RECOMMENDED CITATION
do Nascimento VA, Corado ALG, do Nascimento FO, da Costa AKA, Duarte DCG, Luz SLB, et al. Genomic and phylogenetic characterization of an imported case of SARS-CoV-2 in Amazonas State, Brazil [Submitted]. Mem Inst Oswaldo Cruz E-pub: 15 June 2020.doi: 10.1590/0074-02760200310.

Genomic and phylogenetic characterization of an imported case of SARS-CoV-2 in Amazonas State, Brazil.

Valdinete Alves do Nascimento [1,2], André Lima Guerra Corado [1,2], Fernanda Oliveira do Nascimento [1,3], Ágatha Kelly Araújo da Costa [1,3], Debora Camila Gomes Duarte [1], Michele Silva de Jesus [1], Sérgio Luiz Bessa Luz [1,3], Luciana Mara Fé Gonçalves [4], Cristiano Fernandes da Costa [4], Edson Delatorre [5], Felipe Gomes Naveca [1,2,3,6]

[1] Instituto Leônidas e Maria Deane, Fiocruz, Amazonas, Brasil.

[2] Programa de Pós-Graduação em Biologia Celular e Molecular Fundação - Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brasil.

[3] Programa de Pós-Graduação em Biologia da Interação Patógeno-Hospedeiro - Instituto Leônidas e Maria Deane, Amazonas, Brasil.

[4] Fundação de Vigilância em Saúde do Amazonas, Amazonas, Brasil.

[5] Departamento de Biologia, Centro de Ciências Exatas, Naturais e da Saúde, Universidade Federal do Espírito Santo, Espírito Santo, Brasil.

[6] Rede Genômica em Saúde do Estado do Amazonas, Amazonas, Brasil.

[*]Address correspondence to Felipe Gomes Naveca, felipe.naveca@fiocruz.br.

https://orcid.org/0000-0002-2888-1060

SUMMARY

A new coronavirus (SARS-CoV-2) is currently causing a life-threatening pandemic. In this study, we report the complete genome sequencing and genetic characterization of a SARS-CoV-2 detected in Manaus, Amazonas, Brazil, and the protocol we designed to generate high-quality SARS-CoV-2 full genome data. The isolate was obtained from an asymptomatic carrier returning from Madrid, Spain. Nucleotide sequence analysis showed a total of nine mutations in comparison with the original Human case in Wuhan, China, and support this case as belonging to the recently proposed lineage A.2. Phylogeographic analysis further confirmed the likely European origin of this case. To our knowledge, this is the first SARS-CoV-2 genome obtained from the North Brazilian Region. We believe that the information generated in this study may contribute to the ongoing efforts toward the SARS-CoV-2 emergence.

Key words: Coronavirus; SARS-CoV-2, COVID-19, Brazil; Amazon region; Genome.

Sponsorships: CNPq / CAPES / MS-DECIT / Fiocruz / FAPEAM - REGESAM

The coronavirus disease 19 (COVID-19) is caused by infection with the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) that was identified for the first time in patients with pneumonia in Wuhan, China [1,2]. Since its discovery at the end of 2019, SARS-CoV-2 transmission has been documented as being person-to-person, causing from an asymptomatic status, that may also generate transmission clusters, to a symptomatic disease with the typical symptoms manifested as fever, dry cough, myalgia, fatigue, dyspnea, diarrhea, and nausea [3,4]. In the majority of the patients, the clinical outcome is a mild disease, although McGoogan and Wu (2020) [5] described Chinese patients that developed a severe outcome (16%) or a critical condition (4%). Severe or critical outcomes usually occur in patients with comorbidities, and the disease can progress, presenting arrhythmia and shock, that could evolve to death [5,6].

The SARS-CoV-2 belongs to family *Coronaviridae*, genera *Betacoronavirus*. In 2002 and 2012, two outbreaks occurred caused by new coronaviruses, SARS-CoV and MERS-CoV, with lethality ranging from nine to 33%, respectively [7–9]. The coronaviruses are enveloped viruses, with 80 to 120 nm in diameter. Three proteins compose the virion surface: spike (S), membrane (M), and small membrane protein (E), giving the virus a crown-like visual on electron micrograph [10,11]. This viral family has the largest RNA genome of all other RNA viruses, with a single non-segmented positive-stranded RNA of approximately 30 kb [12].

In Brazil, the first recorded case of SARS-CoV-2 infection occurred in the end of February, in the city of São Paulo, followed by cases in the Northeast (Bahia), Central-west (Brasília) and South (Rio Grande do Sul) regions (https://www.saude.gov.br/images/pdf/2020/April/06/2020-04-06-BE7-Boletim-Especial-do-COE-Atualizacao-da-Avaliacao-de-Risco.pdf).

The northern of Brazil was the last region of the country to detected SARS-CoV-2 in its population. In March 15th, the first case in the state of Amazonas was detected in a woman that traveled to London and return to the Amazonas capital, Manaus (http://www.fvs.am.gov.br/media/publicacao/Boletim_Situação_Epidemiológica_de_COVID-19_e_da_S%C3%ADndrome_Respiratória_Aguda__g9E6Skz.pdf). After three days, the second case of SARS-CoV-2 was detected in Manaus and characterized in the present study.

One 56-years-old man returning from Madrid, Spain, arrived asymptomatic in Manaus, Amazonas State, Brazil at 15-Mar-2020. At that time, a massive outbreak of COVID-19 was already established in several European countries, suggesting that travelers should be kept in quarantine at arrival. Even without the symptoms of a respiratory infection, nasal and oropharyngeal swabs were collected for SARS-CoV-2 testing as a routine established for respiratory virus surveillance at Instituto Leônidas e Maria Deane (ILMD) – Fiocruz, Amazonas State, Brazil, since 2019. The combined swab sample was collected and submitted to total nucleic acid extraction with a commercial kit (Biogene) and immediately evaluated using the reverse transcription real-time polymerase chain reaction (RT-qPCR) protocol developed by the US Centers for Disease Control and Prevention (CDC/USA)[1]. This RT-qPCR assay employs different primers and probes sets aiming three regions of the SARS-CoV-2 nucleocapsid (N) gene (https://www.fda.gov/media/134922/download), and the human RNase P as an internal control.

The analyzed sample tested positive for SARS-CoV-2 with Cts values of 14.43 (N1), 15.39 (N2) and 15.33 (N3). Thus, we immediately generated cDNA with random

[1] On 15-Mar-2020, CDC updated the RT-qPCR protocol removing the N3 target.

primers and Superscript IV reverse transcriptase (ThermoFisher Scientific). We had previously designed a PCR scheme to amplify the entire genome of the SARS-CoV-2 based on an alignment of all complete genome sequences available on GenBank at 03/03/2020. Conserved regions were chosen for primer design with Primer3 v2.3.7. embedded in Geneious Software 10.2.6 [13], spanning around 2.2Kb and with an overlap region between 131 and 225bp. This PCR scheme resulted in 15 amplicons with further details presented in the supplemental file 1 of this manuscript.

The SARS-CoV-2 whole genome was amplified with Platinum SuperFi II Green PCR master mix (ThermoFisher Scientific) using the 15 primers sets in individual reactions. Each amplicon was then visualized as a unique and intense DNA fragment on agarose gel electrophoresis stained with GelRed (Biotium). The PCR amplicons were precipitated with molecular biology grade Polietilenoglicol (PEG) 8,000 (Promega) and then resuspended in nuclease-free water. After a one-hour incubation at 37ºC, all amplicons were quantified in ng/uL using Qubit 2.0 and the dsDNA HS assay kit (Thermo Fisher Scientific). Finally, the number of DNA copies in each purified amplicon was estimated with ENDMEMO (http://www.endmemo.com/bio/dnacopynum.php), normalized, and pooled. A single library was constructed using the Nextera DNA Flex Library Prep and clustered with MiSeq Reagent Micro Kit v2 (300-cycles), following the manufacturer's protocols. Nucleotide sequencing was performed in the MiSeq platform (Illumina), installed at ILMD, in a paired-end run (2x150 cycles).

A total of 10,946,898 reads were trimmed for quality and adapters using BBDUK v37.25, embedded in Geneious software. Thus, 8,362,418 reads were mapped to the SARS-CoV-2 NCBI Reference Sequence NC_045512.2 using Geneious map-to-reference tool. The BR_AM_ILMD_20140001 final consensus genome sequence

contains 29,789 nucleotides, with no gaps, a Q40 score of 100%, with no undetermined "N" bases and high average coverage (>34,000X). To avoid any primers bias, we removed both primer binding sites at the 5′ and 3′ ends. Thus, our final sequence represents the positions between nucleotides 47 and 29,835 of the NCBI RefSeq previously mentioned.

We aligned the BR_AM_ILMD_20140001 genomic sequence with the SARS-CoV-2 NCBI Reference Sequence NC_045512 using MAFFT v7.388 [(14)] to investigate any mutations throughout genome. A total of nine mutations were observed at nucleotide positions 8,782 (C to T); 9,477 (T to A); 12,781 (C to T); 14,805 (C to T); 25,979 (G to T); 26,642 (C to T); 28,144 (T to C) ; 28,657 (C to T) and 28,863 (C to T), with four of these leading to residues substitution in the deduced protein sequences (Table 1).

To put the BR_AM_ILMD_20140001 genome in a global context, we aligned the new genome to a pool of all SARS-CoV-2 genomes with at least 25,000 nucleotides available at GISAID database [(15)] on March 31st, 2020 using MAFFT. We adopted a subsampling strategy to reduce computation time, selecting subsets of 15-20 sequences retaining the most viral diversity from each country using the CD-HIT program [(16)]. This subsampling approach resulted in a final sequence dataset with 490 sequences from 53 countries (supplemental table S1).

Complete coding sequences (CDS) were subjected to maximum likelihood (ML) phylogenetic reconstruction with PhyML v3.0 [(17)], under the HKY+Γ4 nucleotide substitution model. According to the ML phylogenetic tree, the BR_AM_ILMD_20140001 sequence belonged to the lineage A [(18)] (Figure 1A) clustering within a monophyletic cluster (aLRT = 1.00) comprising sequences from Spain, Chile, France, Greece, Georgia, Netherlands, Senegal (Figure 1B). The pangolin

program (github.com/hCoV-2019/pangolin) further assigned the sequences from this cluster to the A.2 lineage (Figure 1B).

After temporal validation with Tempest [19], we conducted a Bayesian discrete spatiotemporal analysis with all lineage A sequences (Figure 2) using the BEAST v1.10 package [20], applying the strict molecular clock and exponential coalescent models. We estimated that the most recent common ancestor (MRCA) of lineage A.2 originated in Spain (posterior state probability, PSP = 0.99) in the beginning of February 2020 (95% highest posterior density, HPD = 2nd – 22nd Feb 2020). The phylogeographic analysis also pointed out that the BR_AM_ILMD_20140001 was most probably introduced from Spain (PSP = 0.43) (Figure 2).

Like other viral infections, the infection by SARS-CoV-2 can be asymptomatic and this fact has been reported in different studies [21–23]. It is noteworthy that according to the definition of suspected case adopted in Brazil at the time that we investigated this asymptomatic carrier, he would not be included in SARS-CoV-2 testing, despite returning from an area with active transmission like Madrid, Spain (https://www.saude.gov.br/images/pdf/2020/marco/24/03--ERRATA---Boletim-Epidemiologico-05.pdf). Until 24-Mar-2020, when we deposited the BR_AM_ILMD_20140001 genome information at https://www.gisaid.org, there were only 17 Brazilian SARS-CoV-2 complete genome sequences, and the sequence reported in the present work is the first complete genome from the northern Brazilian region.

Several laboratories over the world are now sequencing thousands of SARS-CoV-2 genomes, which is undoubtedly the most notable effort of viral sequencing in human history. Like any other virus, the new coronavirus is continuously evolving as more hosts, humans or animals, are getting infected. Thus, it is of paramount importance to

generate and share high-quality full viral genomes from different regions over the world to better understand the SARS-CoV-2 evolution. The information related to the viral evolution is not only necessary for molecular epidemiology studies, but also to monitor if the newly identified mutations are linked to different clinical presentations or may drive into false-negative results when performing nucleic acid amplification assays, like real-time PCR.

Therefore, in this work, we aimed to describe and characterize the complete genome of the SARS-CoV-2 obtained from an asymptomatic carrier returning from Madrid, Spain. The final sequence of the sample BR_AM_ILMD_20140001 showed high quality (Q40 = 100%) and coverage (mean 34,592X), with no ambiguities that could affect further genome analyses.

To achieve this goal, we decided to use a nucleotide sequencing strategy where firstly all the 15 amplicons, encompassing the entire SARS-CoV-2 genome, were confirmed by agarose gel electrophoresis. Subsequently, each amplicon was quantified and normalized in order to prevent that one region could be overrepresented during sequencing. In order to make our approach more straightforward, we decided to evaluate if longer amplicons could be generated in a very similar way. We were successful in generating amplicons around 6Kb, with a minimum overlap of 131bp, reducing the number of PCR reactions to 5 instead of 15 (data not showed). We believe that this approach may be exciting not only to reduce the current protocol costs, but also for those interested in using long reads sequencing technologies like PacBio SMS and nanopore.

Recently, Rambaut and colleagues proposed a rational and dynamic virus classification for SARS-CoV-2 genomes based on a phylogenetic framework [(18)]. Using

this approach authors identified at the root of the phylogeny of SARSCoV-2 two lineages that were simply denoted as lineages A and B. In our analysis, while all Brazilian SARS-CoV-2 sequences belonged to the lineage B, the BR_AM_ILMD_20140001 genome clustered within the lineage A.2, indicating that BR_AM_ILMD_20140001 strain belongs to a distinct transmission cluster than the other full Brazilian genomes reported until March 31st, 2020. The lineage A.2 constitutes a predominantly Spanish lineage, found in at least 12 countries [18] and our phylogeographic analysis corroborates its origins and suggests the importation from Spain to Brazil.

In this study, we report and characterize the first SARS-CoV-2 genome obtained from an infected subject in the Brazilian North Region. Since this case was an asymptomatic carrier, it is not easy to suggest when infection has occurred. However, our phylogeographic analysis strongly indicates this individual was infected in Spain. Finally, we would like to emphasize that more fully genomes studies of the SARS-CoV-2 are necessary to better understand the evolution of this emerging life-threatening virus and the information of the nucleotide sequence described here may contribute to future molecular epidemiological studies in Brazil. In this sense, the protocol that we described in the present study may be useful to aid other researchers to generate other high-quality SARS-CoV-2 genomes.

Nucleotide sequence accession number:

The complete genome sequence of the BR_AM_ILMD_20140001 isolate is available in GISAID since March 24, 2020, under the ID number EPI_ISL_417034.

Acknowledgements

We gratefully acknowledge the following authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based. All GISAID accession numbers and data contributors are described in the supplemental file 2.

FUNDING

FGN is funded by Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM (www.fapeam.am.gov.br, (PCTI-EmergeSaúde/AM), call N°005/2020 and Rede Genômica de Vigilância em Saúde - REGESAM; Conselho Nacional de Desenvolvimento Científico e Tecnológico (http://www.cnpq.br, grant 440856/2016-7) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (http://www.capes.gov.br, grants 88881.130825/2016-01 and 88887.130823/2016-00). The authors thank the Program for Technological Development in Tools for Health-PDTIS FIOCRUZ for use of nucleotide sequencing facilities at ILMD—Fiocruz Amazônia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR'S CONTRIBUTION:

VAN, SLBL, CFC and FGN conceived the study. FGN designed the study protocol. VAN, ALGC, FON, AKAC, DCGD, LMFG, MSJ and FGN performed the molecular tests. VAN, ED and FGN performed the analysis and interpretation of these data. FON and MSJ collected biological sample. VAN, ALGC, AKAC, FON, DCGD, LMFG, ED and FGN wrote the manuscript. SLBL, CFC, ED and FGN critically revised

the manuscript for intellectual content. FGN financed the study. All authors read and approved the final manuscript.

REFERENCES

1. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579(7798): 270–3.
2. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020; 579(7798): 265–9.
3. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med. 2020; 382(13): 1199–207.
4. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020; 368(6490): 489–93.
5. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020; 323(13): 1239–42.
6. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395(10223): 497–506.
7. Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet. 2003; 61(9366): 1319–25.
8. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012; 367(19): 1814–20.
9. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol. 2016; 24(6): 490–502.
10. Bond CW, Leibowitz JL, Robb JA. Pathogenic murine coronaviruses. II. Characterization of virus-specific proteins of murine coronaviruses JHMV and A59V. Virology. 1979 94(2): 371–84.
11. Weiss SR, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. Microbiol Mol Biol Rev. 2005; 69(4): 635–64.
12. Lai MMC, Cavanaght D. The Molecular Biology of Coronaviruses. Adv Viral Res. 1997;48:1–100.
13. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012; 28(12): 1647–9.
14. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30(14): 3059–66.

15. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017; 22(13): 957.
16. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28(23): 3150–2.
17. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online - A web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 2005; 33(SUPPL. 2): 557–9.
18. Rambaut A, Holmes EC, Hill V, O'Toole Á, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. Preprint. 2020; 395(10224) : 514–65. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.04.17.046086
19. Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016; 2(1): 1–7.
20. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018; 4(1): 1–5.
21. Pan X, Chen D, Xia Y, Wu X, Li T, Ou X, et al. Asymptomatic cases in a family cluster with SARS-CoV-2 infection. Lancet Infect Dis. 2020; 20(4): 410–1.
22. Hoehl S, Rabenau H, Berger A, Kortenbusch M, Cinatl J, Bojkova D, et al. Evidence of SARS-CoV-2 Infection in Returning Travelers from Wuhan, China. N Engl J Med. 2020; 382(13): 1278–80.
23. Bai Y, Yao L, Wei T, Tian F, Jin D-Y, Chen L, et al. Presumed Asymptomatic Carrier Transmission of COVID-19. JAMA. 2020; 323(14): 1406–7.

Figure Legends

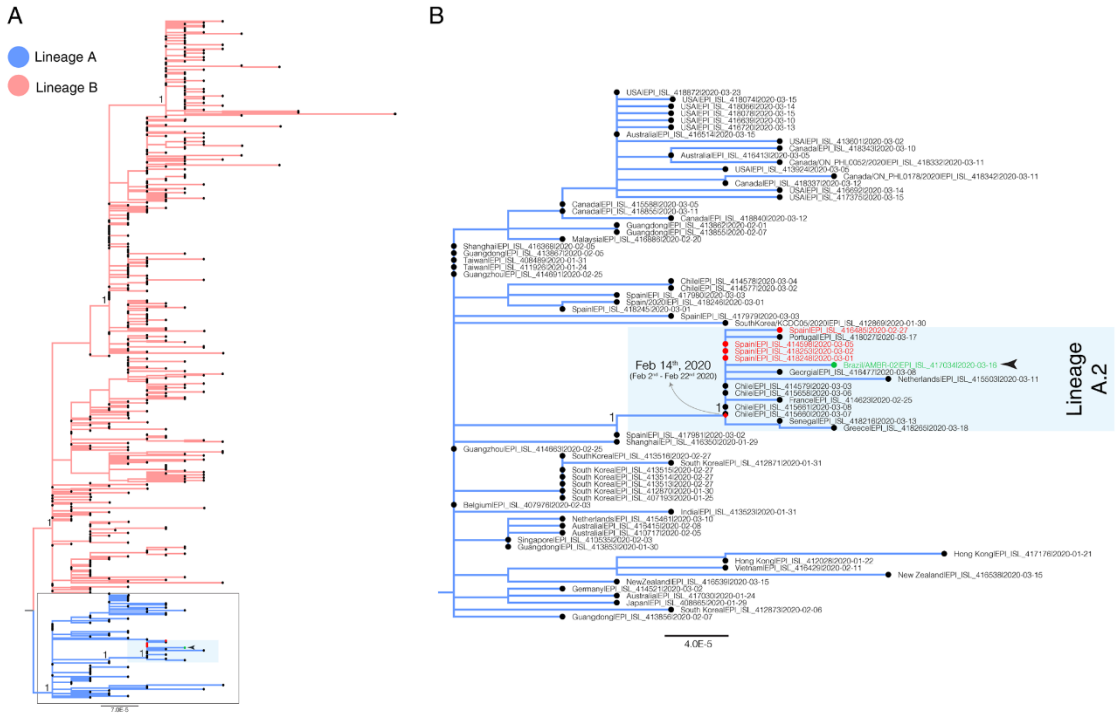

Figure 1. Maximum-likelihood phylogeny of subsampled SARS-CoV-2 genomes. A) ML tree rooted on the branch separating lineages A and B sequences colored following the legend. B) A close view of the Lineage A showing the lineage A.2 (light blue box) that comprises the BR_AM_ILMD_20140001 strain (indicated with an arrow). The nodes representing the MRCA of the lineage A.2 is indicated with a red diamond. In both trees, tips representing the Spanish strains inside lineage A.2 are colored red and the scale bar represents nucleotide substitutions per site.

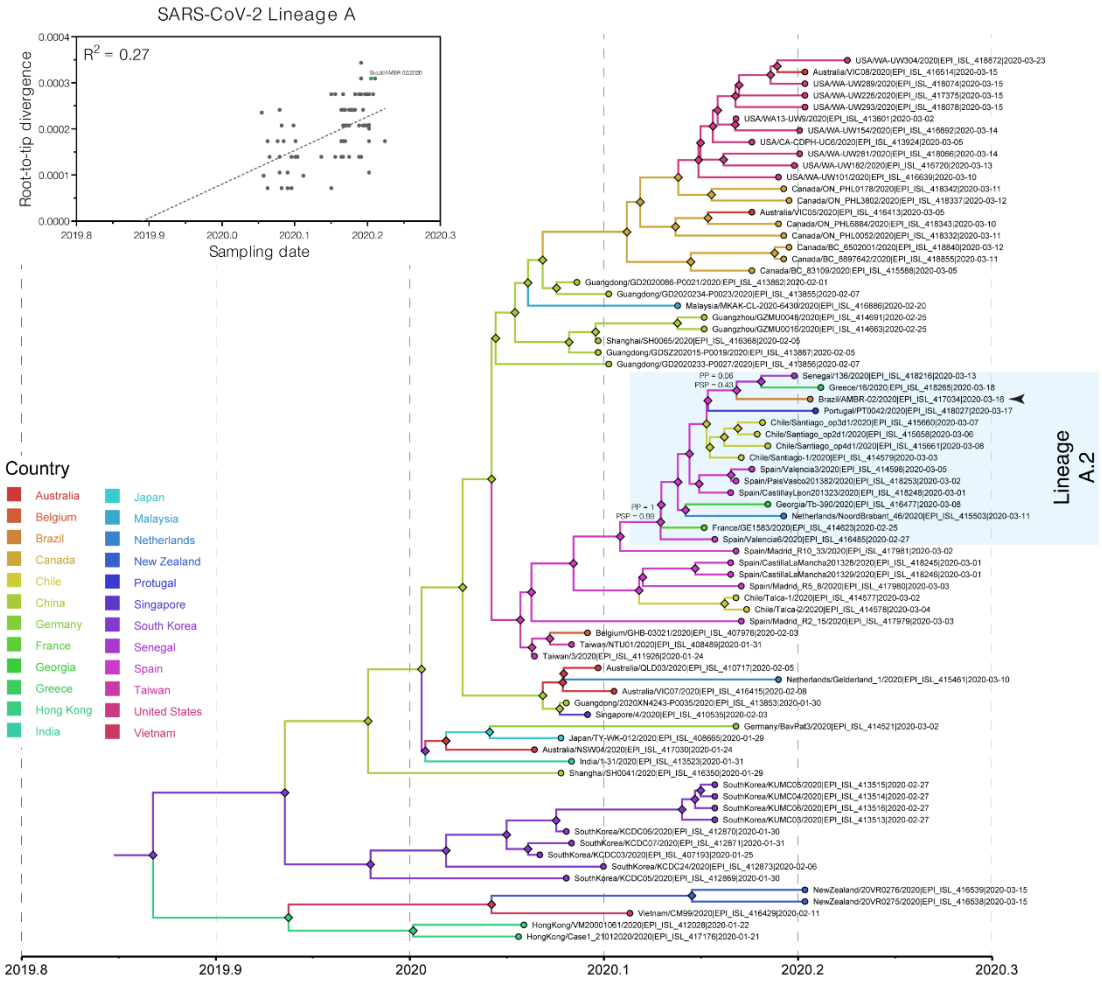

Figure 2. Phylogeography of the SARS-CoV-2 lineage A. A) Temporal signal analysis correlating the sampling date of each sequence and its genetic distance from the root of a maximum likelihood phylogeny. B) Time-scaled Bayesian phylogeographic MCC tree of the SARS-CoV-2 CDS sequences classified as lineage A. The branch's colors represent the most probable location of their descendent nodes as indicated at the legend. Branch support are indicated only at key nodes [posterior (PP) and posterior state probability (PSP)]. The lineage A.2 is highlighted with a light blue box. All horizontal branch lengths are drawn to a scale of years.

Table 1. Differences observed between sample BR_AM_ILMD_20140001 and the SARS-CoV-2 prototype.

Genome position	8782	9477	12781	14805	25979	26642	28144	28657	28863
NC_045512	C	T	C	C	G	C	T	C	C
BR_AM_ILMD	T	A	T	T	T	T	C	T	T
Codon position	AG**C**	T**T**T	TA**C**	TA**C**	GGA	GCC	T**T**A	GA**C**	T**C**A
	AG**T**	T**A**T	TA**T**	TA**T**	GTA	GCT	T**C**A	GA**T**	T**T**A
Mutation type	s	ns	s	s	ns	s	ns	s	ns
Protein	nsp4	nsp4	nsp9	RdRp	ORF3a ptn	M glycoptn	ORF8 ptn	npptn	npptn
Residue	Ser2839	Phe3071Tyr	Tyr4172	Tyr4847	Gly196Val	Ala40	Leu84Ser	Asp128	Ser197Leu

Legend: Nucleotides substituted in each codon are represented in bold. s: silent mutation; ns: non-silent mutation; nsp: non-structural protein; RdRp: RNA-dependent RNA-polymerase; ORF3a ptn: ORF3a protein; M glycoptn: M glycoprotein; ORF8 ptn: ORF8 protein; npptn: nucleocapsid phosphoprotein.