

1 **MOLECULAR EPIDEMIOLOGY TO UNDERSTAND THE SARS-CoV-2**
2 **EMERGENCE IN THE BRAZILIAN AMAZON REGION**

3 Mirleide Cordeiro dos Santos^{1*}, Edivaldo Costa Sousa Junior^{1*}, Jessylene de Almeida
4 Ferreira^{1*}, Sandro Patroca da Silva², Michel Platini Caldas de Souza², Jedson Ferreira
5 Cardoso², Amanda Mendes Silva¹, Luana Soares Barbagelata¹, Wanderley Dias das
6 Chagas Junior¹, James Lima Ferreira¹, Edna Maria Acunã de Souza¹, Patrícia Louise
7 Araújo Vilaça¹, Jainara Cristina dos Santos Alves², Michelle Carvalho de Abreu²,
8 Patrícia dos Santos Lobo², Fabíolla da Silva dos Santos², Alessandra Alves Polaro
9 Lima², Camila de Marco Bragagnolo², Luana da Silva Soares², Patrícia Sousa Moraes
10 de Almeida², Darleise de Souza Oliveira², Carolina Koury Nassar Amorim², Iran Barros
11 Costa², Dielle Monteiro Teixeira², Edvaldo Tavares da Penha Júnior², Delana Andreza
12 Melo Bezerra², Jones Anderson Monteiro Siqueira², Fernando Neto Tavares², Felipe
13 Bonfim Freitas², Janete Taynã Nascimento Rodrigues³, Janaína Mazaro³, Andreia
14 Santos Costa⁴, Márcia Socorro Pereira Cavalcante⁴, Marineide Souza da Silva⁵,
15 Guilherme Alfredo Novelino Araújo⁵, Ilvanete Almeida da Silva⁶, Gleissy Adriane
16 Lima Borges⁶, Lídio Gonçalves de Lima⁷, Hivylla Lorrana dos Santos Ferreira⁷,
17 Miriam Teresinha Furlam Prando Livorati⁹, André Luiz de Abreu⁹, Arnaldo Correia de
18 Medeiros⁹, Hugo Reis Resque², Rita Catarina Medeiros Sousa⁸, Giselle Maria Rachid
19 Viana².

20 *Correspondent authors

21 Mirleide Cordeiro dos Santos – mirleidesantos@iec.gov.br

22 Edivaldo Costa Sousa Junior - costajr.013@gmail.com

23 Jessylene de Almeida Ferreira – jessylene_almeida@hotmail.com

24 **Affiliations**

25 ¹Laboratory of Respiratory Viruses, Evandro Chagas Institute (IEC), National Influenza Center (NIC) for
26 the World Health Organization (WHO), Health Surveillance Office and Brazilian Health Ministry,
27 Ananindeua, Pará, Brazil.

28 ²Virology section, Evandro Chagas Institute (IEC), Health Surveillance Office and Brazilian Health
29 Ministry, Ananindeua, Pará, Brazil.

30 ³Laboratório Central de Saúde Pública do Acre – (Central Laboratory of Public Health of Acre).
31 (LACEN-AC) Rio Branco, AC, Brazil.

32 ⁴Laboratório Central de Saúde Pública do Amapá – (Central Laboratory of Public Health of Amapá).
33 (LACEN-AP). Macapá, AP, Brazil.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

34 ⁵Laboratório Central de Saúde Pública do Amazonas – (Central Laboratory of Public Health of
35 Amazonas). (LACEN-AM). Manaus, AM, Brazil.

36 ⁶Laboratório Central de Saúde Pública do Pará – (Central Laboratory of Public Health of Pará). (LACEN-
37 PA). Belém, PA, Brazil.

38 ⁷Laboratório Central de Saúde Pública do Maranhão – (Central Laboratory of Public Health of
39 Maranhão). (LACEN-MA). São Luís, MA, Brazil.

40 ⁸ Virology Postgraduate Program, Evandro Chagas Institute , Para, Brazil

41 ⁹Secretaria de Vigilância em Saúde, Ministério da Saúde – (Health Surveillance Office, Health Ministry).
42 Brasília, DF, Brazil.

43

44 **ABSTRACT**

45 The COVID-19 pandemic in Brazil has demonstrated an important public health impact,
46 as has been observed in the world. In Brazil, the Amazon Region contributed with a
47 large number of cases of COVID-19, especially in the beginning of the circulation of
48 SARS-CoV-2 in the country. Thus, we describe the epidemiological profile of COVID-
49 19 and the genetic diversity of SARS-CoV-2 strains circulating in the Amazon Region.
50 We observe an extensive spread of virus in this Brazilian site. The data on sex, age and
51 symptoms presented by the investigated individuals were similar to what has been
52 observed worldwide. The genomic analysis of the viruses revealed important amino
53 acid changes, including the D614G and the I33T in Spike and ORF6 proteins,
54 respectively. The latter found in strains originating in Brazil. The phylogenetic analyzes
55 demonstrated the circulation of the lineages B.1 and B.1.1, whose circulation in Brazil
56 has already been previous reported. Our data reveals molecular epidemiology of SARS-
57 CoV-2 in the Amazon Region. These findings also reinforce the importance of
58 continuous genomic surveillance this virus with the aim of providing accurate and
59 updated data to understand and map the transmission network of this agent in order to
60 subsidize operational decisions in public health.

61

62 Keywords: SARS-CoV-2; COVID-19; Amazon Region; Brazil.

63

64 **INTRODUCTION**

65 Coronavirus disease 2019 (COVID-19) is an infectious disease caused by a
66 newly discovered *Betacoronavirus*, now recognized as severe acute respiratory
67 syndrome coronavirus 2 (SARS-CoV-2)^{1,2}, and is responsible for one of the most
68 significant pandemics in this century, causing millions of cases and high rates of
69 hospitalizations and deaths^{3,4}.

70 In South America, Brazil holds the first place of infected individuals, with
71 3,761,391 diagnosed cases and, out of these, approximately 118,649 Brazilians have
72 lost their lives due to COVID-19³. To this end, the Amazon region has effectively
73 contributed to the number of 760,394 infected patients and 19.358 deaths (Update
74 August 28th2020)⁵ with some of the states in this region presenting the worst scenario of

75 cases at the beginning of the pandemic in Brazil with high rates of occupancy in
76 intensive care units (ICU) and deaths. Most persons with COVID-19 experience mild to
77 moderate respiratory symptoms and recover⁶. On the other hand, individuals with
78 underlying medical conditions, such as cardiovascular disease, diabetes, chronic
79 respiratory diseases and cancer are more likely to be severely and possibly in need of
80 intensive care ^{6,7}.

81 In addition to epidemiological information, the SARS-CoV-2 genomic data, as
82 well as evolution datasets to quantify the impact of non-pharmaceutical interventions
83 (NPIs) in virus spatiotemporal spread, are under much investigation, and until then it has
84 been shown that this virus has diversified into several phylogenetic strains⁸, marked by
85 different punctual mutations that reflect ongoing transmission currents⁹.

86 In view of the above, investigations aimed at evaluating the circulation
87 dynamics, genetics and evolutionary characteristics of SARS-CoV-2 are of substantial
88 importance for the global surveillance of this virus and, consequently, will provide a
89 better understanding of the virus, the disease it causes and its circulation, providing
90 relevant information for the development of new therapeutic and control strategies and
91 prevention of infections by the new coronavirus. To this end, we combine genetics and
92 epidemiological data to investigate the genetic diversity, evolution and epidemiology of
93 SARS-CoV-2 in the Amazon region.

94

95 **RESULTS**

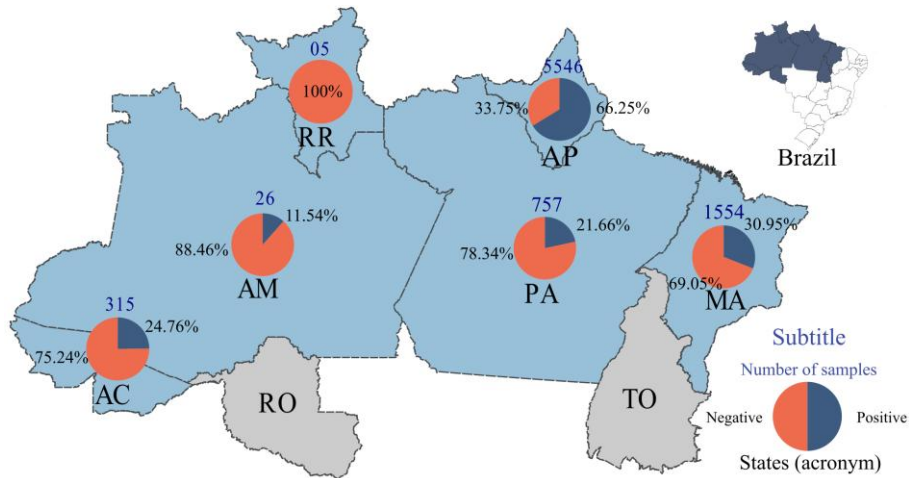
96

97 **Epidemiological data**

98 In the Brazilian Amazon region, 8,203 samples were analyzed and 4,400 of
99 which (53.64%) were positive for SARS-CoV-2. The frequency of detection within the
100 region is shown in figure 1. As for circulation, the highest rate was in epidemiological
101 weeks (EW) 17, 18, 19 and 21, with the highest peak in EW 18 (figure 2).

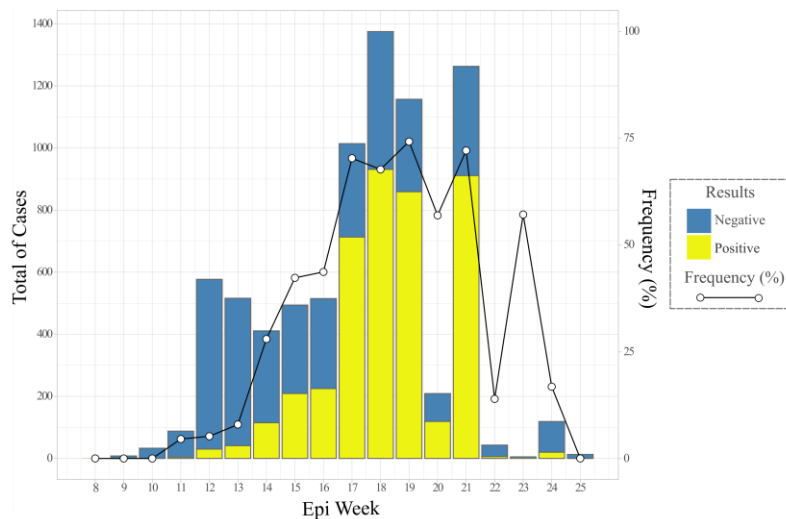
102

103 **Figure 1-** Distribution of SARS-CoV-2 cases in the Amazon region. Total number of samples studied =
 104 8,203.



105
 106
 107

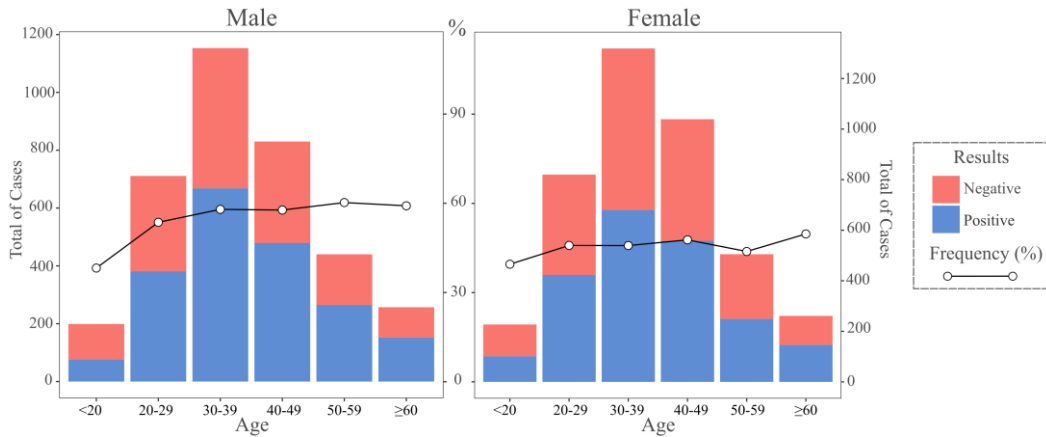
Figure 2- Distribution of SARS-CoV-2 cases according to the epidemiological week.



108
 109
 110
 111
 112
 113
 114
 115
 116

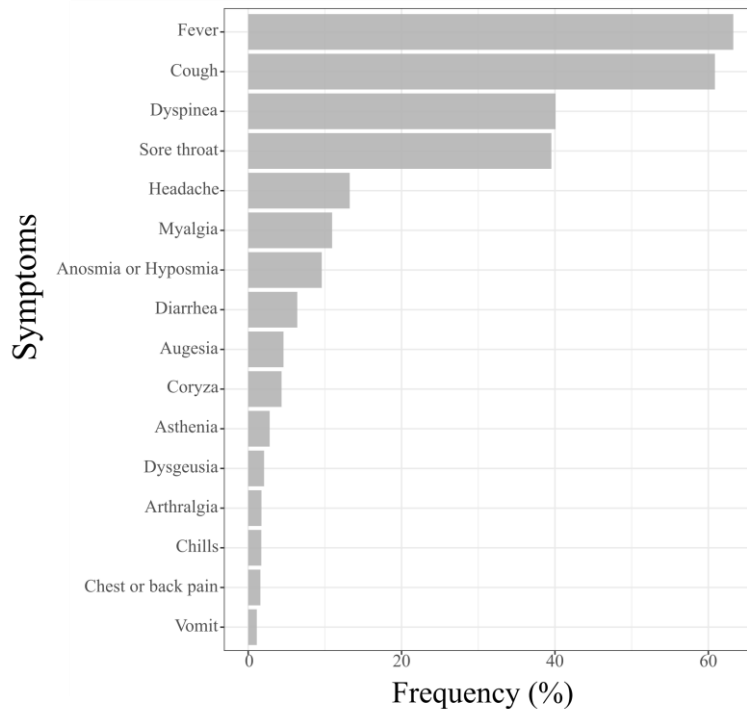
Amongst the 4,400 SARS-CoV-2 positive cases, 214 did not contain age information. In this regard, the distribution of positive samples by age group has demonstrated that the highest frequency of positivity has occurred in the adult population amongst the over 20 age groups; the average age was 47 (figure 3). Regarding sex, 2,273 (51.57%) are female and 2,116 are male (56.04%) and 11 (0.25%) did not inform it. Fever (63.11%) was the most common symptom amongst patients, followed by cough (60.70%), dyspnoea (39.52%) and sore throat (39.45%) (figure 4).

117 **Figure 3** - Absolute and relative frequency of positive and negative cases for SARS-CoV-2, by sex and
 118 age group.



119

120 **Figure 4**- Description of symptoms and signs among positive cases for COVID-19.

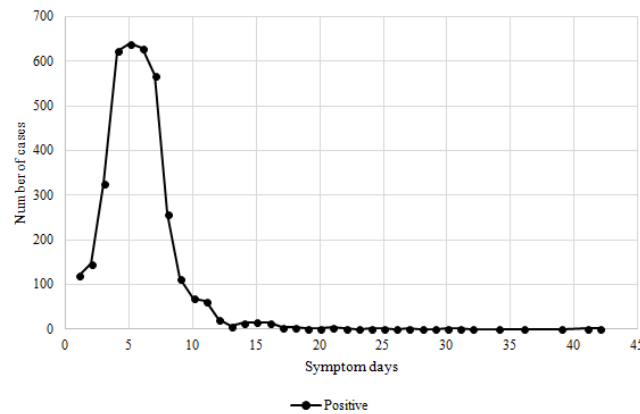


121

122 Amidst the confirmed cases of COVID-19 by molecular assay, it was observed
 123 that the detection range varied between the 1st and the 42nd day after the onset of
 124 symptoms, being the fifth day (62.22%) the best collection day for detection by RT-
 125 qPCR after the onset of symptoms (figure 5).

126

127 **Figure 5** - Positivity for SARS-CoV-2 regarding disease duration.



128

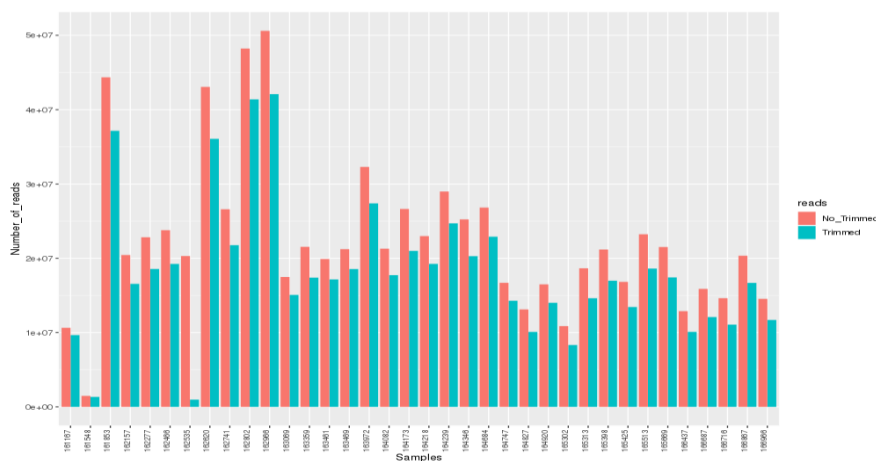
129 NGS sequencing

130 Thirty-three (33) samples were successfully sequenced from the states of Acre
131 (1), Amapá (11), Maranhão (8), Pará (11), Paraíba (1) and Rio Grande do Norte (1).
132 These samples were analyzed and showed an average of 22,594,487 reads per
133 sequenced sample, ranging from 1,476,498 to 50,601,712 reads (supplementary material
134 1).

135 Pre-processing data

136 After the trimming process of regions with low quality, removal of the adapters
137 and reads smaller than 40 bp, these samples had presented an average of 18,207,316
138 reads per sample, extending from 971,498 to 42,083,456 reads and then were used for
139 assembly by De Novo and Reference Mapping (figure 6 and supplementary material 1).

140 **Figure 6** - Number of reads before and after pre-processing. The reads with length less than 40 bp and
141 quality less than Phread 20 were removed.



142

143 **Genome Assembly (De Novo and Reference Mapping)**

144 For the genomic assembly process, the assembly was first done via the De
145 Novo method, which has generated an average of 364,488 contigs per sample,
146 extending from 3,964 to 1,816,455 contigs. As for the minimum size of the generated
147 contigs is 200 bp, all the sequences under this length were discarded, for there is no
148 variation in this item. The average of the maximum length of the contigs was 44,130 bp,
149 extending from 4,574 bp to 312,642 bp. The generated N50 lengths were on average
150 409 bp, ranging from 344 bp to 618 bp (supplementary material 1).

151 For reference mapping assemblies, the average of reads mapped was 479,937
152 reads, extending from 11,935 to 3,087,036 reads per sample, leading to an average
153 coverage of 3,007x, extending from 115x to 36,460x (figure 7 and supplementary
154 material 1). All genomes had 38% GC content. There is no variation in this regard. All
155 SARS-CoV-2 genomes were assembled almost entirely during assembly by De Novo,
156 with only the ends needing editing, assembled via reference mapping and later sequence
157 edition.

158

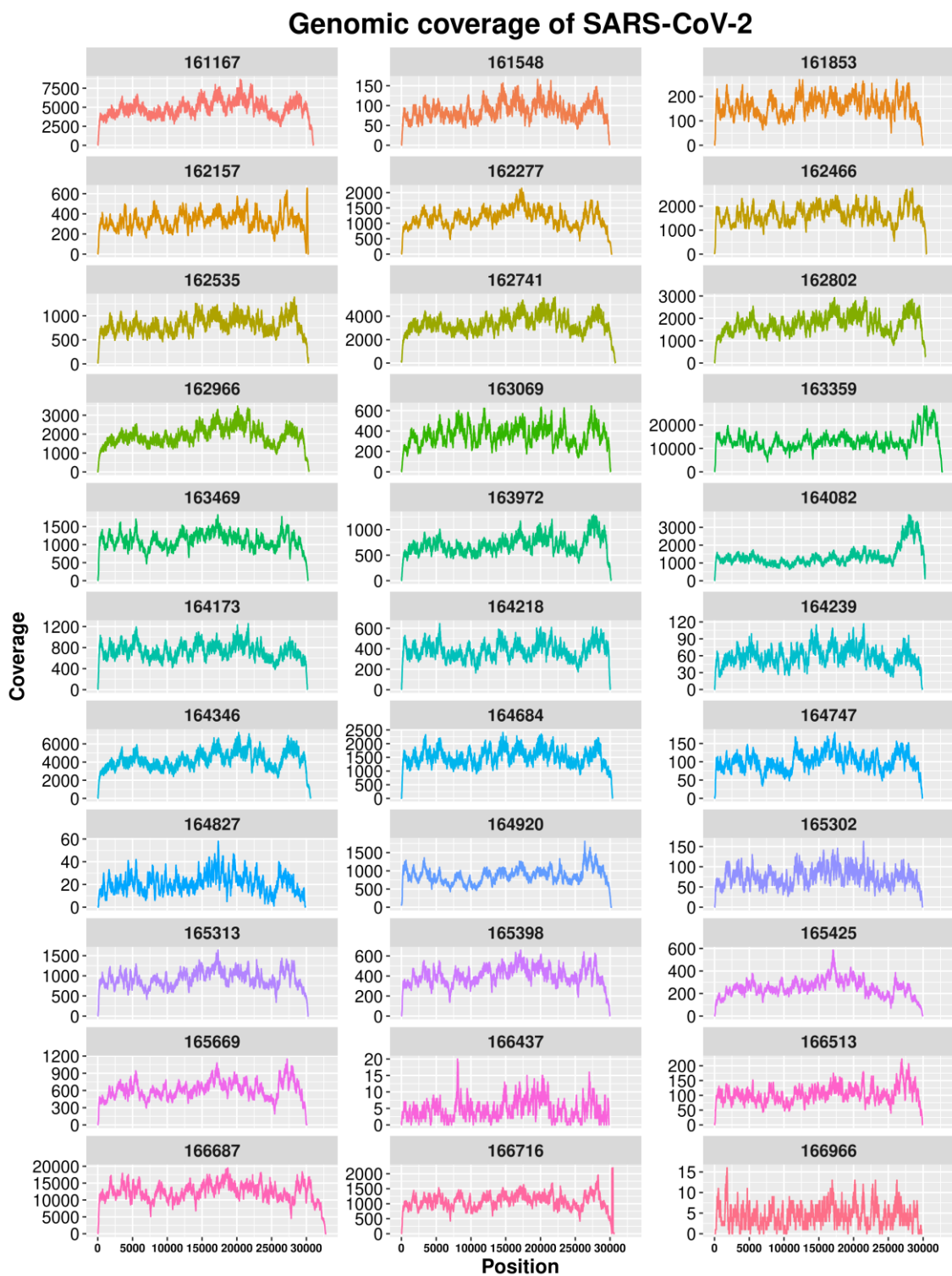
159 **Phylogenetic Analysis and Mutation analysis**

160 The genomes obtained were aligned with the reference strains deposited in
161 GISAID, showing an identity of 99.98% (supplementary material 2). The analysis of the
162 SARS-CoV-2 genome found revealed 62 nucleotide changes in 12 genes leading to 32
163 amino acid changes in 7 proteins (supplementary material 3).

164 The phylogenetic analysis reveals that isolates from present study clustering in
165 three major clades in B.1 (one clade) and B.1.1 (two clades), with moderate statistical
166 values of 70-89%. These clades are characterized by the presence of mutations
167 S:D614G (B.1 and B.1.1) and substitutions in protein N (N:R203K and N:G204R) that
168 classifies the lineage B.1.1 (Figure 08).

169

170 **Figure 7** - Coverage plot by sample sequenced in the present study.



171

172

173 **Figure 8** - SARS-CoV-2 complete genome phylogenetic tree (ML) with 1000 bootstraps, using GTR
 174 evolution model for nucleotide substitutions.



175

176

177 **DISCUSSION**

178 In Brazil, the first case of COVID-19 was detected in the state of São Paulo on
179 February 26, 2020. In the Brazilian Amazon Region, the first detection has occurred in
180 the state of Manaus on March 13, 2020 and, in that month, the Brazilian Health
181 Ministry had already reported community transmission in the country, as well as a
182 pandemic status by the World Health Organization (WHO). After the first detection of
183 SARS-CoV-2 in this Brazilian site, an extensive spread of this virus was observed in the
184 region, demonstrating, in the present study, a detection frequency of 53.64% showing a
185 critical adaptation and circulation of SARS-CoV-2 in this tropical region. The most
186 active circulation of SARS-CoV-2 has occurred mainly in April and May, in EW 17, 18,
187 19 and 21, which coincides with the Amazonian winter season.

188 The detection of SARS-CoV-2 has occurred in all age groups. However, the
189 highest frequency has occurred in adults and the elderly, as described in the world
190 population¹⁰⁻¹³. The low frequency of SARS-CoV-2 in children and teenagers under 20,
191 verified in the present study, has been associated with reduced susceptibility and less
192 likelihood to infection or a combination of both, compared to adults¹⁴⁻¹⁶.

193 As for sex, similar to what has been reported by literature, the frequency
194 amongst men was slightly higher than in women^{17,18}. Estrogen, the main female sex
195 hormone, plays a possible protective role in COVID-19, activating the immune
196 response, and directly suppressing SARS-CoV-2 replication^{19,20}. Indeed, estrogen
197 inhibits the activity or expression of different components of the renin-angiotensin
198 system. Particularly, estrogen can upregulate ACE2 expression^{21,22}. Regarding the
199 clinical analysis presented by the investigated patients, the most described symptoms
200 were fever, cough, sore throat and dyspnoea, respectively, associated or not, commonly
201 reported amongst respiratory infections, as well as in COVID-19^{10,11,23,24}. However,
202 unlike other respiratory infections, in COVID-19 anosmia and dysgeusia have been
203 frequently reported²⁵⁻²⁹, as described amongst patients in the study region. After the
204 onset of symptoms, the period of greatest detection of SARS-CoV-2 by RT-qPCR has
205 occurred on the fifth day, similarly to what has been described in other studies³⁰.

206 The genomic analysis of the viruses found revealed that 61 nucleotide
207 mutations were found in the entire genome when compared to the reference genome,
208 and, out of these, 16 led to amino acid changes, with emphasis on the substitutions in S
209 and N proteins that have a structural role and ORF6, a non-structural protein not yet
210 characterized^{31,32}.

211 Amongst the alterations in the Spike protein that plays a role in binding to the
212 human ACE2 receptor and is also the main antigenic target, it was found the D614G
213 substitution that is described as a factor that antigenically favors the virus, giving it a
214 higher capacity to infection³³ and has been used as a genetic marker for strains of the B-
215 lineage (Pangolin Classification) which has become the largest circulating group
216 worldwide³⁴. Also, it was verified the V1176F mutation described in the literature³⁵ and
217 used as a genetic marker for samples circulating in Brazil (<https://www.gisaid.org>), but
218 no antigenic advantage has yet been attributed.

219 Regarding the N protein plays a role in folding viral genetic material and has
220 been used as a marker for samples from Europe, it was verified R203K, G204R and
221 I292T amino acidic substitutions. However, their molecular roles are still unclear^{36,37}.
222 The change I33T in ORF6, a non-structural protein, has been observed in samples
223 originating in Brazil and that circulate in South America³⁸.

224 The phylogenetic analysis revealed that the samples of this study have formed
225 three distinct groups that cluster with the phylogenetic lineages B.1 and B.1.1 that have
226 samples already sequenced from Brazil³⁹. Within clade B.1, only two samples from Pará
227 were clustering with samples from Europe. In clade B.1.1, it was possible to observe the
228 formation of two distinct groups divided by the I33T ORF6 and V1176F S protein
229 substitutions. These two mutations have been observed to divide the two main strains of
230 SARS-CoV-2 circulating in Brazil³⁴. Since its worldwide circulation on December
231 2019⁴⁰, the SARS-CoV-2 genome has changed wherever it arrives⁴¹, which may mean a
232 likely adaptation to the population⁴².

233 In this study, we did not yet had the chance to analyze how SARSCoV-2
234 became established across the Amazon region and to associate the finding lineages with
235 the population movements, that is, to relate to the proportion of within and between
236 state measured virus movements. Another relevant issue is that the B.1 and B.1.1
237 lineages from the Amazon region were quite similar, making it difficult to trace with
238 precision the origin of these strains in the study site.

239 In conclusion, this study reveals that the highest SARS-CoV-2 circulation has
240 reached its peak in epidemiological week 18. The distribution of positive samples by
241 age group has demonstrated that the median age was 47, with men being the main
242 affected gender and there was a spectrum of symptoms composed of fever, cough,
243 dyspnoea and sore throat. Furthermore, this investigation supports the evidence for the
244 existence of two main lineages (B.1 and B.1.1) associated with genomic epidemiology

245 of SARS-CoV-2 in the Amazon region. Thus, genomic surveillance must be
246 continuously adopted to be able to offer accurate and quality data to understand where
247 this virus emerged from, and map the transmission network to improve operational
248 decisions in public health.

249

250 **METHODS**

251

252 **Samples and ethical aspects**

253 The Laboratory of Respiratory Viruses of the Evandro Chagas Institute (LVR-
254 IEC), located in the Amazon region, works with the World Health Organization (WHO)
255 as a National Influenza Center (NIC) for the surveillance of influenza and other
256 respiratory viruses, amongst them, the SARS-CoV-2. Thus, this laboratory has received
257 8,203 clinical specimens from patients of both sexes and in different age groups (zero to
258 111 years old) between February 27th, 2020 to July 1st, 2020 for the diagnosis of SARS-
259 CoV-2 from the states of Acre, Amapá, Amazonas, Ceará, Maranhão, Pará, Paraíba,
260 Pernambuco, Rio Grande Norte and Roraima. The clinical specimens collected and used
261 for molecular diagnosis and viral genetic analysis were nasopharyngeal swabs plus
262 throat swabs, nasopharyngeal aspirate and sputum. This study was approved by Evandro
263 Chagas Institute Ethical Committee (34931820.0.0000.0019).

264

265 **Extraction and Detection by RT-qPCR of viral nucleic acid**

266 The viral RNA was extracted manually using the QIAamp® Viral RNA
267 Mini Kit (QIAGEN, Hilden, Germany) following the manufacturer's guidelines. The
268 detection of the viral genome by RT-qPCR was performed with the Molecular Kit
269 SARS-CoV-2 (E/RP) Biomanguinhos (Biomanguinhos, Rio de Janeiro, Brazil),
270 according to the protocol described by Corman et al (2020)⁴⁵.

271

272 The amplification reaction was conducted sequentially in the following
273 steps: reverse transcription at 50°C for 15 minutes, followed by transcriptase
274 inactivation and activation of Taq DNA polymerase at 95 ° C for 2 minutes, polymerase
275 chain reaction at 95°C for 15 seconds in 45 cycles, extension and annealing at 55 ° C for
276 30 seconds. At the end of the amplification, all clinical samples should have reaction
277 sigmoid curves for the targets that cross the limit line (*cycle threshold* - Ct) equal to or
before 40 cycles. Positive and negative controls were included in each reaction.

278

279 **Epidemiological analysis**

280 Graphs of epidemiological data (age, sex, state, signs and symptoms) and
281 circulation were performed with support by the LVR-IEC database and the Microsoft
282 Office Excel program. The data were inspected, visualized and plotted using the R
283 programming language script⁴³ together with the libraries ggplot2⁴⁴, geobr⁴⁵, pipeR⁴⁶,
284 readr⁴⁷, lubridate, fmsb⁴⁸, plyr⁴⁹, scales⁵⁰, viridis⁵¹ and hrbrthemes⁵². By international
285 convention, epidemiological weeks were counted from Sunday to Saturday, considering
286 the sample collection date.

287

288 **Sample selection for sequencing**

289 The selection of strains for sequencing the viral genome was conducted so that
290 there was geographical and temporal representativeness. In this aspect, the date of
291 collection and the respective epidemiological week of the sample of each state of origin
292 were considered to reach the minimum representation of each federated unit per
293 epidemiological week. In addition, in order to obtain the highest amount of viral RNA
294 and, thus, a greater chance of success in sequencing, samples that showed $Ct \leq 20$ in the
295 RT-qPCR for SARS-CoV-2 were selected.

296

297 **Library construction and sequencing**

298 The total RNA was converted to complementary DNA (cDNA) using the
299 cDNA Synthesis System Kit and 400 μ M of random primers, following the
300 manufacturer's procedure. The reaction solution was purified with the Agencourt
301 AMPure XP Reagent. The cDNA library was prepared and sequenced using the
302 methodology described in the Nextera XT DNA Library Preparation Kit on a NextSeq
303 (Illumina, Inc) platform by paired-end methodology with 300 cycles (2x150 reads), in
304 the Evandro Chagas Institute, Brazil Ministry of Health.

305

306 **Data pre-processing**

307 The data were evaluated for their quality regions. The adapters sequences reads
308 with a quality lower than Phred 20 and reads with less than 40 bp size, were removed
309 using Trimmomatic⁵³. The processed reads were visualized with FastQC⁵⁴. For
310 Trimmomatic, we have used the following parameters: LEADING:3
311 TRAILING:3 MINLEN:40

312

313 **Genome Assembly (*De Novo* and *Reference Mapping*)**

314 For this step, the reads validated based on quality trimming were used to
315 assembly the SARS-CoV-2 genomes. The De Novo assembly was performed using the
316 Megahit v.1.1.4-2⁵⁵ and for Reference Mapping we have used the software Bowtie2⁵⁶
317 and Geneious Prime, where the respective coverage, gaps and final size of the genome
318 were analyzed. For genome assembly, all programs were performed with default
319 parameters.

320

321 **Taxonomic annotation and submission to GISAID**

322 The generated de novo contigs were compared using the Blastx tool⁵⁷
323 implemented in Diamond v.0.9.33⁵⁸, against the RefSeq database (NCBI's Protein
324 Reference Sequences Database), which is a database of cured protein sequences and
325 which provides a high level of annotation, such as the description of the function of a
326 protein, its domain structure, post-translational modifications, where a statistical value
327 (e-value) of 0.0001 was considered.

328 The viral genome annotation was performed automatically using the Geneious
329 Prime software (*Biomatters, Ltd., New Zealand, 2019*) and cured manually by
330 comparing the starts and stop codons, as well as the sizes of the genes. These genome
331 sequences were subsequently submitted to the GISAID database
332 (<https://www.gisaid.org/>) under accession numbers EPI_ISL_450873-450874,
333 EPI_ISL_458138-EPI_ISL_458149 and EPI_ISL_524783-EPI_ISL_524801.

334 **Phylogenetic Analysis and Mutation analysis**

335 The genomes sequences were aligned with other genomes from all the world
336 using the Mafft v.7.471⁵⁹. For phylogenetic analysis the software RaXML⁶⁰ with 1000
337 bootstraps was used as statistical support, using GTR as a nucleotide substitution model.
338 The genomes obtained were compared to the reference strain (NC_045512) by *in house*
339 python script that compares each base of the entire genome and gives us a mutation list.

340

341 **REFERENCES**

- 342 1. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related
343 coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat.*
344 *Microbiol.* **5**, 536–544 (2020).
- 345 2. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel
346 coronavirus: implications for virus origins and receptor binding. *Lancet* **395**,
347 565–574 (2020).
- 348 3. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard.
349 *WHO Coronavirus Disease (COVID-19) Dashboard* (2020).
- 350 4. World Health Organization. Statement on the second meeting of the International
351 Health Regulations (2005) Emergency Committee regarding the outbreak of
352 novel coronavirus (2019-nCoV). [https://www.who.int/news-room/detail/30-01-](https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov))
353 *2020-statement-on-the-second-meeting-of-the-international-health-regulations-*
354 *(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-*
355 *(2019-ncov)* (2020).
- 356 5. Ministry of health. COVID-19 in Brazil.
357 <http://susanalitico.saude.gov.br/#/dashboard/> (2020).
- 358 6. Huang, C. *et al.* Clinical features of patients infected with 2019 novel
359 coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
- 360 7. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients
361 with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**,
362 1054–1062 (2020).
- 363 8. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist
364 genomic epidemiology. *bioRxiv* 2020.04.17.046086 (2020).
365 doi:10.1101/2020.04.17.046086
- 366 9. Stefanelli, P. *et al.* Whole genome and phylogenetic analysis of two SARSCoV-2
367 strains isolated in Italy in January and February 2020: Additional clues on
368 multiple introductions and further circulation in Europe. *Eurosurveillance* **25**, 1–
369 5 (2020).
- 370 10. Guan, W. *et al.* Clinical characteristics of coronavirus disease 2019 in China. *N.*
371 *Engl. J. Med.* **382**, 1708–1720 (2020).
- 372 11. Jiang, F. *et al.* Review of the Clinical Characteristics of Coronavirus Disease
373 2019 (COVID-19). *J. Gen. Intern. Med.* **35**, 1545–1549 (2020).
- 374 12. Shahid, Z. *et al.* COVID-19 and Older Adults: What We Know. *J. Am. Geriatr.*
375 *Soc.* **68**, 926–929 (2020).
- 376 13. Nikolich-Zugich, J. *et al.* SARS-CoV-2 and COVID-19 in older adults: what we
377 may expect regarding pathogenesis, immune responses, and outcomes.
378 *GeroScience* **42**, 505–514 (2020).
- 379 14. Davies, N. G. *et al.* Age-dependent effects in the transmission and control of
380 COVID-19 epidemics. *Nat. Med.* (2020). doi:10.1038/s41591-020-0962-9

- 381 15. Mantovani, A. *et al.* Coronavirus disease 2019 (COVID-19) in children and/or
382 adolescents: a meta-analysis. *Pediatr. Res.* **2019**, (2020).
- 383 16. Ludvigsson, J. F. Systematic review of COVID-19 in children shows milder
384 cases and a better prognosis than adults. *Acta Paediatr. Int. J. Paediatr.* **109**,
385 1088–1095 (2020).
- 386 17. Li, L. quan *et al.* COVID-19 patients' clinical characteristics, discharge rate, and
387 fatality rate of meta-analysis. *J. Med. Virol.* **92**, 577–583 (2020).
- 388 18. Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R. & Klein, S. L.
389 Impact of sex and gender on COVID-19 outcomes in Europe. *Biol. Sex Differ.*
390 **11**, 1–13 (2020).
- 391 19. Channappanavar, R. *et al.* Sex-Based Differences in Susceptibility to Severe
392 Acute Respiratory Syndrome Coronavirus Infection. *J. Immunol.* **198**, 4046–4053
393 (2017).
- 394 20. Scully, E. P., Haverfield, J., Ursin, R. L., Tannenbaum, C. & Klein, S. L.
395 Considering how biological sex impacts immune responses and COVID-19
396 outcomes. *Nat. Rev. Immunol.* (2020). doi:10.1038/s41577-020-0348-8
- 397 21. Bukowska, A. *et al.* Protective regulation of the ACE2/ACE gene expression by
398 estrogen in human atrial tissue from elderly men. *Exp. Biol. Med.* **242**, 1412–
399 1423 (2017).
- 400 22. Scully, E. P., Haverfield, J., Ursin, R. L., Tannenbaum, C. & Klein, S. L.
401 Considering how biological sex impacts immune responses and COVID-19
402 outcomes. *Nat. Rev. Immunol.* (2020). doi:10.1038/s41577-020-0348-8
- 403 23. Rodríguez-Cola, M. *et al.* Clinical features of coronavirus disease 2019 (COVID-
404 19) in a cohort of patients with disability due to spinal cord injury. *Spinal Cord*
405 *Ser. Cases* **6**, (2020).
- 406 24. Ge, H. *et al.* The epidemiology and clinical information about COVID-19. *Eur. J.*
407 *Clin. Microbiol. Infect. Dis.* **39**, 1011–1019 (2020).
- 408 25. Vaira, L. A., Salzano, G., Deiana, G. & De Riu, G. Anosmia and Ageusia:
409 Common Findings in COVID-19 Patients. *Laryngoscope* **130**, 1787 (2020).
- 410 26. Vaira, L. A., Salzano, G., Fois, A. G., Piombino, P. & De Riu, G. Potential
411 pathogenesis of ageusia and anosmia in COVID-19 patients. *Int. Forum Allergy*
412 *Rhinol.* **00**, 1–2 (2020).
- 413 27. Russell, B. *et al.* Anosmia and ageusia are emerging as symptoms in patients
414 with COVID-19: What does the current evidence say? *Ecancermedalscience*
415 **14**, 9–10 (2020).
- 416 28. Klopfenstein, T. *et al.* Features of anosmia in COVID-19. *Med. Mal. Infect.* 4–7
417 (2020). doi:10.1016/j.medmal.2020.04.006
- 418 29. Whittaker, A., Anson, M. & Harky, A. Neurological Manifestations of COVID-
419 19: A systematic review and current update. *Acta Neurol. Scand.* **142**, 14–22
420 (2020).

- 421 30. Rodríguez-Cola, M. *et al.* Clinical features of coronavirus disease 2019 (COVID-
422 19) in a cohort of patients with disability due to spinal cord injury. *Spinal Cord*
423 *Ser. Cases* **6**, (2020).
- 424 31. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for
425 drug repurposing. *Nature* **583**, (2020).
- 426 32. da Silva, S. J. R., da Silva, C. T. A., Mendes, R. P. G. & Pena, L. Role of
427 Nonstructural Proteins in the Pathogenesis of SARS-CoV-2. *J. Med. Virol.* 3–5
428 (2020). doi:10.1002/jmv.25858
- 429 33. Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: evidence that D614G
430 increases infectivity of the COVID-19 virus. *Cell* (2020).
431 doi:10.1016/j.cell.2020.06.043
- 432 34. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data –
433 from vision to reality. *Eurosurveillance* **22**, 2–4 (2017).
- 434 35. Gonçalves, R. L. *et al.* SARS-CoV-2 mutations and where to find them: An in
435 silico perspective of structural changes and antigenicity of the Spike protein.
436 *bioRxiv* **3**, 2020.05.21.108563 (2020).
- 437 36. Yin, C. Genotyping coronavirus SARS-CoV-2: methods and implications.
438 *Genomics* **19**, 1–12 (2020).
- 439 37. Castillo, A. E. *et al.* Phylogenetic analysis of the first four SARS-CoV-2 cases in
440 Chile. *J. Med. Virol.* 1–5 (2020). doi:10.1002/jmv.25797
- 441 38. Resende, P. C. *et al.* Genomic surveillance of SARS-CoV-2 reveals community
442 transmission of a major lineage during the early pandemic phase in Brazil.
443 *bioRxiv* 2020.06.17.158006 (2020). doi:10.1101/2020.06.17.158006
- 444 39. Candido, D. S. *et al.* Evolution and epidemic spread of SARS-Cov-2 in Brazil.
445 *Science (80-.)*. **21**, 1–9 (2020).
- 446 40. Riou, J. & Althaus, C. L. Pattern of early human-to-human transmission of
447 Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020.
448 *Eurosurveillance* **25**, 1–5 (2020).
- 449 41. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel
450 RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **18**, 1–9 (2020).
- 451 42. Cao, Y. *et al.* Comparative genetic analysis of the novel coronavirus (2019-
452 nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.* **6**, 4–7
453 (2020).
- 454 43. R Foundation for Statistical Computing. R Core Team (2018). R: A language and
455 environment for statistical computing. (2018).
- 456 44. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* (2009).
- 457 45. Pereira, R.H.M.; Gonçalves, C. N. *geobr: Loads Shapefiles of Official Spatial*
458 *Data Sets of Brazil.* GitHub repository. (2019).
- 459 46. Ren, K. *pipeR: Multi-Paradigm Pipeline Implementation.* R package version

- 460 0.6.1.3. (2016).
- 461 47. Hadley Wickham, J. H. and R. F. readr: Read Rectangular Text Data. R package
462 version 1.3.1. (2018).
- 463 48. Nakazawa, M. fmsb: Functions for Medical Statistics Book with some
464 Demographic Data. R package version 0.7.0. [https://CRAN.R-](https://CRAN.R-project.org/package=fmsb)
465 [project.org/package=fmsb](https://CRAN.R-project.org/package=fmsb) (2019).
- 466 49. Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. *Journal of*
467 *Statistical Software*, 40(1), 1-29. (2011).
- 468 50. Seidel, H. W. and D. scales: Scale Functions for Visualization. R package version
469 1.1.1. (2020).
- 470 51. Simon Garnier. viridis: Default Color Maps from ‘matplotlib’. R package version
471 0.5.1. (2018).
- 472 52. Bob Rudis. hrbrthemes: Additional Themes, Theme Components and Utilities for
473 ‘ggplot2’. R package version 0.8.0. 2020
- 474 53. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for
475 Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- 476 54. Andrews, S. FastQC: a quality control tool for high throughput sequence data.
477 <http://www.bioinformatics.babraham.ac.uk/project> (2010).
- 478 55. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-
479 fast single-node solution for large and complex metagenomics assembly via
480 succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
- 481 56. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-
482 efficient alignment of short DNA sequences to the human genome. *Genome Biol.*
483 **10**, (2009).
- 484 57. Madden, T. & Coulouris, G. BLAST+ User Manual. *Ncbi* 1–64 (2008).
- 485 58. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
486 DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
- 487 59. Katoh, K. & Toh, H. Parallelization of the MAFFT multiple sequence alignment
488 program. *Bioinformatics* **26**, 1899–900 (2010).
- 489 60. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-
490 analysis of large phylogenies. *Bioinformatics* **30**, 1312–3 (2014).

491

492

493 **Acknowledgements**

494 The authors would like to thank all the professionals who worked bravely to deal with
495 this pandemic, especially in the Amazon. We thank the Evandro Chagas Institute, where
496 the development of the research was carried out with great contribution from the
497 virology team. We would also like to thank the General Coordination of Laboratories
498 (CGLab) of the Ministry of Health (MS), States of the Brazilian Central Laboratory
499 (LACENs), and local surveillance teams for the partnership in viral surveillance in
500 Brazil.

501

502 **Author contributions**

503 MCS, HRR, RCMS and GMRV coordinated the study; ECSJ, JAF, AMS, SPS, MPCS
504 and JFC performed the sequencing and genomic analysis of SARS-CoV-2 strains; LSB,
505 WDCJ, AMS and JAF performed the detection of SARS-CoV-2 by RT-qPCR; AMS,
506 JLF, EMAS, CKNA and DSO received and checked the samples; PLAV, JCSA, MCA,
507 PSL, FSS, AAPL, CMB, LSS and PSMA performed the registration of epidemiological
508 data in a database and assisted in the release of results; IBC, DMT, ETPJ, DAMB,
509 JAMS, FNT and FBF performed the extraction of the viral genome; MSS, GAN, IAS,
510 GALB, LGL, HLSF collected the samples; MTFPL, ALA and ACM assisted with
511 technical support through the viral surveillance network in Brazil.

512

513 **Additional Information**

514 The authors declare that they have no conflicting interests.

515

516 **Figure 1** - Distribution of SARS-CoV-2 cases in the Amazon region. Total number of
517 samples studied = 8,203.

518 **Figure 2** - Distribution of SARS-CoV-2 cases according to the epidemiological week.

519 **Figure 3** - Absolute and relative frequency of positive and negative cases for SARS-
520 CoV-2, by sex and age group.

521 **Figure 4** - Description of symptoms and signs among positive cases for COVID-19.

522 **Figure 5** - Positivity for SARS-CoV-2 regarding disease duration.

523 **Figure 6** - Number of reads before and after pre-processing. The reads with length less
524 than 40 bp and quality less than Phread 20 were removed.

525 **Figure 7** - Coverage plot by sample sequenced in the present study.

526 **Figure 8** - SARS-CoV-2 complete genome phylogenetic tree (ML) with 1000
527 bootstraps, using GTR evolution model for nucleotide substitutions.