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

Mirleide Cordeiro dos Santos[1]*, Edivaldo Costa Sousa Junior[1]*, Jessylene de Almeida Ferreira[1]*, Sandro Patroca da Silva[2], Michel Platini Caldas de Souza[2], Jedson Ferreira Cardoso[2], Amanda Mendes Silva[1], Luana Soares Barbagelata[1], Wanderley Dias das Chagas Junior[1], James Lima Ferreira[1], Edna Maria Acunã de Souza[1], Patrícia Louise Araújo Vilaça[1], Jainara Cristina dos Santos Alves[2], Michelle Carvalho de Abreu[2], Patrícia dos Santos Lobo[2], Fabíolla da Silva dos Santos[2], Alessandra Alves Polaro Lima[2], Camila de Marco Bragagnolo[2], Luana da Silva Soares[2], Patrícia Sousa Moraes de Almeida[2], Darleise de Souza Oliveira[2], Carolina Koury Nassar Amorim[2], Iran Barros Costa[2], Dielle Monteiro Teixeira[2], Edvaldo Tavares da Penha Júnior[2], Delana Andreza Melo Bezerra[2], Jones Anderson Monteiro Siqueira[2], Fernando Neto Tavares[2], Felipe Bonfim Freitas[2], Janete Taynã Nascimento Rodrigues[3], Janaína Mazaro[3], Andreia Santos Costa[4], Márcia Socorro Pereira Cavalcante[4], Marineide Souza da Silva[5], Guilherme Alfredo Novelino Araújo[5], Ilvanete Almeida da Silva[6], Gleissy Adriane Lima Borges[6], Lídio Gonçalves de Lima[7], Hivylla Lorrana dos Santos Ferreira[7], Miriam Teresinha Furlam Prando Livorati[9], André Luiz de Abreu[9], Arnaldo Correia de Medeiros[9], Hugo Reis Resque[2], Rita Catarina Medeiros Sousa[8], Giselle Maria Rachid Viana[2].

*Correspondent authors

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

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

Jessylene de Almeida Ferreira – jessylene_almeida@hotmail.com

Affiliations

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

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

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

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

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

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

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

[8] Virology Postgraduate Program, Evandro Chagas Institute , Para, Brazil

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

ABSTRACT

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

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

INTRODUCTION

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

In South America, Brazil holds the first place of infected individuals, with 3,761,391 diagnosed cases and, out of these, approximately 118,649 Brazilians have lost their lives due to COVID-19[3]. To this end, the Amazon region has effectively contributed to the number of 760,394 infected patients and 19.358 deaths (Update August 28th2020)[5] with some of the states in this region presenting the worst scenario of cases at the beginning of the pandemic in Brazil with high rates of occupancy in intensive care units (ICU) and deaths. Most persons with COVID-19 experience mild to moderate respiratory symptoms and recover[6]. On the other hand, individuals with underlying medical conditions, such as cardiovascular disease, diabetes, chronic respiratory diseases and cancer are more likely to be severely and possibly in need of intensive care [6,7].

In addition to epidemiological information, the SARS-CoV-2 genomic data, as well as evolution datasets to quantify the impact of non-pharmaceutical interventions (NPIs) in virus spatiotemporal spread, are under much investigation, and until then it has been shown that this virus has diversified into several phylogenetic strains[8], marked by different punctual mutations that reflect ongoing transmission currents[9].

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

RESULTS

Epidemiological data

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

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

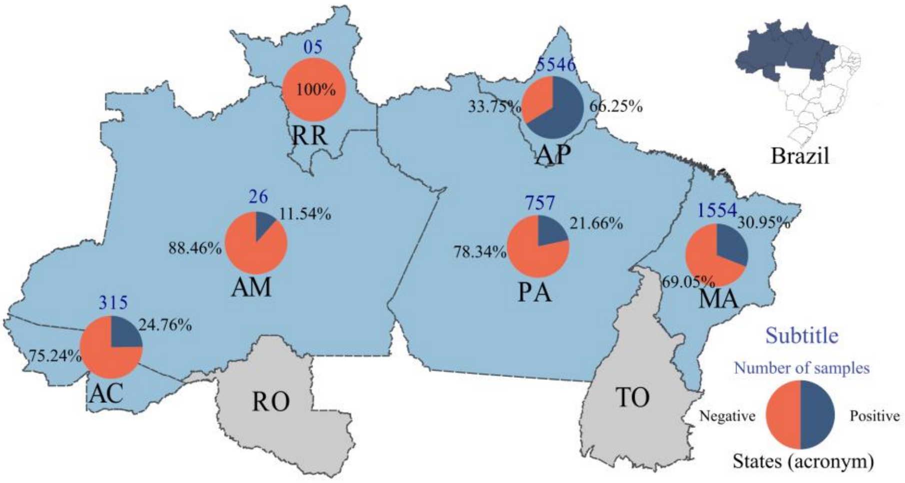
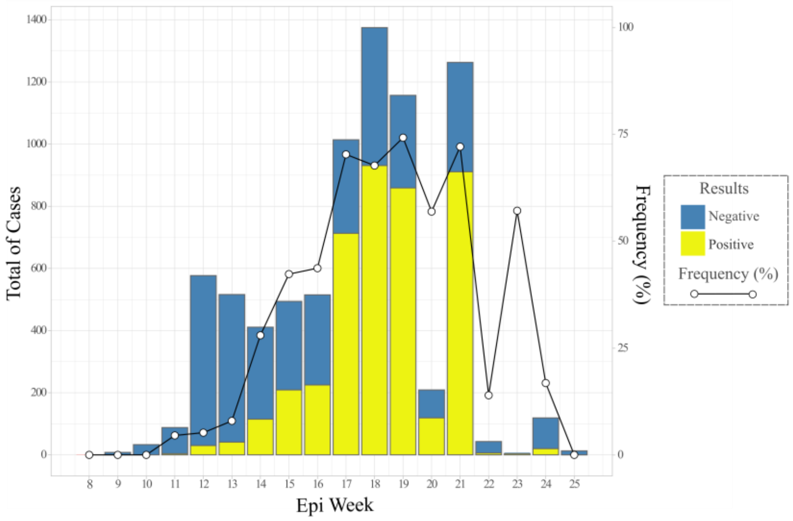


Figure 2- Distribution of SARS-CoV-2 cases according to the epidemiological week.



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).

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

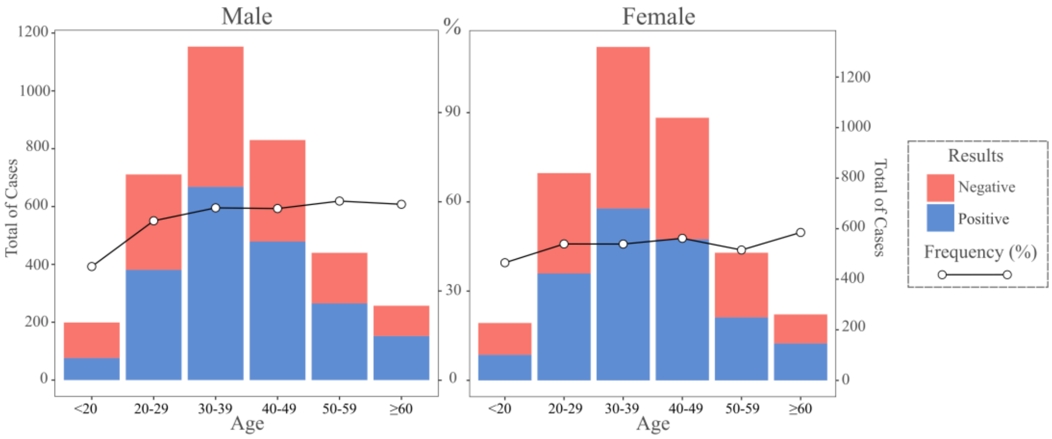
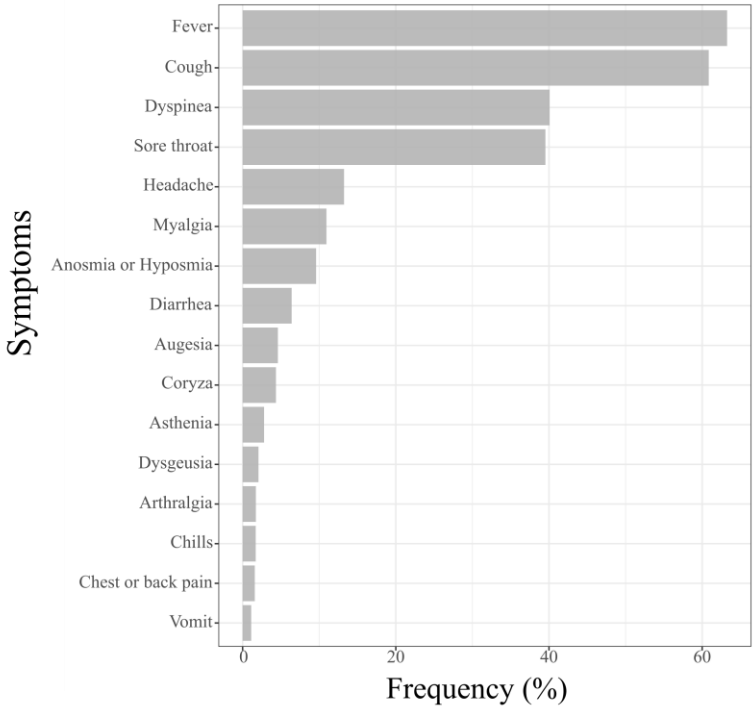
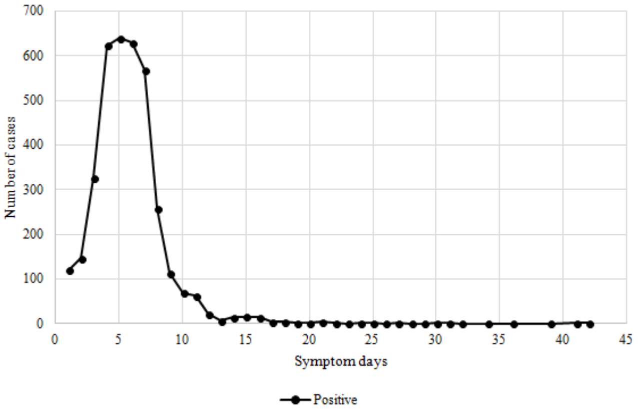


Figure 4- Description of symptoms and signs among positive cases for COVID-19.



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

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



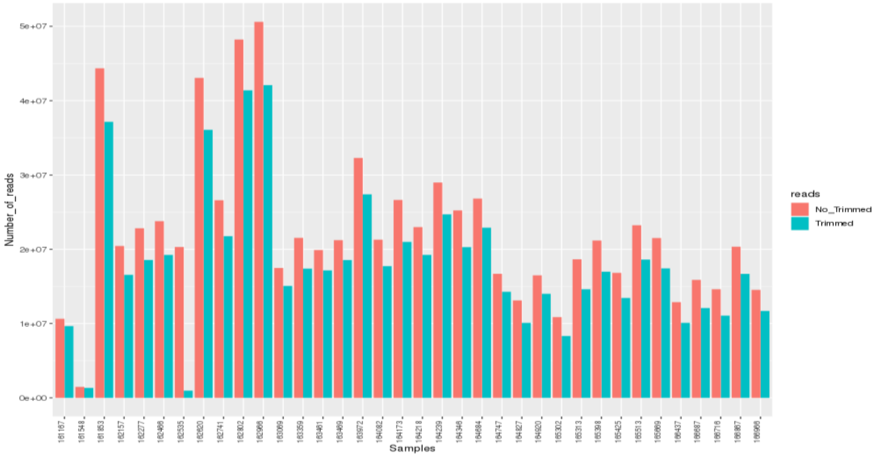
NGS sequencing

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

Pre-processing data

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

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



Genome Assembly (De Novo and Reference Mapping)

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

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

Phylogenetic Analysis and Mutation analysis

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

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

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

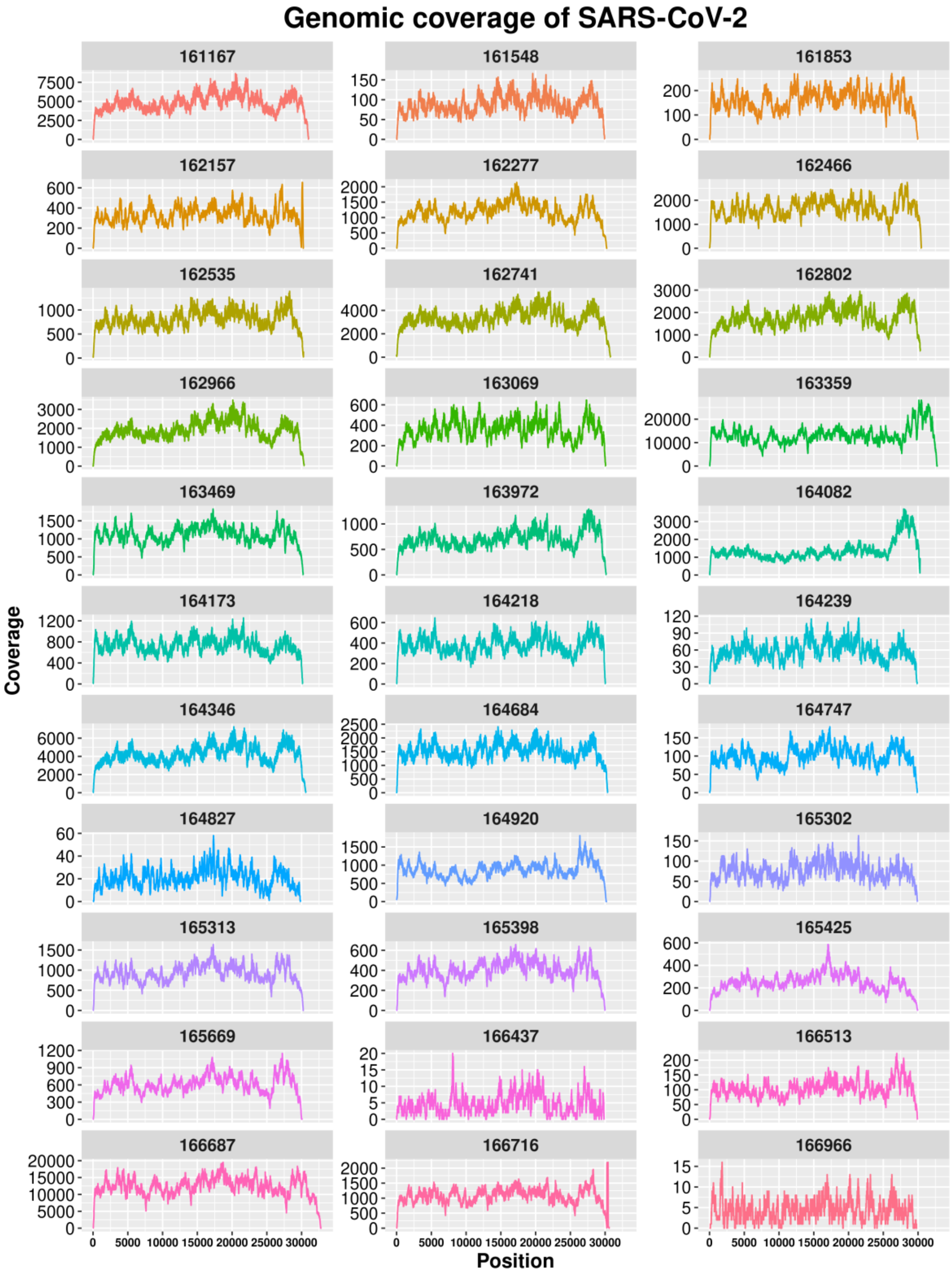
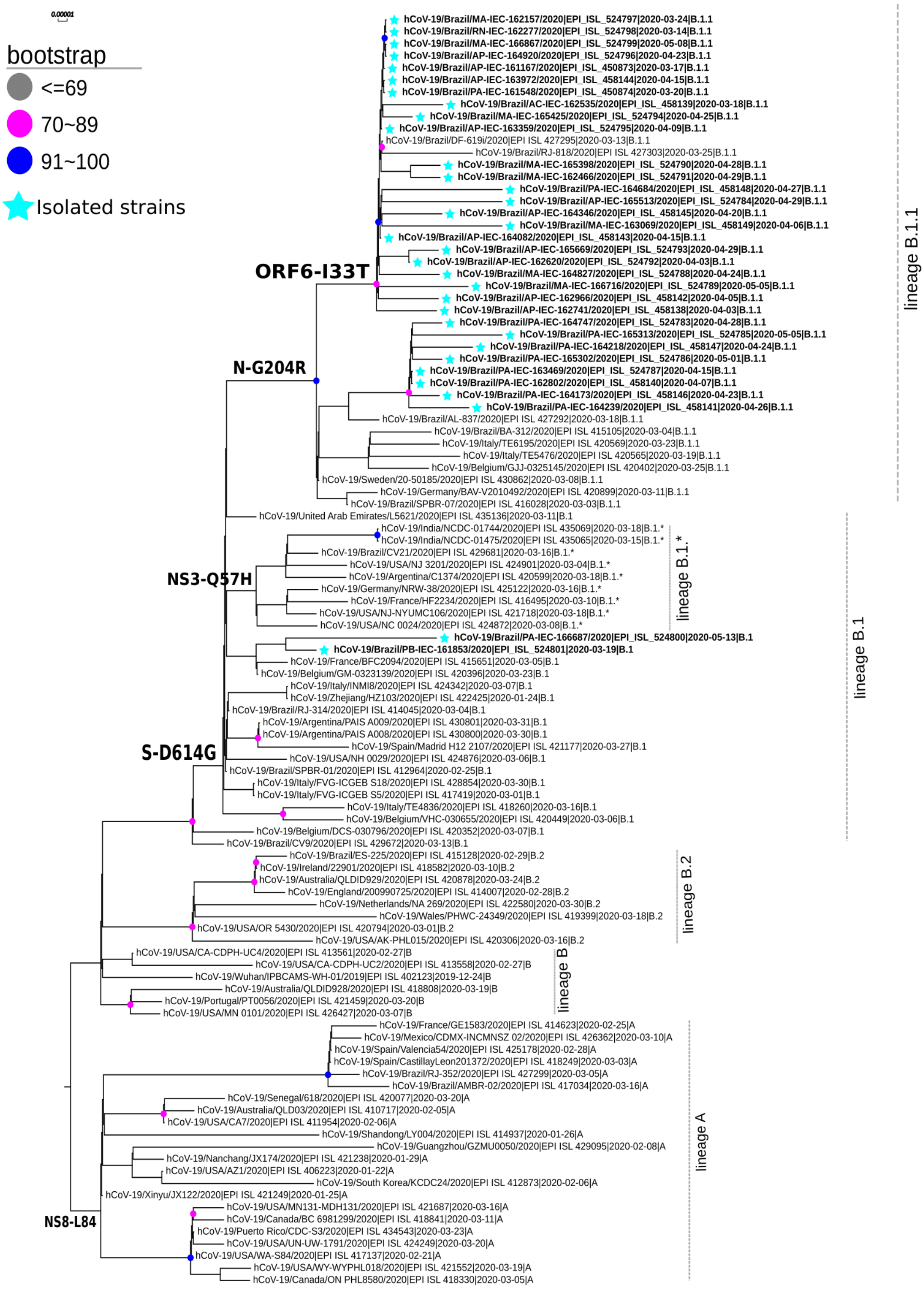


Figure 8 - SARS-CoV-2 complete genome phylogenetic tree (ML) with 1000 bootstraps, using GTR evolution model for nucleotide substitutions.


0.00001
bootstrap
<=69
70~89
91~100
Isolated strains
ORF6-I33T
N-G204R
NS3-Q57H
S-D614G
NS8-L84
lineage B.1.1
lineage B.1.*
lineage B.1
lineage B.2
lineage B
lineage A
hCoV-19/Brazil/MA-IEC-162157/2020|EPI_ISL_524797|2020-03-24|B.1.1
hCoV-19/Brazil/RN-IEC-162277/2020|EPI_ISL_524798|2020-03-14|B.1.1
hCoV-19/Brazil/MA-IEC-166867/2020|EPI_ISL_524799|2020-05-08|B.1.1
hCoV-19/Brazil/AP-IEC-164920/2020|EPI_ISL_524796|2020-04-23|B.1.1
hCoV-19/Brazil/AP-IEC-161167/2020|EPI_ISL_450873|2020-03-17|B.1.1
hCoV-19/Brazil/AP-IEC-163972/2020|EPI_ISL_458144|2020-04-15|B.1.1
hCoV-19/Brazil/PA-IEC-161548/2020|EPI_ISL_450874|2020-03-20|B.1.1
hCoV-19/Brazil/AC-IEC-162535/2020|EPI_ISL_458139|2020-03-18|B.1.1
hCoV-19/Brazil/MA-IEC-165425/2020|EPI_ISL_524794|2020-04-25|B.1.1
hCoV-19/Brazil/AP-IEC-163359/2020|EPI_ISL_524795|2020-04-09|B.1.1
hCoV-19/Brazil/DF-619i/2020|EPI ISL 427295|2020-03-13|B.1.1
hCoV-19/Brazil/RJ-818/2020|EPI ISL 427303|2020-03-25|B.1.1
hCoV-19/Brazil/MA-IEC-165398/2020|EPI_ISL_524790|2020-04-28|B.1.1
hCoV-19/Brazil/MA-IEC-162466/2020|EPI_ISL_524791|2020-04-29|B.1.1
hCoV-19/Brazil/PA-IEC-164684/2020|EPI_ISL_458148|2020-04-27|B.1.1
hCoV-19/Brazil/AP-IEC-165513/2020|EPI_ISL_524784|2020-04-29|B.1.1
hCoV-19/Brazil/AP-IEC-164346/2020|EPI_ISL_458145|2020-04-20|B.1.1
hCoV-19/Brazil/MA-IEC-163069/2020|EPI_ISL_458149|2020-04-06|B.1.1
hCoV-19/Brazil/AP-IEC-164082/2020|EPI_ISL_458143|2020-04-15|B.1.1
hCoV-19/Brazil/AP-IEC-165669/2020|EPI_ISL_524793|2020-04-29|B.1.1
hCoV-19/Brazil/AP-IEC-162620/2020|EPI_ISL_524792|2020-04-03|B.1.1
hCoV-19/Brazil/MA-IEC-164827/2020|EPI_ISL_524788|2020-04-24|B.1.1
hCoV-19/Brazil/MA-IEC-166716/2020|EPI_ISL_524789|2020-05-05|B.1.1
hCoV-19/Brazil/AP-IEC-162966/2020|EPI_ISL_458142|2020-04-05|B.1.1
hCoV-19/Brazil/AP-IEC-162741/2020|EPI_ISL_458138|2020-04-03|B.1.1
hCoV-19/Brazil/PA-IEC-164747/2020|EPI_ISL_524783|2020-04-28|B.1.1
hCoV-19/Brazil/PA-IEC-165313/2020|EPI_ISL_524785|2020-05-05|B.1.1
hCoV-19/Brazil/PA-IEC-164218/2020|EPI_ISL_458147|2020-04-24|B.1.1
hCoV-19/Brazil/PA-IEC-165302/2020|EPI_ISL_524786|2020-05-01|B.1.1
hCoV-19/Brazil/PA-IEC-163469/2020|EPI_ISL_524787|2020-04-15|B.1.1
hCoV-19/Brazil/PA-IEC-162802/2020|EPI_ISL_458140|2020-04-07|B.1.1
hCoV-19/Brazil/PA-IEC-164173/2020|EPI_ISL_458146|2020-04-23|B.1.1
hCoV-19/Brazil/PA-IEC-164239/2020|EPI_ISL_458141|2020-04-26|B.1.1
hCoV-19/Brazil/AL-837/2020|EPI ISL 427292|2020-03-18|B.1.1
hCoV-19/Brazil/BA-312/2020|EPI ISL 415105|2020-03-04|B.1.1
hCoV-19/Italy/TE6195/2020|EPI ISL 420569|2020-03-23|B.1.1
hCoV-19/Italy/TE5476/2020|EPI ISL 420565|2020-03-19|B.1.1
hCoV-19/Belgium/GJJ-0325145/2020|EPI ISL 420402|2020-03-25|B.1.1
hCoV-19/Sweden/20-50185/2020|EPI ISL 430862|2020-03-08|B.1.1
hCoV-19/Germany/BAV-V2010492/2020|EPI ISL 420899|2020-03-11|B.1.1
hCoV-19/Brazil/SPBR-07/2020|EPI ISL 416028|2020-03-03|B.1.1
hCoV-19/United Arab Emirates/L5621/2020|EPI ISL 435136|2020-03-11|B.1
hCoV-19/India/NCDC-01744/2020|EPI ISL 435069|2020-03-18|B.1.*
hCoV-19/India/NCDC-01475/2020|EPI ISL 435065|2020-03-15|B.1.*
hCoV-19/Brazil/CV21/2020|EPI ISL 429681|2020-03-16|B.1.*
hCoV-19/USA/NJ 3201/2020|EPI ISL 424901|2020-03-04|B.1.*
hCoV-19/Argentina/C1374/2020|EPI ISL 420599|2020-03-18|B.1.*
hCoV-19/Germany/NRW-38/2020|EPI ISL 425122|2020-03-16|B.1.*
hCoV-19/France/HF2234/2020|EPI ISL 416495|2020-03-10|B.1.*
hCoV-19/USA/NJ-NYUMC106/2020|EPI ISL 421718|2020-03-18|B.1.*
hCoV-19/USA/NC 0024/2020|EPI ISL 424872|2020-03-08|B.1.*
hCoV-19/Brazil/PA-IEC-166687/2020|EPI_ISL_524800|2020-05-13|B.1
hCoV-19/Brazil/PB-IEC-161853/2020|EPI_ISL_524801|2020-03-19|B.1
hCoV-19/France/BFC2094/2020|EPI ISL 415651|2020-03-05|B.1
hCoV-19/Belgium/GM-0323139/2020|EPI ISL 420396|2020-03-23|B.1
hCoV-19/Italy/INMI8/2020|EPI ISL 424342|2020-03-07|B.1
hCoV-19/Zhejiang/HZ103/2020|EPI ISL 422425|2020-01-24|B.1
hCoV-19/Brazil/RJ-314/2020|EPI ISL 414045|2020-03-04|B.1
hCoV-19/Argentina/PAIS A009/2020|EPI ISL 430801|2020-03-31|B.1
hCoV-19/Argentina/PAIS A008/2020|EPI ISL 430800|2020-03-30|B.1
hCoV-19/Spain/Madrid H12 2107/2020|EPI ISL 421177|2020-03-27|B.1
hCoV-19/USA/NH 0029/2020|EPI ISL 424876|2020-03-06|B.1
hCoV-19/Brazil/SPBR-01/2020|EPI ISL 412964|2020-02-25|B.1
hCoV-19/Italy/FVG-ICGEB S18/2020|EPI ISL 428854|2020-03-30|B.1
hCoV-19/Italy/FVG-ICGEB S5/2020|EPI ISL 417419|2020-03-01|B.1
hCoV-19/Italy/TE4836/2020|EPI ISL 418260|2020-03-16|B.1
hCoV-19/Belgium/VHC-030655/2020|EPI ISL 420449|2020-03-06|B.1
hCoV-19/Belgium/DCS-030796/2020|EPI ISL 420352|2020-03-07|B.1
hCoV-19/Brazil/CV9/2020|EPI ISL 429672|2020-03-13|B.1
hCoV-19/Brazil/ES-225/2020|EPI ISL 415128|2020-02-29|B.2
hCoV-19/Ireland/22901/2020|EPI ISL 418582|2020-03-10|B.2
hCoV-19/Australia/QLDID929/2020|EPI ISL 420878|2020-03-24|B.2
hCoV-19/England/200990725/2020|EPI ISL 414007|2020-02-28|B.2
hCoV-19/Netherlands/NA 269/2020|EPI ISL 422580|2020-03-30|B.2
hCoV-19/Wales/PHWC-24349/2020|EPI ISL 419399|2020-03-18|B.2
hCoV-19/USA/OR 5430/2020|EPI ISL 420794|2020-03-01|B.2
hCoV-19/USA/AK-PHL015/2020|EPI ISL 420306|2020-03-16|B.2
hCoV-19/USA/CA-CDPH-UC4/2020|EPI ISL 413561|2020-02-27|B
hCoV-19/USA/CA-CDPH-UC2/2020|EPI ISL 413558|2020-02-27|B
hCoV-19/Wuhan/IPBCAMS-WH-01/2019|EPI ISL 402123|2019-12-24|B
hCoV-19/Australia/QLDID928/2020|EPI ISL 418808|2020-03-19|B
hCoV-19/Portugal/PT0056/2020|EPI ISL 421459|2020-03-20|B
hCoV-19/USA/MN 0101/2020|EPI ISL 426427|2020-03-07|B
hCoV-19/France/GE1583/2020|EPI ISL 414623|2020-02-25|A
hCoV-19/Mexico/CDMX-INCMNSZ 02/2020|EPI ISL 426362|2020-03-10|A
hCoV-19/Spain/Valencia54/2020|EPI ISL 425178|2020-02-28|A
hCoV-19/Spain/CastillayLeon201372/2020|EPI ISL 418249|2020-03-03|A
hCoV-19/Brazil/RJ-352/2020|EPI ISL 427299|2020-03-05|A
hCoV-19/Brazil/AMBR-02/2020|EPI ISL 417034|2020-03-16|A
hCoV-19/Senegal/618/2020|EPI ISL 420077|2020-03-20|A
hCoV-19/Australia/QLD03/2020|EPI ISL 410717|2020-02-05|A
hCoV-19/USA/CA7/2020|EPI ISL 411954|2020-02-06|A
hCoV-19/Shandong/LY004/2020|EPI ISL 414937|2020-01-26|A
hCoV-19/Guangzhou/GZMU0050/2020|EPI ISL 429095|2020-02-08|A
hCoV-19/Nanchang/JX174/2020|EPI ISL 421238|2020-01-29|A
hCoV-19/USA/AZ1/2020|EPI ISL 406223|2020-01-22|A
hCoV-19/South Korea/KCDC24/2020|EPI ISL 412873|2020-02-06|A
hCoV-19/Xinyu/JX122/2020|EPI ISL 421249|2020-01-25|A
hCoV-19/USA/MN131-MDH131/2020|EPI ISL 421687|2020-03-16|A
hCoV-19/Canada/BC 6981299/2020|EPI ISL 418841|2020-03-11|A
hCoV-19/Puerto Rico/CDC-S3/2020|EPI ISL 434543|2020-03-23|A
hCoV-19/USA/UN-UW-1791/2020|EPI ISL 424249|2020-03-20|A
hCoV-19/USA/WA-S84/2020|EPI ISL 417137|2020-02-21|A
hCoV-19/USA/WY-WYPHL018/2020|EPI ISL 421552|2020-03-19|A
hCoV-19/Canada/ON PHL8580/2020|EPI ISL 418330|2020-03-05|A

DISCUSSION

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

The detection of SARS-CoV-2 has occurred in all age groups. However, the highest frequency has occurred in adults and the elderly, as described in the world population[10–13]. The low frequency of SARS-CoV-2 in children and teenagers under 20, verified in the present study, has been associated with reduced susceptibility and less likelihood to infection or a combination of both, compared to adults[14–16].

As for sex, similar to what has been reported by literature, the frequency amongst men was slightly higher than in women[17,18]. Estrogen, the main female sex hormone, plays a possible protective role in COVID-19, activating the immune response, and directly suppressing SARS-CoV-2 replication[19,20]. Indeed, estrogen inhibits the activity or expression of different components of the renin-angiotensin system. Particularly, estrogen can upregulate ACE2 expression[21,22]. Regarding the clinical analysis presented by the investigated patients, the most described symptoms were fever, cough, sore throat and dyspnoea, respectively, associated or not, commonly reported amongst respiratory infections, as well as in COVID-19[10,11,23,24]. However, unlike other respiratory infections, in COVID-19 anosmia and dysgeusia have been frequently reported[25–29], as described amongst patients in the study region. After the onset of symptoms, the period of greatest detection of SARS-CoV-2 by RT-qPCR has occurred on the fifth day, similarly to what has been described in other studies[30].

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

Amongst the alterations in the Spike protein that plays a role in binding to the human ACE2 receptor and is also the main antigenic target, it was found the D614G substitution that is described as a factor that antigenically favors the virus, giving it a higher capacity to infection[33] and has been used as a genetic marker for strains of the B-lineage (Pangolin Classification) which has become the largest circulating group worldwide[34]. Also, it was verified the V1176F mutation described in the literature[35] and used as a genetic marker for samples circulating in Brazil (https://www.gisaid.org), but no antigenic advantage has yet been attributed.

Regarding the N protein plays a role in folding viral genetic material and has been used as a marker for samples from Europe, it was verified R203K, G204R and I292T amino acidic substitutions. However, their molecular roles are still unclear[36,37]. The change I33T in ORF6, a non-structural protein, has been observed in samples originating in Brazil and that circulate in South America[38].

The phylogenetic analysis revealed that the samples of this study have formed three distinct groups that cluster with the phylogenetic lineages B.1 and B.1.1 that have samples already sequenced from Brazil[39]. Within clade B.1, only two samples from Pará were clustering with samples from Europe. In clade B.1.1, it was possible to observe the formation of two distinct groups divided by the I33T ORF6 and V1176F S protein substitutions. These two mutations have been observed to divide the two main strains of SARS-CoV-2 circulating in Brazil[34]. Since its worldwide circulation on December 2019[40], the SARS-CoV-2 genome has changed wherever it arrives[41], which may mean a likely adaptation to the population[42].

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

In conclusion, this study reveals that the highest SARS-CoV-2 circulation has reached its peak in epidemiological week 18. The distribution of positive samples by age group has demonstrated that the median age was 47, with men being the main affected gender and there was a spectrum of symptoms composed of fever, cough, dyspnoea and sore throat. Furthermore, this investigation supports the evidence for the existence of two main lineages (B.1 and B.1.1) associated with genomic epidemiology of SARS-CoV-2 in the Amazon region. Thus, genomic surveillance must be continuously adopted to be able to offer accurate and quality data to understand where this virus emerged from, and map the transmission network to improve operational decisions in public health.

METHODS

Samples and ethical aspects

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

Extraction and Detection by RT-qPCR of viral nucleic acid

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

The amplification reaction was conducted sequentially in the following steps: reverse transcription at 50°C for 15 minutes, followed by transcriptase inactivation and activation of Taq DNA polymerase at 95 ° C for 2 minutes, polymerase chain reaction at 95°C for 15 seconds in 45 cycles, extension and annealing at 55 ° C for 30 seconds. At the end of the amplification, all clinical samples should have reaction 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.

Epidemiological analysis

Graphs of epidemiological data (age, sex, state, signs and symptoms) and circulation were performed with support by the LVR-IEC database and the Microsoft Office Excel program. The data were inspected, visualized and plotted using the R programming language script[43] together with the libraries ggplot2[44], geobr[45], pipeR[46], readr[47], lubridate, fmsb[48], plyr[49], scales[50], viridis[51] and hrbrthemes[52]. By international convention, epidemiological weeks were counted from Sunday to Saturday, considering the sample collection date.

Sample selection for sequencing

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

Library construction and sequencing

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

Data pre-processing

The data were evaluated for their quality regions. The adapters sequences reads with a quality lower than Phred 20 and reads with less than 40 bp size, were removed using Trimmommatic[53]. The processed reads were visualized with FastQC[54]. For Trimmomatic, we have used the following parameters: LEADING:3 TRAILING:3 MINLEN:40

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

For this step, the reads validated based on quality trimming were used to assembly the SARS-CoV-2 genomes. The De Novo assembly was performed using the Megahit v.1.1.4-2[55] and for Reference Mapping we have used the software Bowtie2[56] and Geneious Prime, where the respective coverage, gaps and final size of the genome were analyzed. For genome assembly, all programs were performed with default parameters.

Taxonomic annotation and submission to GISAID

The generated de novo contigs were compared using the Blastx tool[57] implemented in Diamond v.0.9.33[58], against the RefSeq database (NCBI's Protein Reference Sequences Database), which is a database of cured protein sequences and which provides a high level of annotation, such as the description of the function of a protein, its domain structure, post-translational modifications, where a statistical value (e-value) of 0.0001 was considered.

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

Phylogenetic Analysis and Mutation analysis

The genomes sequences were aligned with other genomes from all the world using the Mafft v.7.471[59]. For phylogenetic analysis the software RaXML[60] with 1000 bootstraps was used as statistical support, using GTR as a nucleotide substitution model. The genomes obtained were compared to the reference strain (NC_045512) by *in house* python script that compares each base of the entire genome and gives us a mutation list.

REFERENCES

1. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544 (2020).

2. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).

3. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. *WHO Coronavirus Disease (COVID-19) Dashboard* (2020).

4. World Health Organization. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). *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)* (2020).

5. Ministry of health. COVID-19 in Brazil. *http://susanalitico.saude.gov.br/#/dashboard/* (2020).

6. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).

7. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).

8. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* 2020.04.17.046086 (2020). doi:10.1101/2020.04.17.046086

9. Stefanelli, P. *et al.* Whole genome and phylogenetic analysis of two SARSCoV-2 strains isolated in Italy in January and February 2020: Additional clues on multiple introductions and further circulation in Europe. *Eurosurveillance* **25**, 1–5 (2020).

10. Guan, W. *et al.* Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).

11. Jiang, F. *et al.* Review of the Clinical Characteristics of Coronavirus Disease 2019 (COVID-19). *J. Gen. Intern. Med.* **35**, 1545–1549 (2020).

12. Shahid, Z. *et al.* COVID-19 and Older Adults: What We Know. *J. Am. Geriatr. Soc.* **68**, 926–929 (2020).

13. Nikolich-Zugich, J. *et al.* SARS-CoV-2 and COVID-19 in older adults: what we may expect regarding pathogenesis, immune responses, and outcomes. *GeroScience* **42**, 505–514 (2020).

14. Davies, N. G. *et al.* Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat. Med.* (2020). doi:10.1038/s41591-020-0962-9

15. Mantovani, A. *et al.* Coronavirus disease 2019 (COVID-19) in children and/or adolescents: a meta-analysis. *Pediatr. Res.* **2019**, (2020).

16. Ludvigsson, J. F. Systematic review of COVID-19 in children shows milder cases and a better prognosis than adults. *Acta Paediatr. Int. J. Paediatr.* **109**, 1088–1095 (2020).

17. Li, L. quan *et al.* COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. *J. Med. Virol.* **92**, 577–583 (2020).

18. Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R. & Klein, S. L. Impact of sex and gender on COVID-19 outcomes in Europe. *Biol. Sex Differ.* **11**, 1–13 (2020).

19. Channappanavar, R. *et al.* Sex-Based Differences in Susceptibility to Severe Acute Respiratory Syndrome Coronavirus Infection. *J. Immunol.* **198**, 4046–4053 (2017).

20. Scully, E. P., Haverfield, J., Ursin, R. L., Tannenbaum, C. & Klein, S. L. Considering how biological sex impacts immune responses and COVID-19 outcomes. *Nat. Rev. Immunol.* (2020). doi:10.1038/s41577-020-0348-8

21. Bukowska, A. *et al.* Protective regulation of the ACE2/ACE gene expression by estrogen in human atrial tissue from elderly men. *Exp. Biol. Med.* **242**, 1412–1423 (2017).

22. Scully, E. P., Haverfield, J., Ursin, R. L., Tannenbaum, C. & Klein, S. L. Considering how biological sex impacts immune responses and COVID-19 outcomes. *Nat. Rev. Immunol.* (2020). doi:10.1038/s41577-020-0348-8

23. Rodríguez-Cola, M. *et al.* Clinical features of coronavirus disease 2019 (COVID-19) in a cohort of patients with disability due to spinal cord injury. *Spinal Cord Ser. Cases* **6**, (2020).

24. Ge, H. *et al.* The epidemiology and clinical information about COVID-19. *Eur. J. Clin. Microbiol. Infect. Dis.* **39**, 1011–1019 (2020).

25. Vaira, L. A., Salzano, G., Deiana, G. & De Riu, G. Anosmia and Ageusia: Common Findings in COVID-19 Patients. *Laryngoscope* **130**, 1787 (2020).

26. Vaira, L. A., Salzano, G., Fois, A. G., Piombino, P. & De Riu, G. Potential pathogenesis of ageusia and anosmia in COVID-19 patients. *Int. Forum Allergy Rhinol.* **00**, 1–2 (2020).

27. Russell, B. *et al.* Anosmia and ageusia are emerging as symptoms in patients with COVID-19: What does the current evidence say? *Ecancermedicalscience* **14**, 9–10 (2020).

28. Klopfenstein, T. *et al.* Features of anosmia in COVID-19. *Med. Mal. Infect.* 4–7 (2020). doi:10.1016/j.medmal.2020.04.006

29. Whittaker, A., Anson, M. & Harky, A. Neurological Manifestations of COVID-19: A systematic review and current update. *Acta Neurol. Scand.* **142**, 14–22 (2020).

30. Rodríguez-Cola, M. *et al.* Clinical features of coronavirus disease 2019 (COVID-19) in a cohort of patients with disability due to spinal cord injury. *Spinal Cord Ser. Cases* **6**, (2020).

31. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, (2020).

32. da Silva, S. J. R., da Silva, C. T. A., Mendes, R. P. G. & Pena, L. Role of Nonstructural Proteins in the Pathogenesis of SARS-CoV-2. *J. Med. Virol.* 3–5 (2020). doi:10.1002/jmv.25858

33. Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* (2020). doi:10.1016/j.cell.2020.06.043

34. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 2–4 (2017).

35. Gonçalves, R. L. *et al.* SARS-CoV-2 mutations and where to find them: An in silico perspective of structural changes and antigenicity of the Spike protein. *bioRxiv* **3**, 2020.05.21.108563 (2020).

36. Yin, C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics* **19**, 1–12 (2020).

37. Castillo, A. E. *et al.* Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *J. Med. Virol.* 1–5 (2020). doi:10.1002/jmv.25797

38. Resende, P. C. *et al.* Genomic surveillance of SARS-CoV-2 reveals community transmission of a major lineage during the early pandemic phase in Brazil. *bioRxiv* 2020.06.17.158006 (2020). doi:10.1101/2020.06.17.158006

39. Candido, D. S. *et al.* Evolution and epidemic spread of SARS-Cov-2 in Brazil. *Science (80-.).* **21**, 1–9 (2020).

40. Riou, J. & Althaus, C. L. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance* **25**, 1–5 (2020).

41. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **18**, 1–9 (2020).

42. Cao, Y. *et al.* Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.* **6**, 4–7 (2020).

43. R Foundation for Statistical Computing. R Core Team (2018). R: A language and environment for statistical computing. (2018).

44. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2009).

45. Pereira, R.H.M.; Gonçalves, C. N. geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil.GitHub repository. (2019).

46. Ren, K. pipeR: Multi-Paradigm Pipeline Implementation. R package version

0.6.1.3. (2016).

47. Hadley Wickham, J. H. and R. F. readr: Read Rectangular Text Data. R package version 1.3.1. (2018).

48. Nakazawa, M. fmsb: Functions for Medical Statistics Book with some Demographic Data. R package version 0.7.0. https://CRAN.R-project.org/package=fmsb (2019).

49. Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. (2011).

50. Seidel, H. W. and D. scales: Scale Functions for Visualization. R package version 1.1.1. (2020).

51. Simon Garnier. viridis: Default Color Maps from 'matplotlib'. R package version 0.5.1. (2018).

52. Bob Rudis. hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'. R package version 0.8.0. 2020

53. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

54. Andrews, S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/project (2010).

55. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

56. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, (2009).

57. Madden, T. & Coulouris, G. BLAST+ User Manual. *Ncbi* 1–64 (2008).

58. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

59. Katoh, K. & Toh, H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**, 1899–900 (2010).

60. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–3 (2014).

Acknowledgements

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

Author contributions

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

Additional Information

The authors declare that they have no conflicting interests.

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

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

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

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

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

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

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

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