes. The three sequences were from lineage B.1.1
205 and collected between March and June 2020. They all had eight mutations relative to
206 Wuhan-Hu-1, and seven of the mutations were shared among them (Figure 4A). The
207 large number of differences suggests that the Omicron lineage was separated from
208 other lineages a long time ago and has never been sequenced since then. This is an
209 uncommon situation considering more than 5 million genomes have been sequenced
210 in over 180 countries and regions. The distribution of the number of differences
211 between all sequences in the public database and their closest sequences showed that
212 53-56 is approximately three times higher than the maximum number of differences
213 observed in the database (20 when at least three sequences were required to eliminate
214 the influence of sequencing or assembly errors, Figure 4B), again emphasizing the
215 distinctiveness of the Omicron variant.

216 Most Omicron lineage-specific mutations (52/54) were identified in public
217 databases (Figure 4C). However, they were unlikely to be presented in one sequence
218 by chance. First, over half of the mutations were rarely detected in the populations,

219 i.e., 33 mutations were detected in less than 1000 samples out of five million
220 sequences (16 mutations were detected in less than 100 samples). Second, the
221 mutation rate (represented by the occurrence number of mutations on the phylogenetic
222 tree) was extremely low for 13 of the mutations (occurring only once in the evolution
223 of SARS-CoV-2, mutation rate = 1). Third, the linkage disequilibrium between these
224 mutations was low, and only four mutation pairs had r^2 greater than 0.8. Moreover, we
225 further examined whether any combination of these mutations appeared in the
226 database and found that the maximum number of mutations in the same genome was
227 six. Therefore, the evolutionary trajectory of the Omicron lineage cannot be resolved
228 by the current genome data.

229 **Discussion**

230 The unique genome features of the Omicron variant make it the most special
231 SARS-CoV-2 variant to date. The excess number of nonsynonymous mutations in the
232 Spike gene implies that the Omicron variant might evolve under selection pressure,
233 which may come from antibodies or adaptation to new hosts. It is speculated that it
234 may have been incubated in a patient chronically infected with SARS-CoV-2, e.g.,
235 HIV patients with immunocompromising conditions. This hypothesis was supported
236 by the accelerated viral evolution observed in immunocompromised patients and has
237 been previously proposed to explain how the Alpha variant was generated [24,25]
238 ([https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-co](https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-co-v-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563)
239 [v-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563](https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-co-v-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563)). If this
240 hypothesis is true for the Omicron variant, we suspect that the original virus that
241 infected the patient might still be missing in the database because the current closest
242 sequences were circulating in population one and a half years ago, the time was too
243 long, even for a chronic infection. Another hypothesis involves a spillover from
244 humans to animals and spills back from animals to humans; such a process has been
245 proposed to be possible in mink [26]. Interestingly, a recent study proposed that the
246 progenitor of the Omicron variant seemed to have evolved in mice for some time
247 before jumping back into humans [27]. The binding affinity test between the Omicron
248 RBD and animal ACE2 may help to test this hypothesis. A third hypothesis is that the
249 virus split with other variants a long time ago and transmitted cryptically in the
250 population. Since viral genome surveillance is poor in many countries, it is difficult to
251 reject this hypothesis, which again underscores the importance of strengthening viral

surveillance on a global scale. Moreover, a hypothesis of acquisition by recombination between different variants is unlikely since the components that make up the Omicron genome could not be found in the current SARS-CoV-2 database, and of course, we cannot reject the possibility that the Omicron genome consists of a combination of components that have not been sequenced. More discussion of the possible origin of the Omicron variant can be found in other studies [28].

Benefiting from the establishment of the viral genome surveillance network and extensive research on the function of viral mutations, it took less than a week to designate the new VOC Omicron since the first identification of its genome, which is much faster than the designation of previous VOCs. However, it will still take several months to verify the risk of the new VOC. There have been over 200,000 new infections per day in the past year. Undoubtedly, we will face more mutant variants in the future, which may result in significant changes in transmissibility, infectivity, and pathogenicity. Unfortunately, it is still impossible to predict the evolutionary direction of the viral genome; hence, we have no hint at what the next VOC will be. To enhance the ability to rapidly respond to the emergence of new VOCs, we should further strengthen genome surveillance on a worldwide scale and develop experimental and computational methods for rapid and high-throughput resolution of mutational functions.

Materials and methods

Data collection

The sequencing data were retrieved from SRA database in NCBI (BioProject: PRJNA784038), which was generated by the Network for Genomic Surveillance in South Africa (NGS-SA) [2,6]. In total 211 samples were downloaded on November 30, 2021 (Table S1). The virus lineage was assigned by Pango [29], four samples that cannot be assigned to the Omicron lineage were discarded. All the remaining 207 samples were assigned to Omicron BA.1.

Quality control and mutation detection

Quality control and adaptor trimming were performed by FASTP [30]. The resultant reads were mapped to Wuhan-Hu-1 (NC_045512.2) using minimap2 (-ax sr) [31].

283 Primer alignment and trimming were performed by the align_trim function from Artic
284 (https://artic-tools.readthedocs.io/en/latest/commands/#align_trim). The mpileup file
285 and the read count file were generated by SAMtools [32] and Varscan2 [33]. The
286 consensus sequence was obtained using the following criteria: 1) depth \geq 5-fold; 2)
287 frequency of the major allele \geq 70%.

288 **Sequence depth analysis**

289 The sequencing depth was calculated for each nonoverlapping window with a size of
290 100 bp, except for the last window, which ranged from 29801 to 29880 bp. The fold
291 change of each primer region was calculated by the sum of the depth of all samples in
292 this region divided by the expected value (assuming no differences among regions).

293 **Identification of epitope regions on the Spike Protein**

294 We downloaded the structures of 182 protein complexes of antibodies that bound to
295 the SARS-CoV-2 *Spike* or its receptor-binding domain (RBD) or N-terminal
296 domain (NTD) from the Protein Data Bank (all structures available before August 8,
297 2021, www.rcsb.org). The residues in the Spike protein involved in binding to
298 antibodies were identified by a distance of less than 4.5 Å between two counterparts
299 in which van der Waals interactions occur. Deep mutational scanning results were
300 obtained from https://jbloomlab.github.io/SARS2_RBD_Ab_escape_maps/, which
301 includes information on sites in the SARS-CoV-2 RBD where mutations reduce
302 binding by antibodies/sera [22]. The escape score at each position was calculated as
303 the mean of the scores of all antibodies belonging to the same class.

304 **Construction of the phylogenetic tree**

305 The amino acid sequences were converted from nucleotide sequences using MEGA-X
306 (10.1.8) [34]. Phylogenetic construction was performed by IQ-TREE (1.6.12) [35].
307 The GTR+F model was used for nucleotide sequences, while the Blosum62 model
308 was used for amino acid sequences.

309 **TMRCA estimation**

310 The estimation of the time to the most recent common ancestor (TMRCA) and
311 mutation rate was performed by BEAST (2.6.4) [36] using 108 sequences collected
312 between November 13, 2021, and November 23, 2021. The HKY85 nucleotide
313 substitution model and strict molecular clock were used.

314 Search for the closest sequences in the database

315 The distance of two SARS-CoV-2 sequences was represented by the mutation
316 difference, which was calculated by an online tool at National Genomics Data Center,
317 China National Center for Bioinformation
318 <https://ngdc.cncb.ac.cn/ncov/online/tool/genome-tracing/?lang=en>. Publicly available
319 SARS-CoV-2 sequences were downloaded from the GISAID, NCBI, and RCoV19
320 databases (November 1, 2021) [1,37].

321 Calculation of linkage disequilibrium

322 The r^2 statistic was used to measure the strength of the linkage disequilibrium
323 between each pair of mutations [38]. The calculation of linkage disequilibrium was
324 based on all unique haplotypes from the public database.

325

326 CRediT author statement

327 **Wentai Ma:** Methodology, Formal analysis, and Writing. **Jing Yang:** Methodology,
328 Formal analysis, and Writing. **Haoyi Fu:** Methodology, Formal analysis. **Chao Su:**
329 Methodology, Formal analysis. **Caixia Yu:** Resources. **Qihui Wang:** Methodology.
330 **Ana Tereza Ribeiro de Vasconcelos:** Resources, Writing. **Georgii A. Bazykin:**
331 Resources, Writing. **Yiming Bao:** Methodology, Writing. **Mingkun Li:**
332 Conceptualization, Methodology, Supervision, Writing. All authors have read and
333 approved the final manuscript.

334 Acknowledgments

335 We thank Dr. Jennifer Giandhari, Dr. Eduan Wilkinson, and Dr. Tulio de Oliveira
336 from Centre for Epidemic Response and Innovation (CERI), Stellenbosch University
337 for sharing the data and workflow, GISAID and associated laboratories and
338 researchers for the shared sequence information, Dr. Aiping Wu from Suzhou Institute
339 of Systems Medicine, Chinese Academy of Medical Sciences, for help in the
340 phylogenetic tree analysis. This study was funded by the National Natural Science
341 Foundation of China (Grant No. 82161148009), the Strategic Priority Research
342 Program of Chinese Academy of Sciences, China (Grant No. XDB38030400), The
343 Capital Health Development and Research of Special (Grant No. 2021-1G-3012),
344 Conselho Nacional de desenvolvimento Científico e Tecnológico (CNPq) -

345 NGS-BRICS - n°: 440931/2020-7, and Russian Foundation for Basic Research
346 (RFBR) (Grant No. 20-54-80014.

347 **Competing interests**

348 The authors declare that they have no competing interests.

349 **ORCID**

350 0000-0003-1931-8687 (Wentai Ma)

351 0000-0002-3934-7883 (Jing Yang)

352 0000-0001-9696-5445 (Haoyi Fu)

353 0000-0002-5824-7968 (Chao Su)

354 0000-0002-3882-9979 (Caixia Yu)

355 0000-0003-3768-0401 (Qihui Wang)

356 0000-0002-4632-2086 (Ana Tereza Ribeiro de Vasconcelos)

357 0000-0003-2334-2751 (Georgii A. Bazykin)

358 0000-0002-9922-9723 (Yiming Bao)

359 0000-0003-1041-1172 (Mingkun Li)

360 **References**

- 361 [1] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's
362 innovative contribution to global health. *Glob challenges (Hoboken, NJ)*
363 2017;1: 33–46.
- 364 [2] Viana R, Moyo S, Amoako DG, et al. Rapid epidemic expansion of the
365 SARS-CoV-2 Omicron variant in southern Africa. *medRxiv*
366 2021;2021.12.19.21268028.
- 367 [3] Lista MJ, Winstone H, Wilson HD, et al. The P681H mutation in the Spike
368 glycoprotein confers Type I interferon resistance in the SARS-CoV-2 alpha
369 (B.1.1.7) variant. *bioRxiv*. Epub ahead of print 2021. DOI:
370 10.1101/2021.11.09.467693.
- 371 [4] Cele S, Jackson L, Khoury DS, et al. Omicron extensively but incompletely
372 escapes Pfizer BNT162b2 neutralization. *Nature*. Epub ahead of print 2021.
373 DOI: doi.org/10.1038/d41586-021-03824-5.
- 374 [5] Pulliam J. Increased risk of SARS-CoV-2 reinfection associated with
375 emergence of the Omicron variant in South Africa. *medRxiv* 2021;1–43.

- 376 [6] Wilkinson E, Giovanetti M, Tegally H, et al. A year of genomic surveillance
377 reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science* 2021;374:
378 423–431.
- 379 [7] Zhang Z, Li J, Zhao X-Q, et al. KaKs_Calculator: calculating Ka and Ks
380 through model selection and model averaging. *Genomics Proteomics
381 Bioinformatics* 2006;4: 259–263.
- 382 [8] Temmam S, Salazar EB, Munier S, et al. Coronaviruses with a
383 SARS-CoV-2-like receptor- binding domain allowing ACE2-mediated entry
384 into human cells isolated from bats of Indochinese peninsula. *Res Sq*. Epub
385 ahead of print 2021. DOI: 10.21203/rs.3.rs-871965/v1.
- 386 [9] Zhou P, Yang X-L, Wang X-G, et al. A pneumonia outbreak associated with a
387 new coronavirus of probable bat origin. *Nature* 2020;579: 270–273.
- 388 [10] Liu P, Jiang J-Z, Wan X-F, et al. Are pangolins the intermediate host of the
389 2019 novel coronavirus (SARS-CoV-2)? *PLoS Pathog* 2020;16: e1008421.
- 390 [11] Lam TT-Y, Jia N, Zhang Y-W, et al. Identifying SARS-CoV-2-related
391 coronaviruses in Malayan pangolins. *Nature* 2020;583: 282–285.
- 392 [12] Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human
393 respiratory disease in China. *Nature* 2020;579: 265–269.
- 394 [13] Wang Q, Zhang Y, Wu L, et al. Structural and Functional Basis of
395 SARS-CoV-2 Entry by Using Human ACE2. *Cell* 2020;181: 894-904.e9.
- 396 [14] Laffebber C, de Koning K, Kanaar R, et al. Experimental Evidence for
397 Enhanced Receptor Binding by Rapidly Spreading SARS-CoV-2 Variants. *J
398 Mol Biol* 2021;433: 167058.
- 399 [15] Liu Y, Liu J, Plante KS, et al. The N501Y spike substitution enhances
400 SARS-CoV-2 infection and transmission. *Nature*. Epub ahead of print
401 November 2021. DOI: 10.1038/s41586-021-04245-0.
- 402 [16] Meng B, Kemp SA, Papa G, et al. Recurrent emergence of SARS-CoV-2 spike
403 deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep* 2021;35:
404 109292.
- 405 [17] Lubinski B, Fernandes MH V, Frazier L, et al. Functional evaluation of the
406 P681H mutation on the proteolytic activation the SARS-CoV-2 variant B . 1 .
407 1 . 7 (Alpha) spike. *bioRxiv* 2021;1–28.
- 408 [18] Wu H, Xing N, Meng K, et al. Nucleocapsid mutations R203K/G204R increase
409 the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host Microbe*
410 2021;29: 1788-1801.e6.
- 411 [19] Hastie KM, Li H, Bedinger D, et al. Defining variant-resistant epitopes targeted
412 by SARS-CoV-2 antibodies: A global consortium study. *Science* 2021;374:
413 472–478.
- 414 [20] Wilhelm A, Widera M, Grikscheit K, et al. Reduced Neutralization of

- 415 SARS-CoV-2 Omicron Variant by Vaccine Sera and Monoclonal Antibodies.
416 *medRxiv*. Epub ahead of print 2021. DOI: 10.1101/2021.12.07.21267432.
- 417 [21] Cao Y, Wang J, Jian F, et al. Omicron escapes the majority of existing
418 SARS-CoV-2 neutralizing antibodies. *Nature*. Epub ahead of print 2021. DOI:
419 doi.org/10.1038/d41586-021-03796-6.
- 420 [22] Greaney AJ, Starr TN, Barnes CO, et al. Mapping mutations to the
421 SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat*
422 *Commun* 2021;12: 4196.
- 423 [23] Barnes CO, Jette CA, Abernathy ME, et al. SARS-CoV-2 neutralizing antibody
424 structures inform therapeutic strategies. *Nature* 2020;588: 682–687.
- 425 [24] Choi B, Choudhary MC, Regan J, et al. Persistence and Evolution of
426 SARS-CoV-2 in an Immunocompromised Host. *The New England journal of*
427 *medicine* 2020;383: 2291–2293.
- 428 [25] Kemp SA, Collier DA, Datir RP, et al. SARS-CoV-2 evolution during
429 treatment of chronic infection. *Nature* 2021;592: 277–282.
- 430 [26] Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, et al. Transmission of
431 SARS-CoV-2 on mink farms between humans and mink and back to humans.
432 *Science* 2021;371: 172–177.
- 433 [27] Wei C, Shan K-J, Wang W, et al. Evidence for a mouse origin of the
434 SARS-CoV-2 Omicron variant. *J Genet Genomics*. Epub ahead of print
435 Desember 2021. DOI: 10.1016/j.jgg.2021.12.003.
- 436 [28] Kupferschmidt K. Where did “weird” Omicron come from? *Science (New York,*
437 *N.Y.)* 2021;374: 1179.
- 438 [29] Rambaut A, Holmes EC, O’Toole Á, et al. A dynamic nomenclature proposal
439 for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*
440 2020;5: 1403–1407.
- 441 [30] Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ
442 preprocessor. *Bioinformatics* 2018;34: i884–i890.
- 443 [31] Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
444 2018;34: 3094–3100.
- 445 [32] Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format
446 and SAMtools. *Bioinformatics* 2009;25: 2078–2079.
- 447 [33] Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and
448 copy number alteration discovery in cancer by exome sequencing. *Genome Res*
449 2012;22: 568–576.
- 450 [34] Kumar S, Stecher G, Li M, et al. MEGA X: Molecular Evolutionary Genetics
451 Analysis across Computing Platforms. *Mol Biol Evol* 2018;35: 1547–1549.
- 452 [35] Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New Models and
453 Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol*

- 454 *Evol* 2020;37: 1530–1534.
- 455 [36] Bouckaert R, Vaughan TG, Barido-Sottani J, et al. BEAST 2.5: An advanced
456 software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*
457 2019;15: e1006650.
- 458 [37] Song S, Ma L, Zou D, et al. The Global Landscape of SARS-CoV-2 Genomes,
459 Variants, and Haplotypes in 2019nCoV. *Genomics Proteomics*
460 *Bioinformatics* 2020;18: 749–759.
- 461 [38] Slatkin M. Linkage disequilibrium--understanding the evolutionary past and
462 mapping the medical future. *Nat Rev Genet* 2008;9: 477–485.

463

464

465 **Figure legends**

466 **Figure 1 Sequence enrichment efficiency of the Omicron variant using different** 467 **protocols**

468 **A.** Distribution of the sequencing depth of the Omicron variant. The average
469 sequencing depth is shown for each non-overlapping window of 100 bp after
470 normalization by the total number of reads in the sample. The primers affected by the
471 mutations in Omicron are labeled on top of the figure. **B.** The efficiency of each
472 primer in amplifying the nucleic acids of the Omicron variant. The color represents
473 the fold change of enrichment efficiency, calculated by the sum of the depth of all
474 samples in this region divided by the expected value (assuming no differences among
475 regions). The overlapping region of adjacent primers was excluded from the analysis.
476 The Omicron mutations that located in the region where primers anneal to are labeled
477 on the right of the primer ID.

478

479 **Figure 2. Mutations in the Omicron genome and its evolutionary relationship** 480 **with other variants and SARS-like coronaviruses.**

481 **A.** Summary of mutations in the Omicron genome. Each row represents a mutation,
482 and changes in nucleotides and amino acids are marked on two sides of the heatmap.
483 Mutations located in the sites critical for viral binding to the human receptor
484 angiotensin-converting enzyme 2 (ACE2) are marked in red [13]. Mutations observed
485 in the *Spike* gene of other variants of concern (VOCs) are listed on the left of the
486 heatmap. **B.** Phylogenetic tree of five VOCs and SARS-like coronaviruses based on

the nucleotide sequences. **C.** Phylogenetic tree of five VOCs and SARS-like coronaviruses based on the amino acid sequences in the *Spike* gene. **D.** Phylogenetic tree of five VOCs and SARS-like coronaviruses based on the amino acid sequences in the RBD region. Two bat coronaviruses (Bat BANAL-20-52 and Bat RaTG3) whose genomes were most similar to SARS-CoV-2 [8,9], two pangolin coronaviruses (Pangolin MP789 and Pangolin GXP5L) [10,11], and sequences of other recently collected VOCs (EPI_ISL_6141707, EPI_ISL_6774033, EPI_ISL_6898988, EPI_ISL_6585201 for the Alpha, Beta, Gamma, and Delta variants, respectively. All sequences were collected in November 2021, and those collected in South Africa were preferred) were included in the analysis of the phylogenetic tree. The Wuhan-Hu-1 sequence is shown as the outgroup of the tree for better visualization [12]. The number of mutations relative to Wuhan-Hu-1 is listed on the right of the tree. Insertions or deletions of multiple bases were considered as a single mutation.

500

Figure 3. Distribution of the Omicron mutations at the antibody binding positions. **A.** The number of binding antibodies at the Omicron mutation sites. **B.** The escape score of the Omicron mutations estimated from deep mutational scanning. The escape score for each position was calculated as the mean of the scores of all antibodies belonging to the same class. All Omicron mutations were labeled on the figure.

507

Figure 4. Evolutionary features of the Omicron variant.

A. The phylogenetic tree of 108 Omicron sequences and their closest sequences in the database. Wuhan-Hu-1 is shown as the outgroup of the tree. The three closest sequences belonging to lineage B.1.1 are highlighted in orange. Nonsynonymous mutations are marked in red. **B.** The distribution of the number of differences between all haplotypes (nonredundant sequences) in the public database and their closest sequences. The minimum number of sequences required for a valid haplotype was set to 3. **C.** Correlation between different Omicron mutations. Only 54 Omicron lineage-specific mutations were included in the analysis. The color in the heatmap represents the linkage disequilibrium coefficient (r^2) between mutations. The mutation rate and the number of sequences in the public database that possess the same

519 mutation are labeled on the left and bottom of the heatmap, respectively. A cross is
520 labeled if the mutation was not observed in the public database.

521

522 **Supplementary material**

523 **Figure S1. The correlation of the major allele frequency between the Illumina**
524 **protocol and the GridION protocol.**

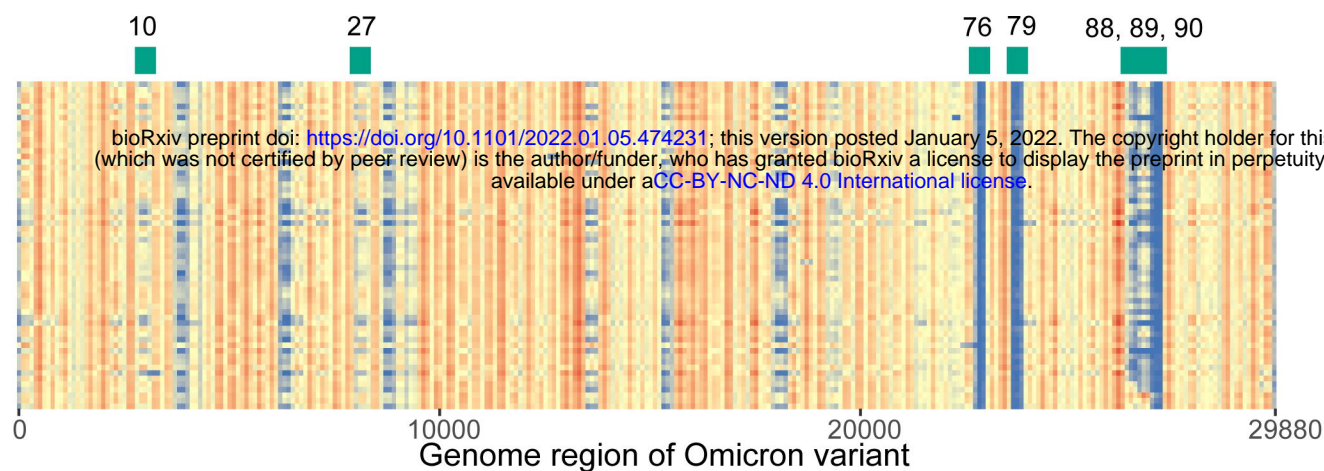
525 Each dot represents a mutation (major allele frequency $\geq 70\%$).

526 **Figure S2. The distribution of the Omicron mutations on the structure of the**
527 **Spike protein (left) and the RBD region (right).**

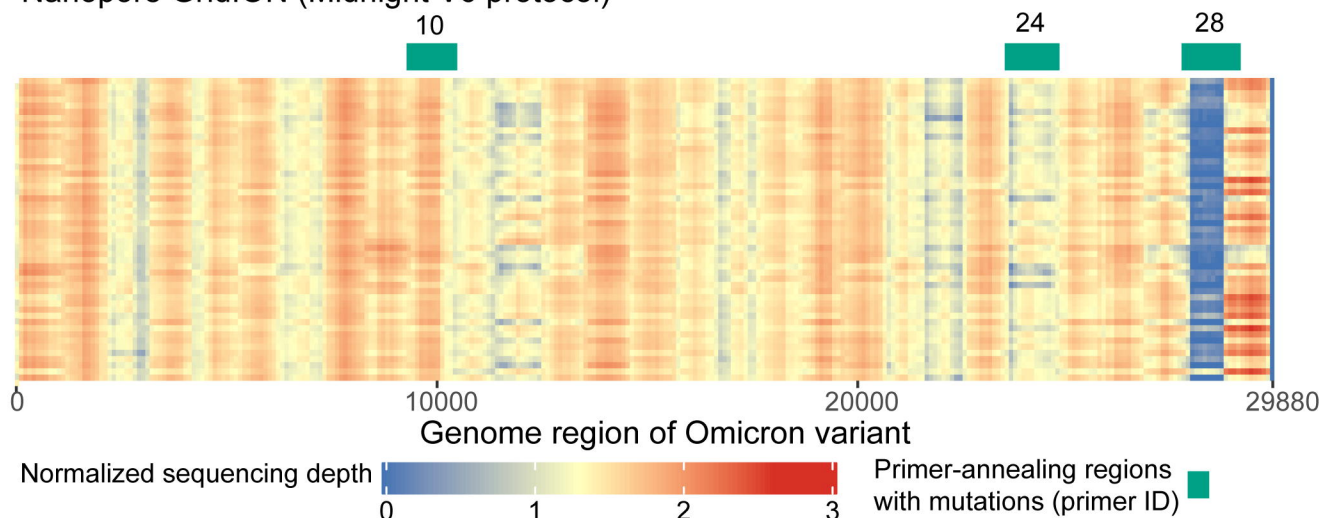
528 **Table S1. The information of the data used in the study.**

A

Illumina MiSeq (ARTIC V4 protocol)



Nanopore GridION (Midnight V6 protocol)



B

Nanopore GridION
(Midnight V6 protocol)

Illumina MiSeq (ARTIC V4 protocol)

Primer ID: region [mutation]

1: 30-1205
2: 1100-2266
3: 2153-3257
4: 3144-4262
5: 4167-5359
6: 5257-6380
7: 6283-7401
8: 7298-8385
9: 8253-9400
10: 9303-10451 [10449]
11: 10343-11469
12: 11372-12560
13: 12450-13621
14: 13509-14641
15: 14540-15735
16: 15608-16720
17: 16624-17754
18: 17622-18706
19: 18596-19678
20: 19574-20698
21: 20553-21642
22: 21532-22612
23: 22511-23631
24: 23518-24736 [23525]
25: 24633-25790
26: 25690-26857
27: 26744-27894
28: 27784-29007 [27807]
29: 28677-29790

Primer ID: region [mutation]

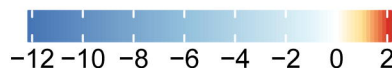
1: 25-431
2: 324-727
3: 644-1044
4: 944-1362
5: 1245-1650
6: 1540-1948
7: 1851-2250
8: 2154-2571
9: 2483-2885
10: 2826-3210 [2832]
11: 3078-3492
12: 3390-3794
13: 3683-4093
14: 3992-4409
15: 4312-4710
16: 4620-5017
17: 4923-5331
18: 5230-5643
19: 5561-5957
20: 5867-6272
21: 6184-6582
22: 6478-6885
23: 6747-7148
24: 7057-7467
25: 7381-7770
26: 7672-8092
27: 7997-8395 [8393]
28: 8304-8714
29: 8596-9013
30: 8919-9329
31: 9168-9564
32: 9470-9866
33: 9782-10176

Primer ID: region [mutation]

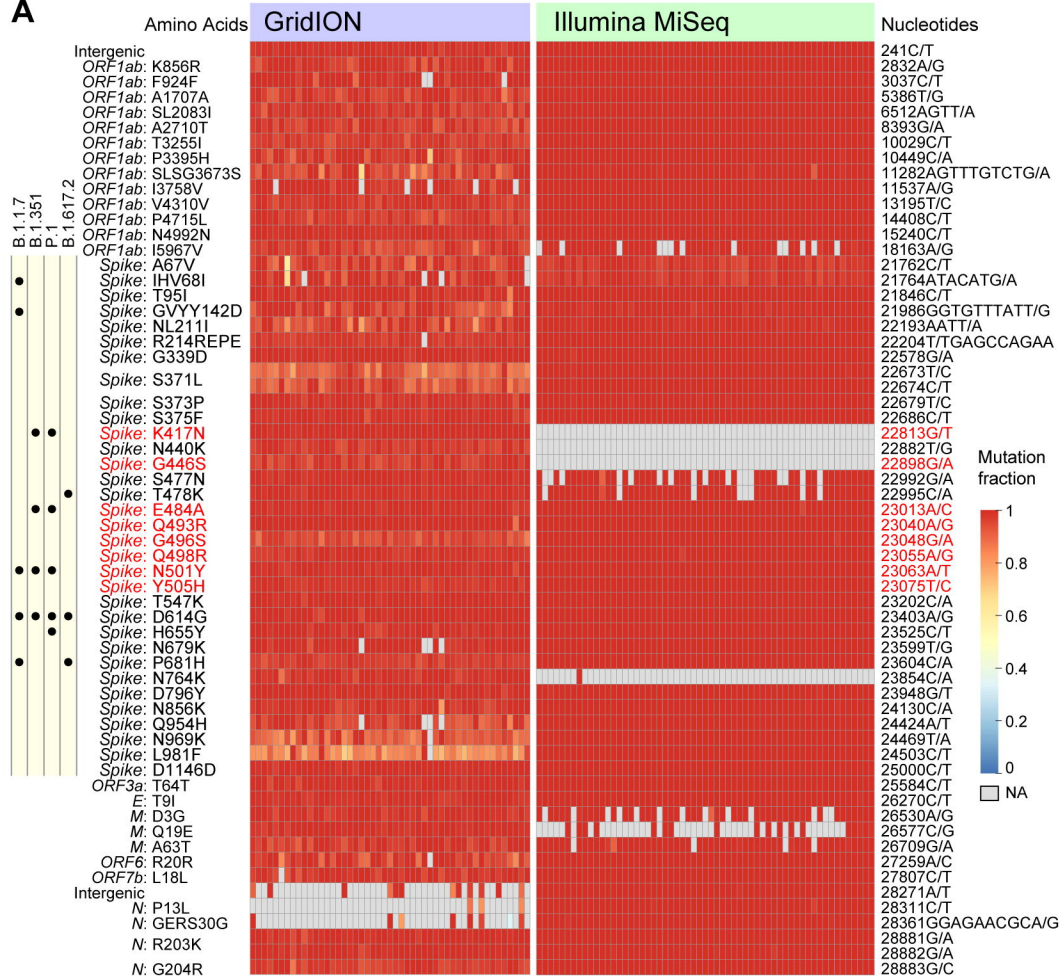
34: 10076-10491
35: 10393-10810
36: 10713-11116
37: 11000-11414
38: 11305-11720
39: 11624-12033
40: 11937-12339
41: 12234-12643
42: 12519-12920
43: 12831-13240
44: 13124-13528
45: 13463-13859
46: 13752-14144
47: 14045-14457
48: 14338-14743
49: 14647-15050
50: 14953-15358
51: 15214-15619
52: 15535-15941
53: 15855-16260
54: 16112-16508
55: 16386-16796
56: 16692-17105
57: 16986-17405
58: 17323-17711
59: 17615-18022
60: 17911-18328
61: 18244-18652
62: 18550-18961
63: 18869-19277
64: 19183-19586
65: 19485-19901
66: 19810-20216

Primer ID: region [mutation]

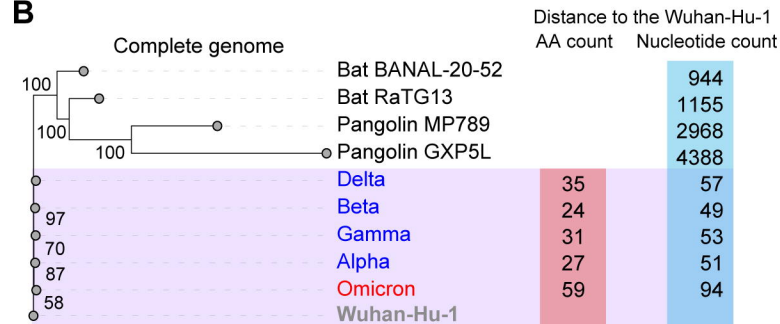
67: 20090-20497
68: 20377-20792
69: 20677-21080
70: 20988-21387
71: 21294-21700
72: 21532-21933
73: 21865-22274
74: 22091-22503
75: 22402-22805
76: 22648-23057 [22673]
77: 22944-23351 [22674]
78: 23219-23635 [23040]
79: 23553-23955 [23048]
80: 23853-24258 [23055]
81: 24171-24567
82: 24426-24836
83: 24750-25150
84: 25051-25461
85: 25331-25740
86: 25645-26050
87: 25951-26360
88: 26255-26661 [26270]
89: 26564-26979 [26577]
90: 26873-27283 [27259]
91: 27152-27560
92: 27447-27855
93: 27700-28104
94: 27996-28416
95: 28190-28598
96: 28512-28914
97: 28827-29227
98: 29136-29534
99: 29452-29854

Log₂FC of sequencing depth

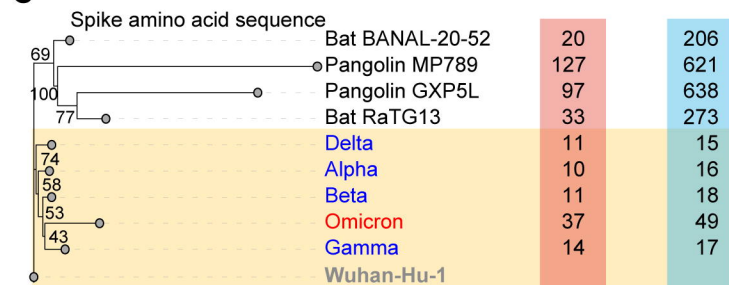
A



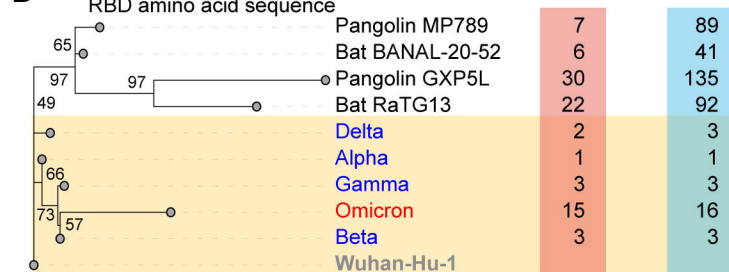
B

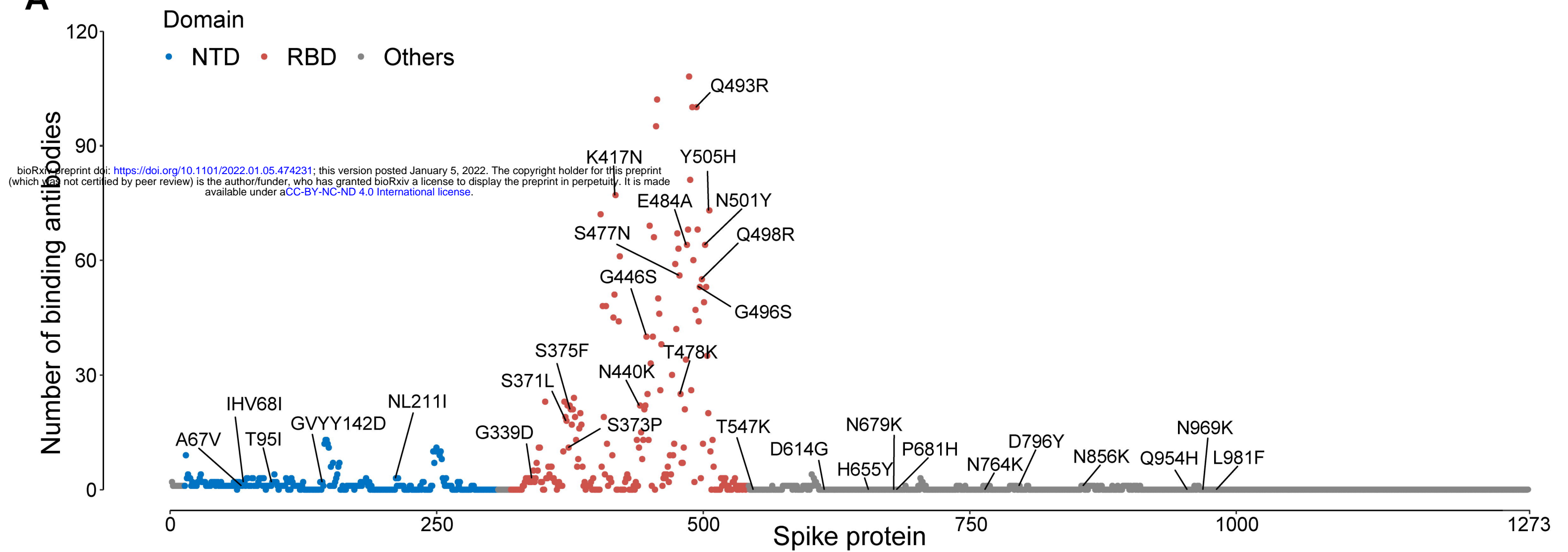
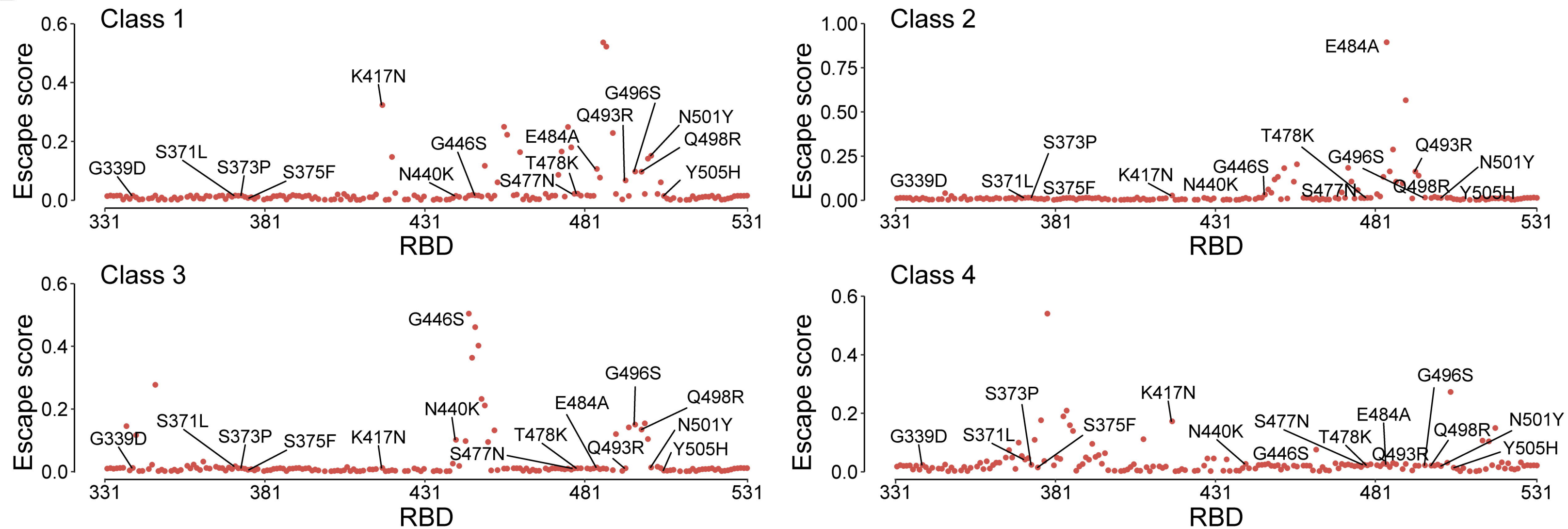


C

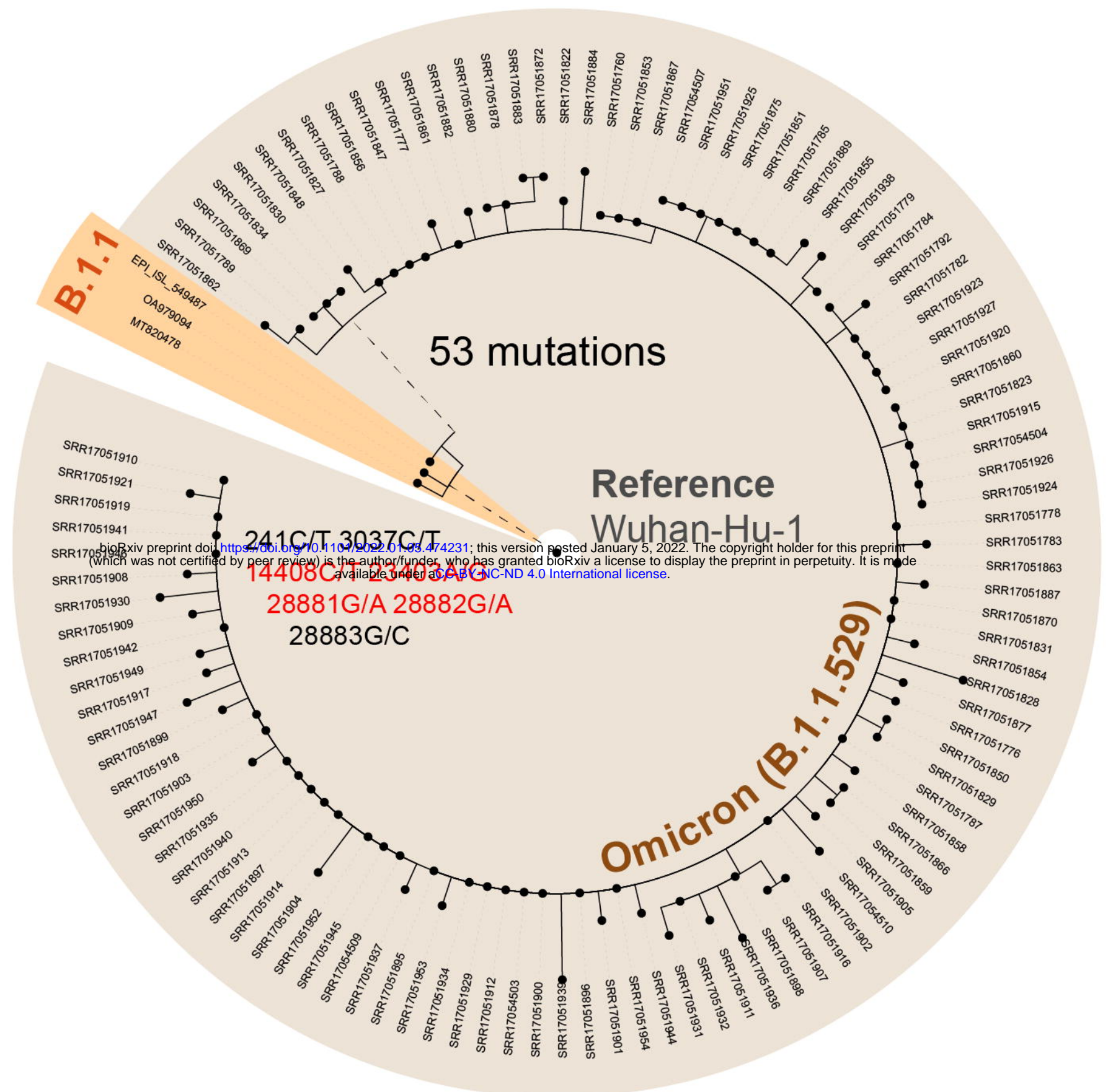


D

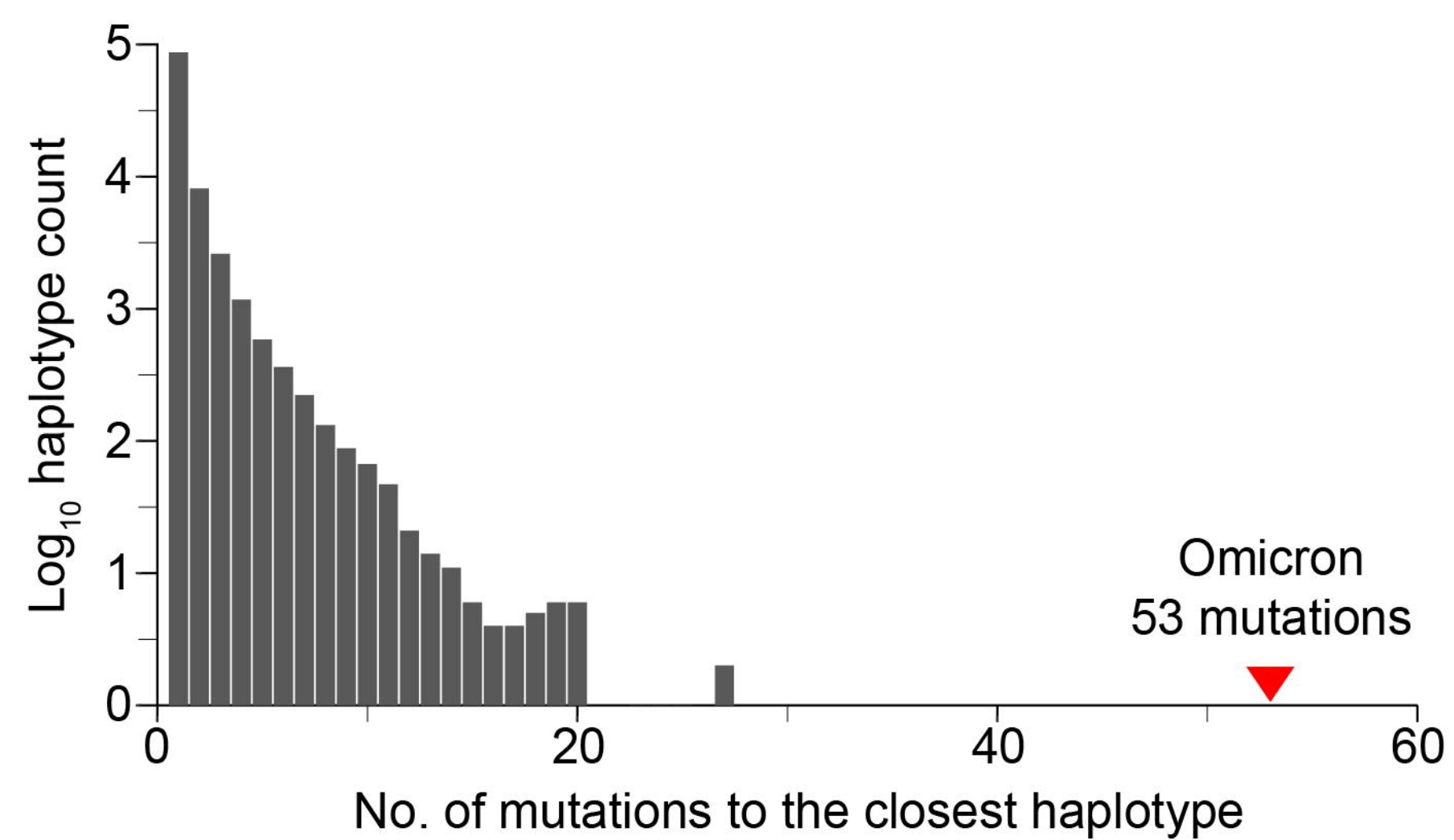


A**B**

A



B



C

