

1 **The global population of SARS-CoV-2 is composed of six major subtypes**

2

3 Ivair José Morais Júnior<sup>a</sup>, Richard Costa Polveiro<sup>b</sup>, Gabriel Medeiros Souza<sup>c</sup>, Daniel Inserra  
4 Bortolin<sup>c</sup>, Flávio Tetsuo Sasaki<sup>d</sup>, Alison Talis Martins Lima<sup>c\*</sup>

5

6 <sup>a</sup>Departamento de Fitopatologia/Universidade de Brasília, Brasília, DF 70910-900, Brazil

7 <sup>b</sup>Departamento de Veterinária/Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brazil

8 <sup>c</sup>Instituto de Ciências Agrárias/Universidade Federal de Uberlândia, Uberlândia, MG 38410-337,  
9 Brazil

10 <sup>d</sup>Instituto de Biotecnologia/Universidade Federal de Uberlândia, Monte Carmelo, MG 38500-000,  
11 Brazil

12

13 \*E-mail: [atmlima@ufu.br](mailto:atmlima@ufu.br)

14

15 **Abstract**

16

17 The World Health Organization characterized the COVID-19 as a pandemic in March 2020, the second  
18 pandemic of the 21<sup>st</sup> century. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a  
19 positive-stranded RNA betacoronavirus of the family *Coronaviridae*. Expanding virus populations, as  
20 that of SARS-CoV-2, accumulate a number of narrowly shared polymorphisms imposing a  
21 confounding effect on traditional clustering methods. In this context, approaches that reduce the  
22 complexity of the sequence space occupied by the SARS-CoV-2 population are necessary for a robust  
23 clustering. Here, we proposed the subdivision of the global SARS-CoV-2 population into sixteen well-  
24 defined subtypes by focusing on the widely shared polymorphisms in nonstructural (*nsp3*, *nsp4*, *nsp6*,  
25 *nsp12*, *nsp13* and *nsp14*) cistrons, structural (*spike* and *nucleocapsid*) and accessory (*ORF8*) genes.  
26 Six virus subtypes were predominant in the population, but all sixteen showed amino acid  
27 replacements which might have phenotypic implications. We hypothesize that the virus subtypes  
28 detected in this study are records of the early stages of the SARS-CoV-2 diversification that were  
29 randomly sampled to compose the virus populations around the world, a typical founder effect. The  
30 genetic structure determined for the SARS-CoV-2 population provides substantial guidelines for  
31 maximizing the effectiveness of trials for testing the candidate vaccines or drugs.

## 32 Main

33 In December 2019, a local pneumonia outbreak of initially unknown etiology was detected in  
34 Wuhan (Hubei, China) and quickly determined to be caused by a novel coronavirus<sup>1</sup>, named Severe  
35 acute respiratory syndrome coronavirus 2 (SARS-CoV-2)<sup>2</sup> and the disease as COVID-19<sup>3</sup>. SARS-  
36 CoV-2 is classified in the family *Coronaviridae*, genus *Betacoronavirus*, which comprises enveloped,  
37 positive stranded RNA viruses of vertebrates<sup>2</sup>. Two-thirds of SARS-CoVs genome is covered by the  
38 ORF1ab, that encodes a large polypeptide which is cleaved into 16 nonstructural proteins (NSPs)  
39 involved in replication-transcription in vesicles from endoplasmic reticulum (ER)-derived  
40 membranes<sup>4,5</sup>. The last third of the virus genome encodes four essential structural proteins: spike (S),  
41 envelope (E), membrane (M), nucleocapsid (N) and several accessory proteins that interfere with the  
42 host innate immune response<sup>6</sup>.

43 Populations of RNA viruses evolve rapidly due to their large population sizes, short generation  
44 times, and high mutation rates, this latter being a consequence of the RNA-dependent RNA  
45 polymerase (RdRP) which lacks the proofreading activity<sup>7</sup>. In fact, virus populations are composed of  
46 a broad spectrum of closely related genetic variants resembling one or more master sequences<sup>8-10</sup>.  
47 Mutation rates inferred for SARS-CoVs are considered moderate<sup>11,12</sup> due to the independent  
48 proofreading activity<sup>13</sup>. However, the large SARS-CoV genomes (from 27 to 31 kb)<sup>14</sup> provide to them  
49 the ability to explore the sequence space<sup>15</sup>. In order to better understand the diversification of SARS-  
50 CoV-2 genomes during the pandemics (from December 2019 to March 25, 2020), we applied a simple,  
51 but robust approach to reduce the complexity of the sequence space occupied by the virus population  
52 by detecting its widely shared polymorphisms.

53 The 767 SARS-CoV-2 genomes with high sequencing coverage obtained from GISAID  
54 (<https://www.gisaid.org/>) and GenBank were clustered into 593 haplotypes (Table S1). We conducted  
55 a fine-scale sequence variation analysis on the 593 genomes-containing alignment by calculating the  
56 nucleotide diversity ( $\pi$ ) using a sliding window and step size of 300 and 20 nucleotides, respectively  
57 (multiple sequence alignments generated in this study are available from the authors upon request).  
58 Such an approach allows to identify genomic regions of increased genetic variation from polymorphic  
59 sites harboring two or more distinct nucleotide bases. Noticeably, one or more large clusters of closely  
60 related sequences, when analyzed by this approach, show locally increased nucleotide diversity. We  
61 observed a contrasting distribution of the genetic variation across the full-length genomes of SARS-  
62 CoV-2 (Fig. 1) with eight segments showing increased genetic variation, arbitrarily defined as

63 nucleotide (nt) segments with  $\pi \geq 0.001$ . Seven out of eight segments (S) had about 280 nucleotides  
64 in length, corresponding approximately to the size of a single sliding window, except the S10 whose  
65 length was equivalent to two sliding windows (600 nt). To further investigate the diversification of  
66 segments with contrasting content of genetic variability, we constructed maximum likelihood (ML)  
67 phylogenetic trees and analyzed the diversification patterns of eight segments with higher (S2, 4, 6, 8,  
68 10, 12, 14 and 16), and nine with lower (S1, 3, 5, 7, 9, 11, 13, 15 and 17) content of genetic variation,  
69 respectively.

70 Although the data set was composed of hundreds of SARS-CoV-2 genomes sampled from  
71 around the world, in the S2-based tree we observed two clusters (Fig. S1a). Markedly, each cluster  
72 was composed of very closely related, if not identical, sequences. Therefore, the increased content of  
73 genetic variation at the S2 was a result of the inter-cluster sequence comparisons. Similar results were  
74 obtained for the other seven ML-trees based on segments with increased genetic variation (Fig. S1b -  
75 h). In contrast, the ML-trees based on segments with lower content of genetic variation did not show  
76 a consistent number of well-defined clusters (Fig. S2).

77 We mapped the polymorphic sites in segments with increased content of genetic variation  
78 responsible for the segregation of ML-trees into two well-defined clusters (Table 1). Only a few (from  
79 one to three) nt positions with polymorphisms shared by a number of SARS-CoV-2 genomes could be  
80 identified within each segment with increased genetic variation. These polymorphisms were  
81 henceforth referred to as ‘widely shared polymorphisms’ (WSPs), while the remaining nt positions in  
82 virus genomes were designed as ‘non widely shared polymorphisms’ (nWSPs).

83 We compared the topologies of the seventeen ML-trees by computing their pairwise distances  
84 followed by a multivariate analysis to group similar trees (Fig. 2). The seventeen trees were subdivided  
85 into seven groups, the largest one containing those nWSPs-containing segments-based trees (S1, 3, 5,  
86 7, 9, 11, 13, 15 and 17; Fig. 2, Group 7). Given the low content of genetic variation in these segments,  
87 the resulting trees were poorly resolved suggesting that such regions represent a wide mutant spectrum  
88 of narrowly shared polymorphisms. It is important to note that there are minor clusters in nWSPs-  
89 containing segments-based ML-trees, *e.g.*, in those for S1, S13 and S17. This is a consequence of our  
90 conservative threshold in which we focused on segments with  $\pi \geq 0.001$ . S1, S13 and S17 also show  
91 locally increased genetic variation with  $\pi$  higher than 0.0005 but lower than 0.001. For example,  
92 stretches 889 - 1,169; 1,409 - 1,509 (within the S1); 25,403 - 25,693 (S13); 29,538 - 29,610 (S17)  
93 (Fig. 1).

94 The S2, S8 and S12-based ML-trees (Fig. 2, Group 1) were considerably congruent and the  
95 nucleotides at their WSPs tended to co-segregate (UUG or CCA, Table 1), which results into two  
96 major subtypes of SARS-CoV-2. Reciprocally, the incongruency among the S4, 6, 10, 14 and 16-  
97 based trees (Fig. 2, Groups 2-6) suggests the segregation of nucleotides at their WSPs, which increases  
98 the possible combinations of virus genotypes.

99 Therefore, our approach reduced the complexity of the sequence space occupied by the SARS-  
100 CoV-2 genomes and provided a robust clustering solution based on the combination of 12 WSPs  
101 (Table 1) to identify the major viral genotypes spread worldwide (Table 2 and Table S2). The global  
102 population of SARS-CoV-2 is structured into six major subtypes (I - VI), comprising 578 out of 593  
103 (about 97.5%) isolates analyzed in this study. The Subtype I (N=132) was represented by the  
104 combination of the most frequent nucleotides at all WSPs, *i.g.*, the canonical genotype:  
105 CCGCCACAUGGG. The SARS-CoV-2 reference isolate (GISAID accession ID: EPI\_ISL\_402124,  
106 GenBank accession: MN908947) is a representative member of this subtype. Subtype IV (N=91) was  
107 represented by the combination of the most frequent nucleotides at eleven out of 12 WSPs  
108 (**CCUCCACAUGGG**; the most frequent nucleotides at each WSP are highlighted in bold and  
109 underlined). Subtypes V (N=74, **CUGCCACACGGG**), II (N=122, **UCGUCACGUGGG**), III  
110 (N=101, **CUGCUGUACGGG**) and VI (N=58, **UCGUCACGUAAAC**) were represented by the  
111 combination of the most frequent nucleotides at ten, nine, seven and six out of 12 WSPs, respectively.  
112 It is important to emphasize that the contrasting sample sizes (Subtypes I - VI vs. VII - XVI) are not  
113 necessarily associated with fitness variation and might be due to a sampling bias. For example, the  
114 three isolates composing the Subtype VIII showed a genotype (**CUGCCAUACGGG**) very similar to  
115 that of the canonical reference isolate. In addition, even though our data set was composed exclusively  
116 by genomes with high sequencing coverage, we cannot rule out that the virus subtypes X to XVI,  
117 which were represented by a single genome might be a consequence of poor sampling or sequencing  
118 errors.

119 The phylogenetic tree depicting all 593 SARS-CoV-2 haplotypes (Fig. 3) showed some  
120 geographical structure with two clusters: a smaller one comprised of isolates mostly sampled from  
121 Western hemisphere (Subtypes II, VI, IX, X and XI) and a larger one whose isolates were sampled  
122 from Western and Eastern hemispheres (I, III, IV, V, VII, VIII, XII, XIII, XIV, XV and XVI).

123 We hypothesize that our clustering method for the SARS-CoV-2 population could involve at  
124 some extent a biological context. Nine out of 12 WSPs led to amino acid replacements (Table 2), *e.g.*,

125 the WSP *nsp6*-[111] in nine SARS-CoV-2 subtypes led to a leucine at the aa residue#37 of the protein  
126 and a phenylalanine in seven other subtypes. NSP6 is an integral membrane protein that interferes in  
127 the autophagosome formation during the SARS-CoV infection. Additionally, in yeast two-hybrid  
128 experiments, NSP6 has been shown to interact with NSP3<sup>16</sup>. Some evidence demonstrates that NSP6  
129 protein limits the expansion of autophagosomes or, alternatively, might remove host proteins involved  
130 in inhibition of viral replication by activating autophagy from the ER<sup>17</sup>.

131 The WSP *nsp12*-[967] resulted in a proline in eleven subtypes of SARS-CoV-2 and a leucine  
132 in others five subtypes at the aa residue #323 of the NSP12 (RNA-dependent RNA polymerase, RdRP)  
133 protein. It is located at the Interface domain of RdRP of SARS-CoV-2, which is responsible for the  
134 connection between the nidovirus RdRP-associated nucleotidyltransferase domain (NiRAN) and the  
135 “Right hand” polymerase domain<sup>18</sup>. The S protein mediates viral entry into host cells by first binding  
136 to a receptor, angiotensin-converting enzyme 2 (ACE2), through the receptor-binding domain (RBD)  
137 in the S1 subunit and then fusing the viral and host membranes through the S2 subunit<sup>19–22</sup>. Sites of  
138 glycosylation are important for S protein folding<sup>23</sup>, affecting priming by host proteases<sup>24</sup> and might  
139 modulate antibody recognition<sup>25,26</sup>. The WSP *S*-[1,841] resulted in a glycine and an aspartate at the aa  
140 residue#614 of the S protein in six and ten subtypes of SAR-CoV-2, respectively. The replacement  
141 was mapped in the intermediate region between the S1 and S2 subunits. This WSP is near a  
142 glycosylation site (N616CT)<sup>27</sup>.

143 The WSP *ORF8*-[251] involved a non-synonymous mutation at the codon#84 encoding for  
144 leucine and serine in nine and seven subtypes, respectively. The SARS-CoV *ORF8* encodes for an  
145 ER-associated protein that induces the activation of ATF6, and this latter is an ER stress-regulated  
146 transcription factor that stimulates the production of chaperones<sup>28</sup>. In addition, the ORF8 protein has  
147 been demonstrated to induce apoptosis<sup>29</sup>. In SARS epidemics, the *ORF8* from different coronaviruses  
148 was targeted by a number of mutations and recombination events during transmission from animals to  
149 humans<sup>30</sup>.

150 Three WSPs mapped in the *N* gene led to two amino acid replacements at residues#203 and  
151 #204. The multifunctional N protein is composed of three domains<sup>31</sup>, two of which are structurally  
152 independent: the N-terminal domain (NTD) and the C-terminal domain (CTD). Both amino acid  
153 replacements were mapped in an intermediary domain referred to as the linker region (LKR), a  
154 positively charged serine-arginine-rich region. As an intrinsically disordered region (IDR) it allows  
155 the independent folding of the NTD and CTD<sup>32</sup> and is also functionally implicated in RNA binding

156 activity<sup>31</sup>. Key determinants of the interaction between the N and NSP3 proteins were also mapped at  
157 the LKR<sup>33</sup>. The SARS-CoV N protein is also responsible for an antigenic response in humans  
158 predominantly involving the immunoglobulin G<sup>34</sup>. Although the host biological factors involved in  
159 the response to SARS-CoV-2 infection are still poorly known, the existence of distinct virus subtypes,  
160 all of them exhibiting amino acid replacements, could alter important aspects of COVID-19.

161 We hypothesized that in the early stages of the SARS-CoV-2 epidemics, due to the rapid virus  
162 population expansion, a number of genetic variants might have arisen followed by a spread of non-  
163 representative sampling of variants to other countries and continents, *i.g.*, a founder effect. We argue  
164 that the virus subtypes and their associated WSPs detected in this study would be records of  
165 diversification in these early stages of the epidemics after transmission from animal to humans. After  
166 the virus introduction in a given geographic region, a number of unique or narrowly shared mutations  
167 is accumulated, however, most of them reduce the fitness and are removed by purifying selection in a  
168 medium to long term evolutionary scale, tending to a decreasing genetic variability<sup>8</sup>.

169 We propose a classification into at least sixteen distinct subtypes of SARS-CoV-2, six of them  
170 accounting for more than 97% of the sampled isolates from around the world. Such classification  
171 might guide the validation of candidate vaccines or drugs for the widest range of virus subtypes. In  
172 this context, our clustering solution provides a robust approach to effectively reduce the complexity  
173 of the mutant spectrum composed of closely related SARS-CoV-2 genomes focusing on WSPs.  
174 Additionally, through the exhaustive sequencing, it would be possible to identify novel virus subtypes  
175 and follow the evolutionary dynamics of the SARS-CoV-2 population during the adaptive process  
176 imposed by the human host.

177 **Table 1** | Characterization of the WSPs detected in genomes of SARS-CoV-2  
178

Segment ID	Segment position* (begin - end)	WSPs†	nt mutation (# isolates)	Position in the codon	#codon	Amino acid
<b>S2</b>	2,872 - 3,152	<i>nsp3</i> -[318]	<b>U</b> (184) / <b>C</b> (409)	Third	106	Phenylalanine/Phenylalanine
<b>S4</b>	8,612 - 8,892	<i>nsp4</i> -[228]	<b>U</b> (183) / <b>C</b> (410)	Third	76	Serine/Serine
<b>S6</b>	10,932 - 11,192	<i>nsp6</i> -[111]	<b>C</b> (1) / <b>U</b> (99) / <b>G</b> (493)	Third	37	Phenylalanine/Phenylalanine/Leucine
<b>S8</b>	14,232 - 14,512	<i>nsp12</i> -[967]	<b>U</b> (184) / <b>C</b> (409)	Second	323	Leucine/Proline
		<i>nsp13</i> -[1,511]	<b>U</b> (101) / <b>C</b> (492)	Second	504	Leucine/Proline
<b>S10</b>	17,573 - 18,173	<i>nsp13</i> -[1,622]	<b>G</b> (101) / <b>A</b> (492)	Second	541	Cysteine/Tyrosine
		<i>nsp14</i> -[21]	<b>U</b> (105) / <b>C</b> (488)	Third	7	Leucine/Leucine
<b>S12</b>	23,243 - 23,523	<i>S</i> -[1,841]	<b>G</b> (185) / <b>A</b> (408)	Second	614	Glycine/Aspartate
<b>S14</b>	27,977 - 28,258	<i>ORF8</i> -[251]	<b>C</b> (184) / <b>U</b> (409)	Second	84	Serine/Leucine
<b>S16</b>	28,718 - 28,998	<i>N</i> -[608]	<b>A</b> (60) / <b>G</b> (533)	Second	203	Lysine/Arginine
		<i>N</i> -[609]	<b>A</b> (60) / <b>G</b> (533)	Third		
		<i>N</i> -[610]	<b>C</b> (60) / <b>G</b> (533)	First	204	Glycine/Arginine

179 \*Relative to the multiple sequence alignment constructed for full-length genomes

180 †Widely Shared Polymorphisms (WSPs) are conventionally indicated by *cistron/gene*-[nt position within the cistron or gene]

181 **Table 2** | Unique genotypes of SARS-CoV-2 based on 12 WSPs and their associated amino acid replacements

182

SUBTYPES	N*	S2 <sup>‡</sup>	S4	S6	S8	S10			S12	S14	S16		
		<i>nsp3</i>	<i>nsp4</i>	<i>nsp6</i>	<i>nsp12</i>	<i>nsp13</i>		<i>nsp14</i>	<i>S</i>	<i>ORF8</i>	<i>N</i>		
		#318	#228	#111	#967	#1511	#1622	#21	#1,841	#251	#608	#609	#610
I	132	C [Phe] <sup>†</sup>	C [Ser]	G [Leu]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	A [Asp]	U [Leu]	G [Arg]	G [Arg]	G [Arg]
II	122	U [Phe]	C [Ser]	G [Leu]	U [Leu]	C [Pro]	A [Tyr]	C [Leu]	G [Gly]	U [Leu]	G [Arg]	G [Arg]	G [Arg]
III	101	C [Phe]	U [Ser]	G [Leu]	C [Pro]	U [Leu]	G [Cys]	U [Leu]	A [Asp]	C [Ser]	G [Arg]	G [Arg]	G [Arg]
IV	91	C [Phe]	C [Ser]	U [Phe]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	A [Asp]	U [Leu]	G [Arg]	G [Arg]	G [Arg]
V	74	C [Phe]	U [Ser]	G [Leu]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	A [Asp]	C [Ser]	G [Arg]	G [Arg]	G [Arg]
VI	58	U [Phe]	C [Ser]	G [Leu]	U [Leu]	C [Pro]	A [Tyr]	C [Leu]	G [Gly]	U [Leu]	A [Lys]	A [Lys]	C [Gly]
VII	3	C [Phe]	U [Ser]	U [Phe]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	A [Asp]	C [Ser]	G [Arg]	G [Arg]	G [Arg]
VIII	3	C [Phe]	U [Ser]	G [Leu]	C [Pro]	C [Pro]	A [Tyr]	U [Leu]	A [Asp]	C [Ser]	G [Arg]	G [Arg]	G [Arg]
IX	2	U [Phe]	C [Ser]	U [Phe]	U [Leu]	C [Pro]	A [Tyr]	C [Leu]	G [Gly]	U [Leu]	A [Lys]	A [Lys]	C [Gly]
X	1	U [Phe]	C [Ser]	U [Phe]	U [Leu]	C [Pro]	A [Tyr]	C [Leu]	G [Gly]	U [Leu]	G [Arg]	G [Arg]	G [Arg]
XI	1	U [Phe]	C [Ser]	G [Leu]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	G [Gly]	U [Leu]	G [Arg]	G [Arg]	G [Arg]
XII	1	C [Phe]	U [Ser]	C [Phe]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	A [Asp]	C [Ser]	G [Arg]	G [Arg]	G [Arg]
XIII	1	C [Phe]	C [Ser]	G [Leu]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	A [Asp]	C [Ser]	G [Arg]	G [Arg]	G [Arg]
XIV	1	C [Phe]	U [Ser]	G [Leu]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	G [Gly]	C [Ser]	G [Arg]	G [Arg]	G [Arg]
XV	1	C [Phe]	C [Ser]	U [Phe]	U [Leu]	C [Pro]	A [Tyr]	C [Leu]	A [Asp]	U [Leu]	G [Arg]	G [Arg]	G [Arg]
XVI	1	C [Phe]	C [Ser]	U [Phe]	C [Pro]	C [Pro]	A [Tyr]	U [Leu]	A [Asp]	U [Leu]	G [Arg]	G [Arg]	G [Arg]

183 \*Sample size

184 <sup>‡</sup>Segment containing the WSP

185 #Nucleotide position

186 <sup>†</sup>Nucleotide base and the encoded amino acid residue

187



188 **Figure legends**

189 **Fig. 1** | Mean pairwise number of nucleotide differences per site (nucleotide diversity,  $\pi$ ) calculated  
190 using a sliding window of 300 nucleotides across the multiple sequence alignment for full-length  
191 genomes of SARS-CoV-2. The red dashed line at  $\pi = 0.001$  represents an arbitrary threshold to  
192 subdivide the segments (S) with increased (S2, 4, 6, 8, 10, 12, 14 and 16) and lower (S1, 3, 5, 7,  
193 9, 11, 13,15 and 17) content of genetic variation. The SARS-CoV-2 genome organization is  
194 represented on top of the plot.

195 **Fig. 2** | Multidimensional Scaling (MDS) visualization of tree distances based on Kendall Colijn  
196 metrics ( $\lambda = 0$ ). The seventeen ML-trees (each one with 593 tips) are represented as dots. Seven  
197 groups of trees are represented with different colors. The WSP-containing segments-based trees  
198 (S2, S8 and S12) formed six groups (1 - 6 are indicated in blue, red, green, orange, purple and  
199 brown, respectively). All nWSPs-containing segments-based ML-trees formed a single group  
200 indicated in pink.

201 **Fig. 3** | Maximum likelihood phylogenetic tree based on 12 WSPs detected across the SARS-CoV-  
202 2 genomes.

## 203 **Methods**

204 A total of 1,137 full-length genomes of SARS-CoV-2 sampled from December 2019 to  
205 March 25, 2020 (at 2:30 pm) were obtained from Genbank<sup>35</sup> and GISAID<sup>36</sup> (Table S1). Only  
206 genomes with high sequencing coverage, intact ORFs (no frameshifts, except that of *nsp12* cistron)  
207 and without any indeterminate nucleotide base (indicated by ‘N’s or ambiguous codes) totaling  
208 767 high quality full-length sequences were effectively analyzed in this study. We wish to  
209 acknowledge all researchers that deposited the SARS-CoV-2 genomes in GISAID and/or GenBank  
210 databases.

211 The genomic data set was aligned using MAFFT-FFT-NS-2<sup>37</sup>. The calculation of the  
212 average number of nucleotide differences per site (nucleotide diversity,  $\pi$ ) was conducted in  
213 DnaSP v.6<sup>38</sup> using a sliding window and step size of 300 and 20 nucleotides, respectively. Sites  
214 with gaps alignment were not considered in analysis.

215 Maximum likelihood (ML) phylogenetic trees were constructed using RAxML<sup>39</sup> under the  
216 nucleotide substitution model General Time-Reversible with gamma distribution (GTRGAMMA).  
217 The branch support for ML-trees based on 300-nucleotides and larger segments was assessed with  
218 1,000 and 5,000 bootstrap replications, respectively. ML-trees were used in this study essentially  
219 as a clustering method due to the weak phylogenetic signal in the data set. All phylogenetic trees  
220 were edited using iTOL<sup>40</sup>. In order to assess the similarity among ML-tree topologies, we  
221 computed all possible pairwise distances using the Kendall–Colijn metric<sup>41</sup> followed by Principal  
222 Coordinates Analysis (PCoA) using the package TREE SPACE<sup>42</sup> in R<sup>43</sup>.

223 The detection of polymorphic sites was conducted using PAUP\* v. 4.0<sup>44</sup> and MEGA X<sup>45</sup>.  
224 Those sites responsible for the segregation of the isolates into two clusters in the ML-trees were  
225 referred to as “widely shared polymorphisms” (WSPs), while the remaining nt positions in the  
226 virus genomes were designed as “non widely shared polymorphisms” (nWSPs). The WSPs were  
227 conventionally indicated by *cistron/gene* name-[nt position within the cistron or gene]. We opted  
228 by indicating the nt position within the cistron or gene due to their highly conserved lengths (no  
229 gap was introduced during the construction of sequence alignments), in contrast to the full-  
230 genomes whose 5’- and 3’-untranslated regions (UTRs) were highly variable in terms of length.

231  
232 **Data Availability:** multiple sequence alignments and ML-phylogenetic trees generated in this  
233 study are available from the authors upon request.

234 **References**

- 235
- 236 1. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature*
- 237 **579**, 265–269 (2020).
- 238 2. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The
- 239 species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and
- 240 naming it SARS-CoV-2. *Nat. Microbiol.* **5**, (2020).
- 241 3. WHO. WHO Director-General’s remarks at the media briefing on 2019-nCoV on 11
- 242 February 2020. *WHO website* <https://www.who.int/dg/speeches/detail/who-directo> (2020).
- 243 4. Sawicki, S. G. & Sawicki, D. L. Coronavirus Transcription: A Perspective. in *Current topics*
- 244 *in microbiology and immunology* vol. 287 31–55 (2005).
- 245 5. de Wilde, A. H., Snijder, E. J., Kikkert, M. & van Hemert, M. J. Host Factors in Coronavirus
- 246 Replication. in *Assessment & Evaluation in Higher Education* vol. 37 1–42 (Springer
- 247 International Publishing, 2017).
- 248 6. Kim, D. *et al.* *The architecture of SARS-CoV-2 transcriptome.* (2020) doi:10.1088/1751-
- 249 8113/44/8/085201.
- 250 7. Peck, K. M. & Luring, A. S. Complexities of Viral Mutation Rates. *J. Virol.* **92**, (2018).
- 251 8. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war — host adaptation and its
- 252 constraints on virus evolution. *Nat. Rev. Microbiol.* **17**, 321–328 (2019).
- 253 9. Domingo, E., Sheldon, J. & Perales, C. Viral Quasispecies Evolution. *Microbiol. Mol. Biol.*
- 254 *Rev.* **76**, 159–216 (2012).
- 255 10. Domingo, E. & Perales, C. Viral quasispecies. *PLoS Genet.* **15**, 1–20 (2019).
- 256 11. Zhao, Z. *et al.* Moderate mutation rate in the SARS coronavirus genome and its
- 257 implications. *BMC Evol. Biol.* **4**, 1–9 (2004).
- 258 12. Gorbalenya, A. E., Enjuanes, L., Ziebuhr, J. & Snijder, E. J. Nidovirales: Evolving the
- 259 largest RNA virus genome. *Virus Res.* **117**, 17–37 (2006).
- 260 13. Ma, Y. *et al.* Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10
- 261 complex. *Proc. Natl. Acad. Sci.* **112**, 9436–9441 (2015).
- 262 14. Knipe, D. M. & Howley, P. M. *Fields virology. Viruses and the Lung: Infections and Non-*
- 263 *Infectious Viral-Linked Lung Disorders* (2013).
- 264 15. Forni, D., Cagliani, R., Clerici, M. & Sironi, M. Molecular Evolution of Human
- 265 Coronavirus Genomes. *Trends Microbiol.* **25**, 35–48 (2017).
- 266 16. Angelini, M. M., Akhlaghpour, M., Neuman, B. W. & Buchmeier, M. J. Severe Acute
- 267 Respiratory Syndrome Coronavirus Nonstructural Proteins 3, 4, and 6 Induce Double-
- 268 Membrane Vesicles. *MBio* **4**, 1–10 (2013).
- 269 17. Cottam, E. M., Whelband, M. C. & Wileman, T. Coronavirus NSP6 restricts
- 270 autophagosome expansion. *Autophagy* **10**, 1426–1441 (2014).
- 271 18. Gao, Y. *et al.* Structure of RNA-dependent RNA polymerase from 2019-nCoV, a major
- 272 antiviral drug target. *bioRxiv* 2020.03.16.993386 (2020) doi:10.1101/2020.03.16.993386.
- 273 19. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is
- 274 Blocked by a Clinically Proven Protease Inhibitor. *Cell* 1–10 (2020)
- 275 doi:10.1016/j.cell.2020.02.052.
- 276 20. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS
- 277 coronavirus. *Nature* vol. 426 450–454 (2003).
- 278
- 279

- 280 21. Matsuyama, S. *et al.* Efficient Activation of the Severe Acute Respiratory Syndrome  
281 Coronavirus Spike Protein by the Transmembrane Protease TMPRSS2. *J. Virol.* **84**, 12658–  
282 12664 (2010).
- 283 22. Shulla, A. *et al.* A Transmembrane Serine Protease Is Linked to the Severe Acute  
284 Respiratory Syndrome Coronavirus Receptor and Activates Virus Entry. *J. Virol.* **85**, 873–  
285 882 (2011).
- 286 23. Rossen, J. W. A. *et al.* The Viral Spike Protein Is Not Involved in the Polarized Sorting of  
287 Coronaviruses in Epithelial Cells †. *J. Virol.* **72**, 497–503 (1998).
- 288 24. Yang, Y. *et al.* Two Mutations Were Critical for Bat-to-Human Transmission of Middle  
289 East Respiratory Syndrome Coronavirus. *J. Virol.* **89**, 9119–9123 (2015).
- 290 25. Pallesen, J. *et al.* Immunogenicity and structures of a rationally designed prefusion MERS-  
291 CoV spike antigen. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7348–E7357 (2017).
- 292 26. Walls, A. C. *et al.* Unexpected Receptor Functional Mimicry Elucidates Activation of  
293 Coronavirus Fusion. *Cell* **176**, 1026–1039 (2019).
- 294 27. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike  
295 Glycoprotein. *Cell* **180**, 1–12 (2020).
- 296 28. Sung, S.-C., Chao, C.-Y., Jeng, K.-S., Yang, J.-Y. & Lai, M. M. C. The 8ab protein of  
297 SARS-CoV is a luminal ER membrane-associated protein and induces the activation of  
298 ATF6. *Virology* **387**, 402–413 (2009).
- 299 29. Chen, C. *et al.* Open Reading Frame 8a of the Human Severe Acute Respiratory Syndrome  
300 Coronavirus Not Only Promotes Viral Replication but Also Induces Apoptosis. *J. Infect.*  
301 *Dis.* **196**, 405–415 (2007).
- 302 30. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev.*  
303 *Microbiol.* **17**, 181–192 (2019).
- 304 31. Parker, M. M. & Masters, P. S. Sequence comparison of the N genes of five strains of the  
305 coronavirus mouse hepatitis virus suggests a three domain structure for the nucleocapsid  
306 protein. *Virology* **179**, 463–468 (1990).
- 307 32. Huang, Q. *et al.* Structure of the N-Terminal RNA-Binding Domain of the SARS CoV  
308 Nucleocapsid Protein. *Biochemistry* **43**, 6059–6063 (2004).
- 309 33. Verheije, M. H. *et al.* The Coronavirus Nucleocapsid Protein Is Dynamically Associated  
310 with the Replication-Transcription Complexes. *J. Virol.* **84**, 11575–11579 (2010).
- 311 34. Leung, D. T. M. *et al.* Antibody Response of Patients with Severe Acute Respiratory  
312 Syndrome (SARS) Targets the Viral Nucleocapsid. *J. Infect. Dis.* **190**, 379–386 (2004).
- 313 35. Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res.* **47**, D94–D99 (2019).
- 314 36. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from  
315 vision to reality. *Eurosurveillance* **22**, 2–4 (2017).
- 316 37. Katoh, K., Misawa, K., Kei-ichi, K. & Miyata, T. MAFFT: a novel method for rapid  
317 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–  
318 3066 (2002).
- 319 38. Rozas, J. *et al.* DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol.*  
320 *Biol. Evol.* **34**, 3299–3302 (2017).
- 321 39. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large  
322 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- 323 40. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new  
324 developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
- 325

- 326 41. Kendall, M. & Colijn, C. Mapping Phylogenetic Trees to Reveal Distinct Patterns of  
327 Evolution. *Mol. Biol. Evol.* **33**, 2735–2743 (2016).
- 328 42. Jombart, T., Kendall, M., Almagro-Garcia, J. & Colijn, C. Treespace: Statistical exploration  
329 of landscapes of phylogenetic trees. *Mol. Ecol. Resour.* **17**, 1385–1392 (2017).
- 330 43. R Core Team. R: A language and environment for statistical computing. (2018).
- 331 44. Swofford, D. L. PAUP\*: Phylogenetic Analysis Using Parsimony (\*and Other Methods).  
332 (2002) doi:10.1111/j.0014-3820.2002.tb00191.x.
- 333 45. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary  
334 Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
- 335

336 **Acknowledgments**

337 This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível  
338 Superior - Brasil (CAPES) - Finance Code 001. IJM and RCP were recipients of CNPq and CAPES  
339 doctoral fellowships, respectively. DIB was the recipient of a FAPEMIG master fellowship. We  
340 wish to acknowledge all researchers that deposited the SARS-CoV-2 genomes in GISAID and/or  
341 GenBank databases.

342

343 **Author Contributions**

344 ATML designed the bioinformatics analyses. IJM, ATML, RCP, GMS, DIB and FTS conducted  
345 the analyses. ATML, IJM, RCP and FTS analyzed data and results. IJM, RCP, FTS and ATML  
346 wrote the manuscript. All authors contributed to the content and writing of the Supplementary  
347 Information.

348

349 **Competing interest declaration**

350 The authors declare that they have no competing interests.

351

352 **Materials & Correspondence**

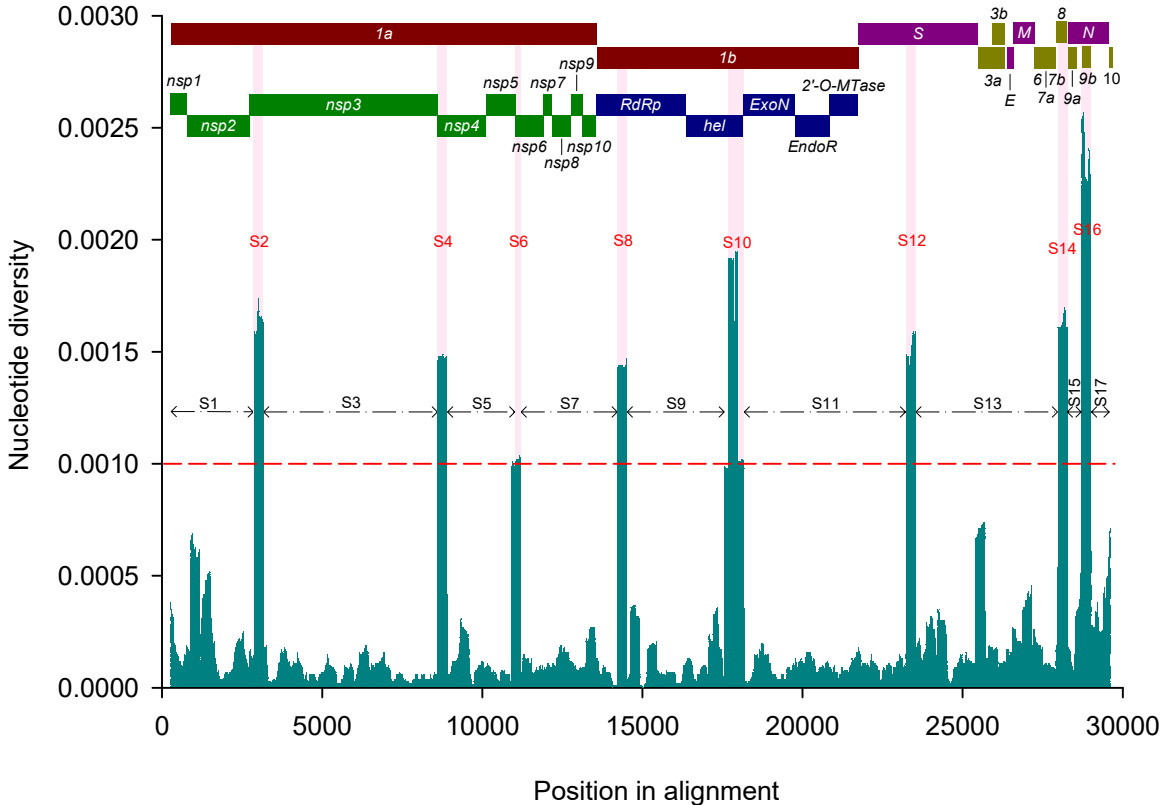
353 Alison Talis Martins Lima

354

355 **Additional information**

356 Supplementary information is available for this manuscript.

# SARS-CoV-2



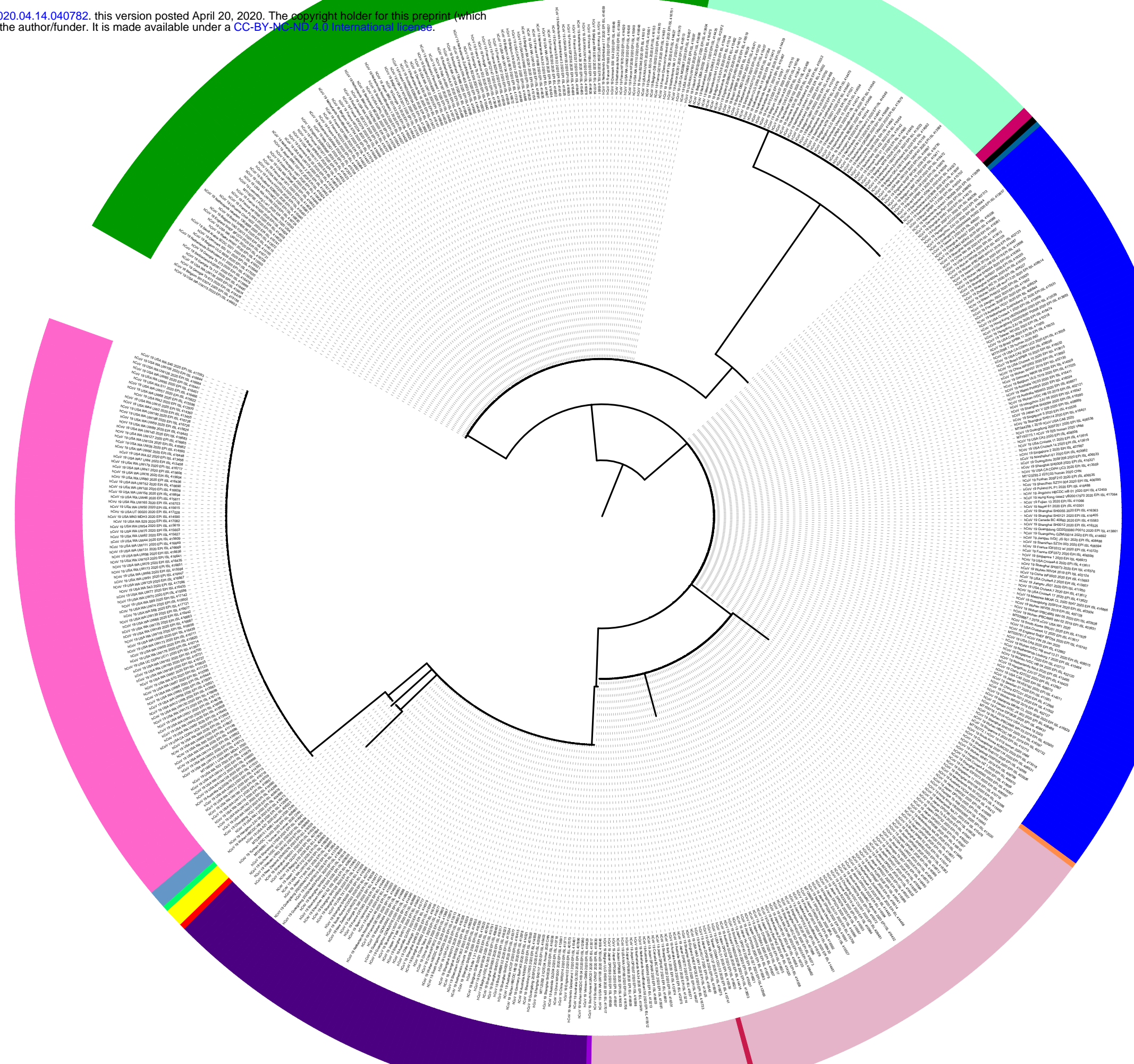




# Subtypes of SARS-CoV-2

- I [N=132]
- II [N=122]
- III [N=101]
- IV [N=91]
- V [N=74]
- VI [N=58]
- VII [N=3]
- VIII [N=3]
- IX [N=2]
- X [N=1]
- XI [N=1]
- XII [N=1]
- XIII [N=1]
- XIV [N=1]
- XV [N=1]
- XVI [N=1]

bioRxiv preprint doi: <https://doi.org/10.1101/2020.04.14.040782>; this version posted April 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).



Tree scale: 1