NATHÁLIA DOS SANTOS RAMOS

EVOLUTIONARY HISTORY OF SARS-COV-2 IN SOUTH AMERICA

Dissertation presented to the Universidade Federal de Viçosa as part of the requirements of the Graduate Program in Applied Biochemistry to obtain the degree of *Magister Scientiae*.

Advisor: Francisco Murilo Zerbini

VIÇOSA – MINAS GERAIS

2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
R175e
2022

Ramos, Nathália dos Santos, 1996-
Evolutionary history of SARS-COV-2 in South America / Nathália dos Santos Ramos. – Viçosa, MG, 2022.
1 dissertação eletrônica (26 f.): il. (algumas color.).

Orientador: Francisco Murilo Zerbini Júnior.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Bioquímica e Biologia Molecular, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.287
Modo de acesso: World Wide Web.

1. COVID-19 (Doença) - Evolução - América do Sul. 2. Mutação (Biologia). I. Júnior, Francisco Murilo Zerbini, 1966-. II. Universidade Federal de Viçosa. Departamento de Bioquímica e Biologia Molecular. Programa de Pós-Graduação em Bioquímica Aplicada. III. Título.

CDD 22. ed. 612.2414

Bibliotecário(a) responsável: Alice Regina Pinto CRB6 2523

NATHÁLIA DOS SANTOS RAMOS

EVOLUTIONARY HISTORY OF SARS-COV-2 IN SOUTH AMERICA

Dissertation presented to the Universidade Federal de Viçosa, as part of the requirements of the Graduate Program in Applied Biochemistry to obtain the degree of *Magister Scientiae*.

APPROVED: February 17st, 2022

Assent:

Nathália dos Santos Ramos
Author

Francisco Murilo Zerbini
Advisor

ACKNOWLEDGEMENTS

I would like to initially thank my family who always supported me and held my hand so that I didn't give up in difficult times, to remember my moments of victory and believe in me when I didn't even believe anymore. So, I dedicate this work to the loves ofmy life.

I would also like to thank the Applied Biochemistry Program and the People Improvement Coordination (Capes) for the opportunity. To Professor Francisco Murilo Zerbini Junior for the opportunity and trust, the UFV Biomolecules Nucleus (Nubiomol),especially to coach Pedro. To my labmates, especially Anelise Orílio who welcomed meon arrival and taught me so much, to Ruither who participated in some stages of the project and João Paulo who helped me with numerous issues throughout the process.

To all, thank you very much!

“Words are windows, or they’re walls.
They sentence us or set us free.
When I speak and when I hear,
Let the love light shine through me”.
(Ruth Bebermeyer)

ABSTRACT

RAMOS, Nathália dos Santos, M.Sc., Universidade Federal de Viçosa, February, 2022. **Evolutionary history of SARS-CoV-2 in South America**. Advisor: Francisco Murilo Zerbini.

This is not the first time that humankind has been targeted by a coronavirus. The history of repeated introductions of animal viruses into human populations, resulting in disease outbreaks, suggests that future similar pandemics are inevitable. Therefore, understanding the possible molecular origin and ongoing evolution of SARS-CoV-2 willprovide critical information for preventing future outbreaks. Coronaviruses have a propensity for genetic recombination across host species boundaries. Consequently, the SARS-CoV-2 genome harbors signatures of multiple recombination events, likely spanning multiple species and wide geographic regions. Other regions of the SARS-CoV-2 genome show the impact of purifying and diversifying selection. The spike (S) proteinof SARS-CoV-2, which allows the virus to enter host cells, exhibits both purifying and diversifying selection signatures, leading to a more effective S protein in infecting humancells and many other mammals and explaining the rapid emergence of new variants. Theglobal spread and explosive growth of the SARS-CoV-2 population within human hostshas contributed to an increase in mutational variability, increasing opportunities for futurerecombination. Differently from what was reported in other studies in which P.1 emerged in Manaus, Amazonas, Brazil between November and December 2020, the present work detects its circulation since August 2020, in São Paulo. Thus, the importance of tracking and monitoring variants is evident, even in the case of low prospecting strains, as their early detection and monitoring of their evolutionary history can enable the prevention of more aggravating conditions or even future new epidemics.

Keywords: SARS-CoV-2. Evolutionary. Variants.

RESUMO

RAMOS, Nathália dos Santos, M.Sc., Universidade Federal de Viçosa, fevereiro de 2022. **História evolutiva do SARS-CoV-2 na América do Sul**. Orientador: Francisco Murilo Zerbini.

Esta não é a primeira vez que a humanidade é alvo de um coronavírus. O histórico de repetidas introduções de vírus animais em populações humanas, resultando em surtos de doenças, sugere que futuras pandemias semelhantes são inevitáveis. Portanto, entender a possível origem e a evolução molecular do SARS-CoV-2 fornecerá informações importantes para prevenir futuros surtos. Os coronavírus têm uma propensão à recombinação genética envolvendo isolados de diferentes espécies hospedeiras. Consequentemente, o genoma do SARS-CoV-2 contém assinaturas de vários eventos de recombinação, provavelmente abrangendo várias espécies e amplas regiões geográficas. Outras regiões do genoma do SARS-CoV-2 evidenciam o impacto de seleção purificadora e diversificadora. A proteína spike (S) do SARS-CoV-2, que permite que o vírus entre nas células hospedeiras, exibe assinaturas de seleção purificadora e diversificadora, levando a uma proteína S mais eficaz na infecção de células humanas e muitos outros mamíferos e explicando o rápido surgimento de novas variantes. A disseminação global e o crescimento explosivo da população de SARS-CoV-2 em hospedeiros humanos contribuíram para um aumento na variabilidade mutacional, aumentando as oportunidades de recombinação futura. Diferentemente do relatado em outros estudos em que o P.1 surgiu em Manaus, Amazonas, Brasil entre novembro e dezembro de 2020, o presente trabalho detecta sua circulação desde agosto de 2020, em São Paulo. Assim, fica evidente a importância do rastreamento e monitoramento de variantes, mesmo no caso daquelas de baixa prevalência, pois sua detecção precoce e o acompanhamento de sua história evolutiva podem possibilitar a prevenção de condições mais agravantes ou mesmo futuras novas epidemias.

Palavras-chave: SARS-CoV-2. Evolução. Variantes.

CONTENTS

INTRODUCTION

In December 2019, a pneumonia outbreak of unknown etiology was reported fromthe city of Wuhan (Hubei Province, China) (Zhou *et al.*, 2020). Soon afterwards it was shown to be caused by a new coronavirus (Wu *et al.*, 2020), which was named "severe acute respiratory syndrome coronavirus 2" (SARS-CoV-2) by the International Virus Taxonomy Committee (ICTV) in February 2020 (Gorbalenya *et al.*, 2020). The disease was named "coronavirus disease 2019" or COVID-19 by the World Health Organization(World Health Organization, 2020a). Later, with the rapid and constant emergence of variants, the WHO started to use letters from the Greek alphabet to identify those considered "variants of concern", of which five have so far been named: alpha, beta, gamma, delta and omicron (Konings *et al.*, 2021; Centers for Disease Control and Prevention, 2021).

After its initial expansion in China, the first cases began to emerge in Western Asia, North America and Europe, seriously impacting these regions. Brazil registered thefirst case in Latin America at the end of February 2020, and in less than a month there were over 7,000 confirmed cases in Latin America and the Caribbean (Miller *et al.*, 2020).The disease was declared a pandemic by the WHO on March 11, 2020 (World Health Organization, 2020b).

A lineage is a group of closely related viruses with a common ancestor, therefore formed by a set of variants. In addition, a variant is a viral genome that may contain one or more mutations. In certain cases, a group of variants with similar genetic changes, such as a lineage or group of lineages, may be designated by public health organizations as a Variant Being Monitored (VBM), Variant of Concern (VOC) or a Variant of Interest (VOI) due to shared attributes and characteristics that may require public health action, as is the case of SARS-CoV-2 (Centers for Disease Control and Prevention, 2021).

This is not the first time that humanity was the target of a coronavirus, as these have already been the causative agents of two serious epidemics in the last two decades. The first occurred in 2002-2003, in Guangdong Province, China, caused by severe acute respiratory syndrome coronavirus (SARS-CoV) (Drosten *et al.*, 2003; Ksiazek *et al.*, 2003; Peiris *et al.*, 2003). The second started in 2012 in the Middle East, caused by Middle East respiratory syndrome coronavirus (MERS-CoV) (Zaki *et al.*, 2012). SARS-CoV-2 has a nucleotide sequence identity of 80% and 50%, respectively, with SARS-CoV and MERS-CoV, and like these two coronaviruses, is also considered to be a case of zoonotic spillover (Zhou *et al.*, 2020; Andersen *et al.*, 2020). Although the natural host of SARS-CoV-2 has not yet been identified, credible candidates include bats and pangolins (Wacharapluesadee *et al.*, 2021; Andersen *et al.*,

2020; Zhou *et al.*, 2020). In addition, other animals could have acted as intermediate hosts, including domestic animals. By analyzing the recurrent outbreaks of SARS, MERS and now SARS-CoV-2 it is possible to visualize the remarkable ability of coronaviruses to cross species barriers, enabling transmission in humans (Menachery *et al.*, 2017).

SARS-CoV-2 is a member of the species *Severe acute respiratory syndrome- related coronavirus*, classified in the subgenus *Sarbecovirus*, genus *Betacoronavirus*, family *Coronaviridae* and order *Nidovirales* (Gorbalenya *et al.*, 2020). SARS-CoV-2 is an enveloped virus, with an approximately 30,000 bp-long, positive sense, single- stranded RNA genome. Two-thirds of the genome are comprised by ORF 1a/1b, whose product processing generates 16 non-structural proteins (nsp1-16), including the RNA- directed RNA polymerase (RdRP). The other one-third of the genome, next to the 3'- end, contains nine ORFs which encode structural and accessory proteins, including spike (ORF S), envelope (ORF E), membrane or matrix (ORF M) and nucleocapsid (ORF N) (Cui et al., 2019).

Populations of RNA viruses evolve at high rates, mostly due to their high recombination and mutations rates (Duffy *et al.*, 2008; Dolan *et al.*, 2018). The latter is a consequence, mainly, of the lack of proofreading activity of the RdRP (Drake, 1993), and is magnified in those RNA viruses with larger genomes (Duffy *et al.*, 2008; Sanjuan, 2012). Coronaviruses such as SARS-CoV-2 share many of these features: they have the largest known genomes among RNA viruses and have high rates of recombination and of synonymous and non-synonymous mutations (Gorbalenya *et al.*, 2006; Mercatelli and Giorgi, 2020; Chan *et al.*, 2020; Chen *et al.*, 2020; Koyama *et al.*, 2020; Wang *et al.*, 2020; Yin, 2020). However, due to the presence of an independent proofreading activity (Ma *et al.*, 2015), mutations occur at lower rates compared to other RNA viruses (Zhao *et al.*, 2004; Gorbalenya *et al.*, 2006). Thus, inasmuch as these three aspects (genome size, recombination rate and mutation rate) contribute to the continuous emergence of a wide range of closely related variants derived from ancestral lineages, recombination may play a fundamental role, enabling viral adaptability and conferring a wide level of diversity to coronavirus populations (Domingo and Perales, 2019; Domingo *et al.*, 2012; Simmonds *et al.*, 2019).

The present study aimed to comprehend the diversity and evolution of SARS- CoV-2 populations in South America during the COVID-19 pandemic, from December 2019 to July 2021, and the implications on transmission dynamics during this period.

MATERIALS AND METHODS

Construction and curation of the data sets

A total of 35,319 complete genomes were downloaded from GISAID (gisaid.org) on July 8th, 2021, and submitted to a screening process that can be divided into six steps:

(1) A parallel processing was carried out to create lists of 500 sequences to check the genome size, discard those sequences with >1% ambiguity (i.e., only sequences with >99% of the nucleotides identified as A, C, G or T were used) and identify the ORFs. These processes were carried out based on the Wuhan reference sequence (GenBank accession number MN908947);

(2) Subsequent analyses were performed based on the sequences of ORFs 1a/1b, S, E and N. For each ORF a BLASTp analysis was carried out and sequences with <90% identity each were discarded;

(3) After the identification of the four regions of interest, these were submitted to a trimming process to remove the other regions and the sequences of each isolate were concatenated;

(4) Then, CD-HIT v. 4.8.1 was used with a 100% identity threshold to remove redundant (identical) sequences, resulting in 19,749 sequences;

(5) Subsequently, sequences that had more than two consecutive "N" were removed. Thus, after curating the initial data set, 15,076 better quality sequences with the four concatenated ORFs was obtained.

(6) The Pangolin software (https://pangolin.cog-uk.io/) was used to classify the 15,076 sequences of the second data set into variants, requiring them to be subdivided into 18 groups before web submission. The results obtained came out in Excel format. The information was thus organized using MS Excel, with the number of variants circulating in South America until July 8, 2021 plotted as charts for better visualization and understanding.

A third data set of 464 sequences, was obtained from the data set of 15,076 sequences by applying a 99.935% threshold in CD-HIT.

After analyzing the data set collected in the previous six steps, it was possible to observe five main variants circulating in South America, making it possible to constructing a fourth data set with complete sequences of the five variants (P.1, P.2, B.1, P.1.14 and B.1.1.7) plus four additional variants which are considered variants of concern (B.1.617.2, delta and B.1.351, beta) or variants of interest (B.1.1.28 and B.1.1.33). Also, there is robust evidence that P.1 was originated from B.1.1.28 (Naveca *et al.*, 2021).

Thus, complete sequences of the nine variants were taken from the complete dataset (35,319 sequences), classified using Pangolin to make sure they were one of the nine variants,

and used to collect data on viral incidence and population expansion, mapping their presence and proportion over time throughout the continent. Then, all redundant sequences were eliminated using CD-HIT (100% identity threshold). Next, CD-HIT was used again with a 99,95% identity threshold, separately, in each of the nine variants groups. Subsequently, the nine variant groups were merged in aligned in MAFFT, resulting in a total of 764 complete genomic sequences.

Diversity and evolutionary analyses

In order to verify the consistency of the second data set obtained after concatenating the ORFs (15,076 sequences), a phylogenetic tree was built to compare its topology with the one available in GISAID. For this, a multiple sequence alignment was performed using MAFFT v. 7.471, applying ModeltestNG v. 0.1.7 to verify which would be the best fit nucleotide substitution model, confirmed to be the GTR model. A phylogenetic tree of the 15,076 sequences was constructed using Fasttree v. 2.1.9, with 10,000 bootstrap repeats. Tajima's neutrality test, implemented in MEGA v. 5.2.2, was performed with the same data set.

The SLAC method (single-likelihood ancestor counting) implemented in the Datamonkey server (https://www.datamonkey.org/analyses) wasused to infer non-synonymous (dN) and synonymous (dS) substitution rates on a per-site basis on the third data set of 464 sequences. The fourth data set composed of the 764 complete sequences of the nine variants was analyzed using the software DnaSP v. 5.10 to calculate polymorphic regions along the complete genome of the nine variants.

RESULTS AND DISCUSSION

Number and relative proportion of circulating SARS-CoV-2 variants in SouthAmerica from January 2020 to July 2021

A total of 184 SARS-CoV-2 variants circulated in South America from January 2020 until July 8, 2021. During this period, the five most abundant were B.1, P.2 (zeta), B.1.1.7 (alpha), P.1 (gamma) and P.1.14 (Figure 1).

The B.1 variant was the only one present in South America during the first two months of the pandemic (Jan/Feb 2020) and was the most prevalent variant throughout 2020 (Figure 1). The P.2 (zeta) variant emerged in April 2020 and stayed at low prevalence until September, when it started to increase substantially, becoming the most common variant in January 2021. In December 2020, B.1.1.7 (alpha) and P.1 (gamma) started to appear in considerable proportions. The P.1 (gamma) variant quickly gained space and became the most prevalent variant in February 2021, maintaining this status until the end of the surveillance period (July 2021). P.1.14 appeared in considerable proportion in January 2021, but its prevalence did not increase significantly during the surveillance period (Figure 1). By May 2021, B.1 and P.2 were present at extremely low proportions, with P.1 (gamma), B.1.1.7 (alpha) and P.1.14 representing >75% of the circulating viral population (Figure 1).

Another point that draws attention is the increase in the number of variants circulating at high proportion between December 2020 and January 2021 (Figure 2), with the end of the year festivities as a possible explanation. With the increase in people's physical contact due to the celebrations, an increase in transmission of many variants mayhave occurred proportionally, resulting in an increase in the number of circulating variants.

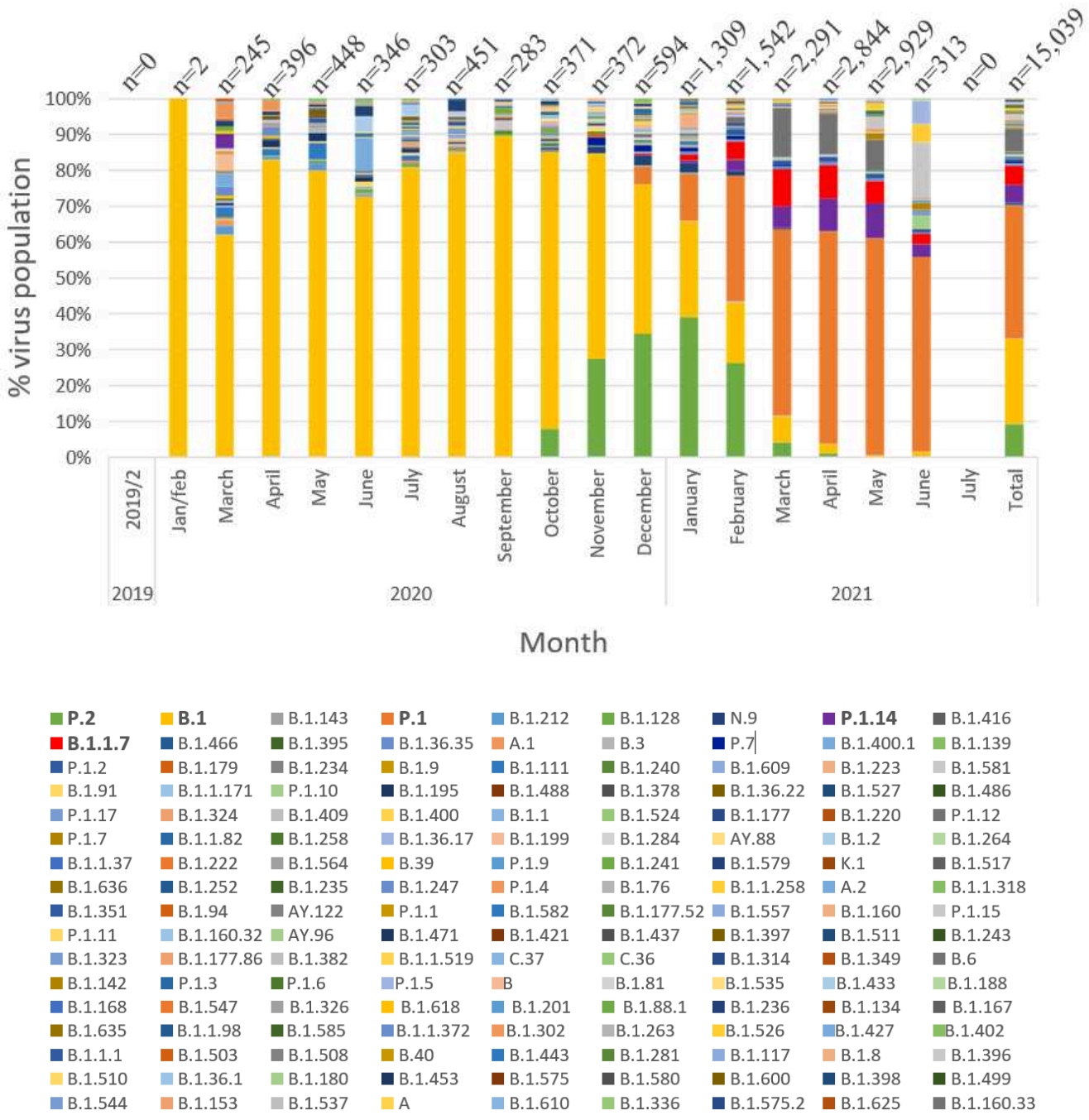

Figure 1. Relative proportion of circulating SAR-CoV-2 variants in South America from January 2020 until July 2021. Variants were classified and named according to Pangolin (https://pangolin.cog-uk.io/). The five major variants (B.1, P.2, P.1, P.1.14 and B.1.1.7), occurring in greater proportion during the surveillance period, are highlighted in bold. "n" is the number of sequences sampled in each month. Out of the 15,076 sequences in the data set, 37 sequences were not used to plot the graph because they did not have a complete (day/month/year) collection date.

No correlation between the number of circulating SARS-CoV-2 variants and COVID-19 mortality

Data on confirmed deaths was taken from the Our World in Data website, with the numbers of deaths per million people in Latin America being collected at three points of each month: beginning, middle and end. The three points were averaged, and the values plotted together with the number of circulating variants (Figure 2). A relationship between viral population expansion and number of confirmed deaths was tested. The value of the Pearson correlation coefficient obtained was low (r=0.401), but suggestive of a moderately positive association. However, the *p*-value was high (0.099), indicating a statistically non-significant correlation. Thus, it is concluded that there is no relationship between the number of circulating variants and COVID-19 mortality.

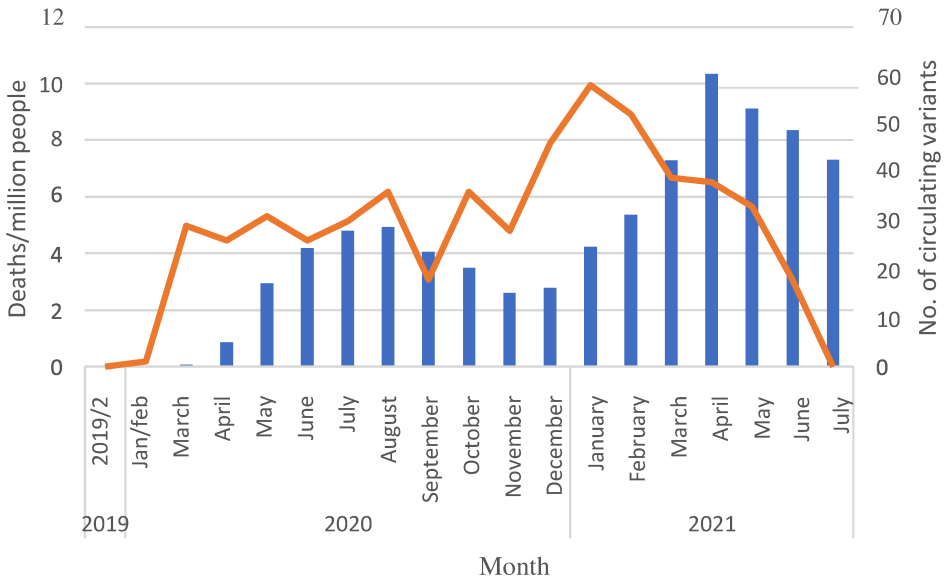

Figure 2. Number of circulating SARS-CoV-2 variants in South America (orange line) plotted against deaths per million people (blue bars) from January 2020 until July 2021.

One possible explanation for the fluctuation in the number of variants over the months would be genetic drift. The first significant reduction in the number of circulating variants occurred in September 2020 (Figure 2), which could have been the result of a bottleneck effect followed by a founder effect. This hypothesis is supported by the observation that P.2 had been circulating in Latin America since April 2020 in low proportion, but after the reduction in September, it started to gain space, becoming moresignificant after October 2020 (Figure 1). Moreover, a second reduction in the number ofcirculating variants was observed in November

2020 (Figure 2), this time followed by theemergence and gradual increase of P.1 (Figure 1). Although we are unable to speculate on the nature of the bottleneck in September 2020, there is a possible explanation for the second bottleneck in November. The emergence of P.1 was proposed to have occurred at the end of November 2020 in Amazonas, Brazil (Naveca *et al.*, 2021). Previously, Manaus underwent a strict lockdown (facilitated by its geographical isolation), which could have caused a bottleneck effect. The rapid increase in the relative proportion of P.1 following its emergence suggests the role of selection, which is supported by its greater transmissibility. Together with the arrival of the end-of-year festivities, this would explain its rapid dissemination throughout Brazil during the first semester of 2021 (Figure 1).

Phylogeny of circulating SARS-CoV-2 variants in South America

The phylogenetic trees based on both ours and the GISAID data sets are strongly congruent. In the tree based on the GISAID data set, two clades corresponding to sequences from Latin America can be observed (Figure 3A). Clade 2 (magnified in Figure3B) includes four of the five main circulating variants in South America (P1, P2, P.1.14and B.1.1.7). In the tree based on our own data set (Figure 4), the South American sequences are also divided into two main clades corresponding to clades A and B of the GISAID tree. It is worth mentioning that in the GISAID tree (Figure 3) there are sequences from countries which are not included in our data set. Moreover, in the tree based on our data set (Figure 4) the B.1 sequences are located at the most basal point of the tree, which will later give rise to other variants. Thus, it is possible to verify the evolution of the B.1 branch, giving rise to the other four major variants in South America.

To verify the demographic dynamics of the viral populations, the nucleotide diversity (π) and Tajima's D test were calculated for the data set of 15,076 sequences. Theobtained values (0.00032 and -2.6502693, respectively) indicate a low degree of geneticvariability and strong evidence of a recent population expansion following a genetic bottleneck, supporting the hypothesis raised above.

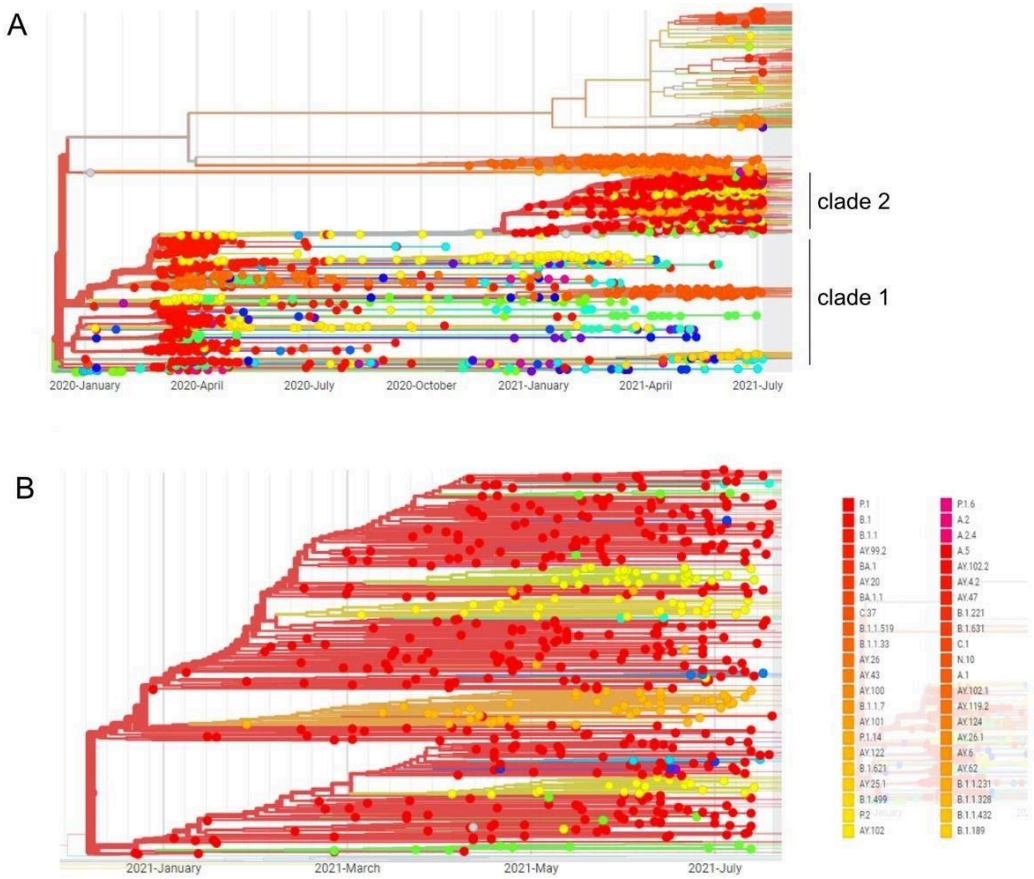

Figure 3. Phylogenetic tree of complete SARS-CoV-2 genomes available in the GISAID database. **A.** Sequences from South America form two distinct clades (1 and 2). **B.** Closerview of clade 2, which includes sequences four of the five major circulating variants in South America, including P.1, which emerged in late November of 2020.

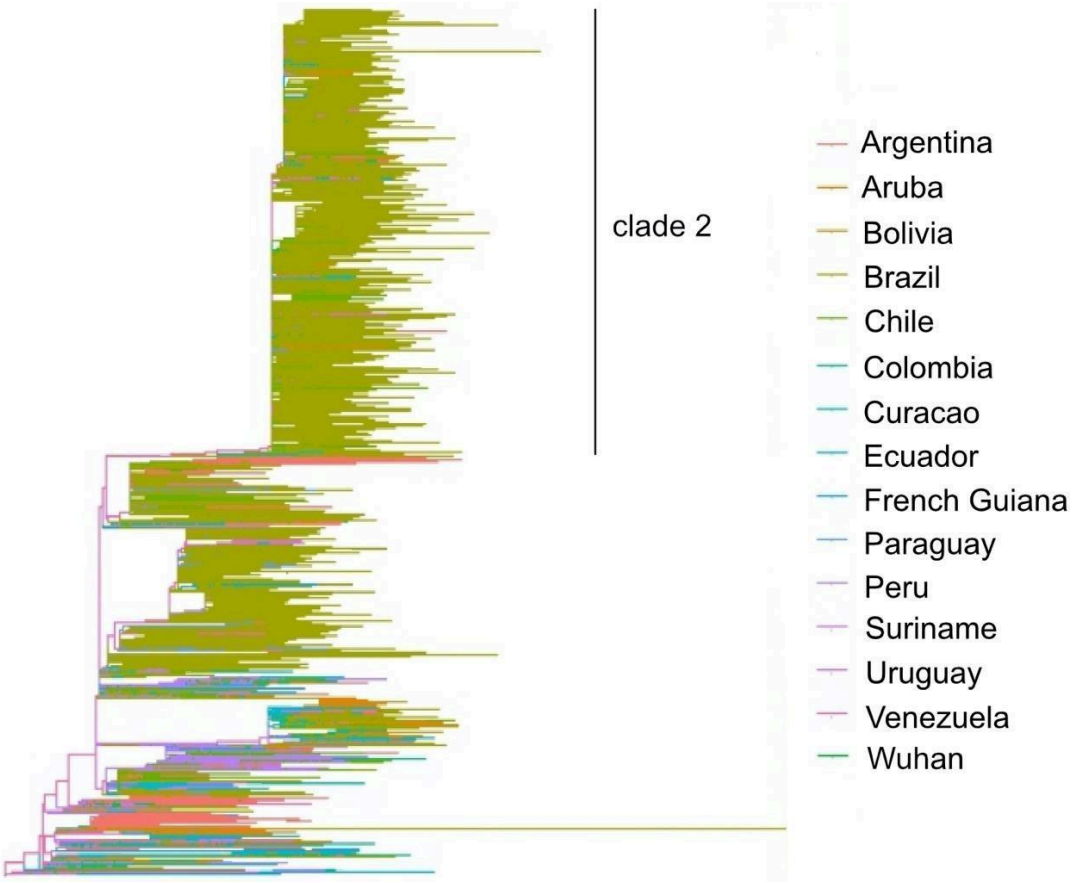

Figure 4. Phylogenetic tree of complete SARS-CoV-2 genomes based on the data set assembled in this work. The clade that corresponds to clade 2 in the GISAID tree is indicated.

Negative selection predominates in the South American SARS-CoV-2 variants

To assess the occurrence of selection in the South American SARS-CoV-2 population, the SLAC method available in the Datamonkey server was used. Amino acid sites were considered to be under positive selection when dN/dS ratios were >1 with $p \leq 0.1$. The results are presented in Table 1. In ORF1a/1b, which encodes the viral RdRp, 12 amino acid sites were found to be under positive (diversifying) selection, and 197 sites were under negative (purifying) selection. In ORF S, which codes for the spike protein, eight amino acid sites were found to be under positive selection and 33 sites were under negative selection. In ORF E, which encodes the envelope protein, no amino acid sites were found to be under either positive or negative selection. In ORF N, which encodes the nucleocapsid protein, five amino acid sites were under positive selection and 19 under negative selection. Thus, for three of the four analyzed ORFs, a number of amino acid sites were positive selection but there is a higher

number of sites under negative selection.

Dates and locations of first detection of the nine main circulating SARS-CoV-2variants in South America

With the database containing the complete sequences of the nine variants obtained until July 2021, it was possible to determine the dates and locations where the nine main circulating variants were sequenced for the first time in South America. Five variants were first sequenced in São Paulo state, Brazil: B.1 on 2/25/2020, B.1.1.28 on 3/5/2020, P.1.14 on 11/3/2020, B.1.1.7 on 12/21/2020 and, surprisingly, P.1 on 8/24/2020. Two variants were first sequenced in Rio de Janeiro state, Brazil: P.2 on 4/13/2020 and B.1.1.33 on 3/1/2020. Two variants were first sequenced in the Caribbean island of Aruba: B.1.617.2 on 4/16/2021 and B.1.351 on 2/25/2021.

Table 1. Amino acid sites under positive (diversifying) and negative (purifying) selection in ORFs 1a/1b, S, E and N of SARS-CoV-2 sequences from South America from January 2020 until July 2021.

ORF	**Amino acid sites under positive selection**	**Amino acid sites under negative selection**
1a/1b	360, 1921, 3201, 3255, 3353, 3468, 3606, 3718, 3750, 3930, 6831, 6957	11, 148, 156, 160, 188, 189, 190, 196, 202, 334, 341, 402, 416, 443, 491, 548, 549, 615, 717, 748, 786, 810, 811, 815, 828, 954, 1157, 1167, 1174, 1184, 1273, 1323, 1345, 1426, 1439, 1554, 1565, 1652, 1673, 1744, 1749, 1903, 1907, 1922, 1925, 1935, 1960, 1980, 1985, 1995, 2007, 2015, 2018, 2067, 2091, 2108, 2116, 2155, 2272, 2385, 2433, 2445, 2489, 2557, 2584, 2625, 2638, 2678, 2682, 2780, 2839, 2851, 2884, 2894, 2895, 2975, 3055, 3089, 3153, 3180, 3238, 3271, 3291, 3297, 3329, 3368, 3397, 3414, 3459, 3460, 3479, 3492, 3511, 3535, 3563, 3568, 3581, 3603, 3744, 3769, 3785, 3796, 3839, 3853, 3867, 3899, 3906, 3936, 3937, 3945, 4050, 4058, 4068, 4168, 4171, 4172, 4196, 4205, 4235, 4361, 4423, 4466, 4531, 4637, 4639, 4665, 4768, 4803, 4807, 4819, 4846, 4886, 4983, 4990, 4991, 5004, 5019, 5091, 5092, 5127, 5151, 5157, 5181, 5323, 5342, 5347, 5357, 5387, 5436, 5540, 5579, 5635, 5679, 5745, 5761, 5803, 5818, 5822, 5833, 5838, 5860, 5954, 5995, 6027, 6073, 6096, 6112, 6159, 6163, 6187, 6207, 6238, 6250, 6275, 6299, 6301, 6306, 6321, 6334, 6369, 6415, 6429, 6445, 6524, 6535, 6572, 6627, 6637, 6665, 6723, 6728, 6804, 6852, 6888, 6940, 6952, 6996

S	5, 95, 142, 452, 484, 501, 681, 1176	28, 32, 43, 53, 55, 58, 68, 77, 99, 186, 200, 234, 391, 432, 489, 541, 562, 606, 682, 692, 723, 789, 824, 840, 856, 940, 960, 1018, 1101, 1119, 1178, 1215, 1238
E	no sites under positive selection	no sites under negative selection
N	13, 67, 204, 292, 377	16, 53, 110, 111, 126, 128, 157, 172, 192, 232, 248, 274, 302, 315, 318, 327, 333, 341, 360

Linked to the fact that São Paulo is the financial center of Brazil and the city of São Paulo one of the most populous in the world, it is reasonable to assume that the region has a higher detection rate of viral lineages, both due to probable introductions from othercontinents and due to carrying out a greater number of tests. Regardless, the point that draws the most attention is the detection of P.1. It has been widely assumed that P.1 emerged in the state of Amazonas between late November and early December of 2020 (Naveca *et al.*, 2021). Its detection in August of 2020 raises two possibilities. The first is that P.1 did not emerge in the Amazon but was actually introduced there after emerging in São Paulo. The second would be that P.1 emerged elsewhere and circulated undetected in Brazil until becoming prevalent in Manaus due to a founder effect following a bottleneck (as hypothesized above). Underreporting of cases and low testing rates could have contributed for its undetected circulation.

The first detection of P.2 and B.1.1.33 occurred in Rio de Janeiro state, Brazil. Variant P.2 was detected in considerable proportions in South America during the secondsemester of 2020 but was eventually replaced by P.1. Variants B.1.351 (beta) and B.1.617.2 (delta) were detected initially in Aruba. These two variants ended up not gaining much space in South America until July of 2021, but were chosen to be monitored in the present study since they are considered variants of concern.

Genetic variability analyses of the SARS-CoV-2 population from South America

Genetic variability indices were calculated for the data set of 764 complete sequences. This data set had 754 haplotypes, with a haplotype diversity of 0.987. The nucleotide diversity for the whole genome was 0.00117, indicating a low degree of genetic variability. The combination of a high value for haplotype diversity and a low value for nucleotide diversity is consistent with the hypothesis of a genetic bottleneck followed by recent population expansion.

Despite the overall low degree of genetic variability, this variability is not evenly distributed along the SARS-CoV-2 genome. Nucleotide diversity values calculated along the

full-length genome (Figure 5) indicate that the S ORF (encoding the spike protein, corresponding approximately to nucleotides 20,000 to 26,000) has a higher variability than the rest of the genome. This is actually expected, since it is the protein responsible for interacting with host ACE2 receptors to enter the cell.

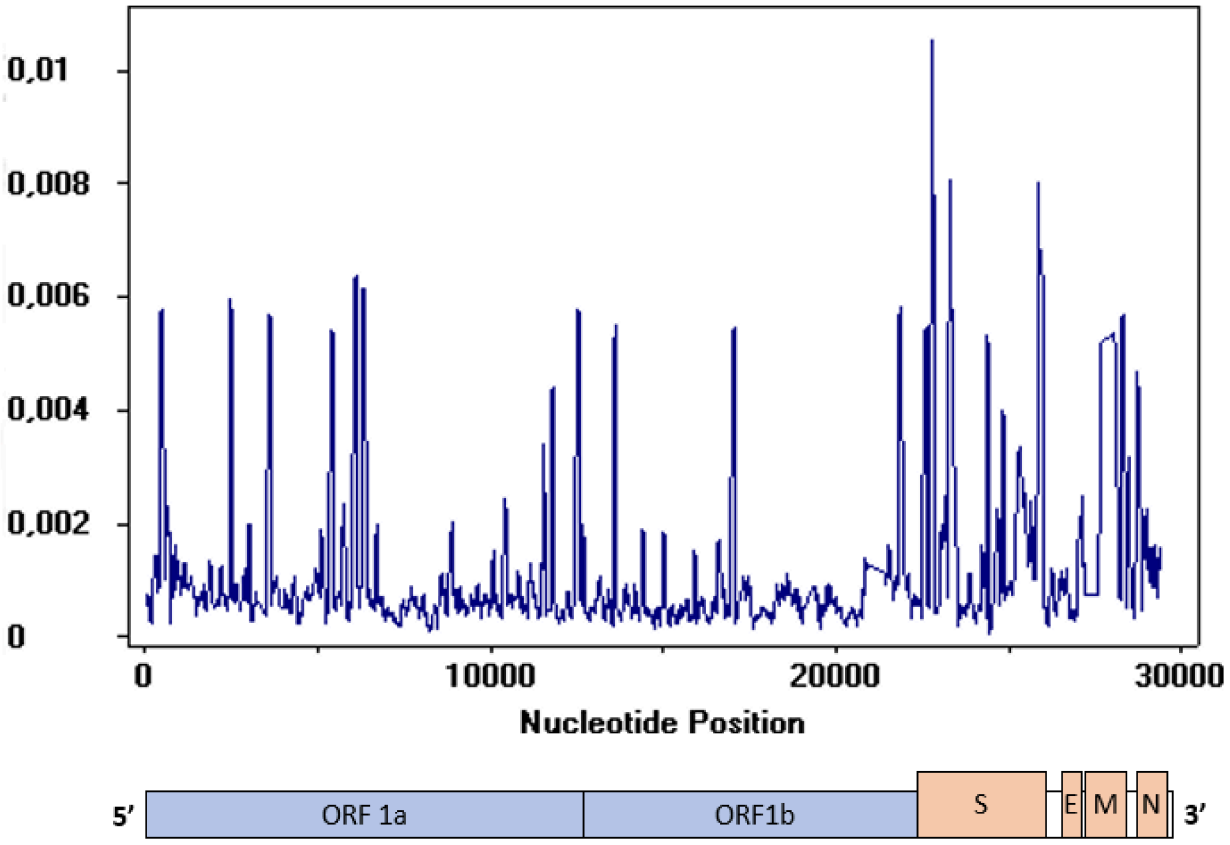

Figure 5. Nucleotide diversity (□) calculated along SARS-CoV-2 sequences from South America obtained from January 2020 until July 2021.

CONCLUSIONS

A total of 184 SARS-CoV-2 variants circulated in South America from January2020 until July 2021. Variants B.1, P.2 (zeta), B.1.1.7 (alpha), P.1 (gamma) and P.1.14 were the most prevalent during this period.

Two possible bottleneck events occurred in September and November of 2020, followed by the quick dissemination of P.2 and P.1, respectively.

The four ORFs analyzed in this work (1a/1b, S, N and E) have a small number of amino acid sites under positive selection (especially ORF S), and a larger number of sites under negative selection.

A low degree of genetic variability was observed for the SARS-CoV-2 population from South America, with ORF S displaying a higher degree of variability compared to other regions of the viral genome.

A sequence corresponding to the P.1 variant was obtained in August of 2020 in São Paulo state, three months before its supposed emergence in Manaus. This initial detection probably went unnoticed as P.1 was not a predominant variant at the time, highlighting the need for increased sequence surveillance and greater monitoring of rare variants.

REFERENCES

Andersen K. G., Rambaut A., Lipkin W. I., Holmes E. C., Garry R. F. (2020) Theproximal origin of SARS-CoV-2. *Nature Medicine,* **26**, 450-452.

Centers for Disease Control and Prevention (2021) *SARS-CoV-2 variant classifications and definitions* [Online]. Available at: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html. Accessed on Feb 9, 2022.

Chan J. F.-W., Kok K.-H., Zhu Z., Chu H., To K. K.-W., Yuan S., Yuen K.-Y. (2020) Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *EmergingMicrobes & Infections,* **9**, 221-236.

Chen Z.-w., Li Z., Li H., Ren H., Hu P. (2020) Global genetic diversity patterns andtransmissions of SARS-CoV-2. *medRxiv*, 2020.2005.2005.20091413.

Cui J., Li F., Shi Z.-L. (2019) Origin and evolution of pathogenic coronaviruses. *NatureReviews Microbiology,* **17**, 181-192.

Dolan P. T., Whitfield Z. J., Andino R. (2018) Mechanisms and concepts in RNA viruspopulation dynamics and evolution. *Annual Review of Virology,* **5**, 69-92.

Domingo E., Perales C. (2019) Viral quasispecies. *PLoS Genetics,* **15**, e1008271.

Domingo E., Sheldon J., Perales C. (2012) Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews,* **76**, 159-216.

Drake J. W. (1993) Rates of spontaneous mutation among RNA viruses. *Proceedings ofthe National Academy of Sciences, USA,* **90**, 4171.

Drosten C., Günther S., Preiser W., van der Werf S., Brodt H.-R., Becker S., Rabenau H., Panning M., Kolesnikova L., Fouchier R. A. M., Berger A., Burguière A.- M., Cinatl J., Eickmann M., Escriou N., Grywna K., Kramme S., Manuguerra J.-C., Müller S., Rickerts V., Stürmer M., Vieth S., Klenk H.-D., Osterhaus A. D., Schmitz H., Doerr H. W. (2003) Identification of a novel coronavirus inpatients with severe acute respiratory syndrome. *New England Journal of Medicine,* **348**, 1967-1976.

Duffy S., Shackelton L. A., Holmes E. C. (2008) Rates of evolutionary change inviruses: Patterns and determinants. *Nature Reviews Genetics,* **9**, 267-276.

Gorbalenya A. E., Baker S. C., Baric R. S., de Groot R. J., Drosten C., Gulyaeva A. A.,Haagmans B. L., Lauber C., Leontovich A. M., Neuman B. W., Penzar D., Perlman S., Poon L. L. M., Samborskiy D. V., Sidorov I. A., Sola I., Ziebuhr J.,*Coronaviridae* Study Group of the International Committee on Taxonomy of V.(2020) The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature*

Microbiology, **5**, 536-544.

Gorbalenya A. E., Enjuanes L., Ziebuhr J., Snijder E. J. (2006) *Nidovirales*: evolving the largest RNA virus genome. *Virus Research,* **117**, 17-37.

Konings F., Perkins M. D., Kuhn J. H., Pallen M. J., Alm E. J., Archer B. N., Barakat A., Bedford T., Bhiman J. N., Caly L., Carter L. L., Cullinane A., de Oliveira T.,Druce J., El Masry I., Evans R., Gao G. F., Gorbalenya A. E., Hamblion E., Herring B. L., Hodcroft E., Holmes E. C., Kakkar M., Khare S., Koopmans M., Korber B., Leite J., MacCannell D., Marklewitz M., Maurer-Stroh S., Rico J. A. M., Munster V. J., Neher R., Munnink B. O., Pavlin B. I., Peiris M.,Poon L., Pybus O., Rambaut A., Resende P., Subissi L., Thiel V., Tong S., vander Werf S., von Gottberg A., Ziebuhr J., Van Kerkhove M. D. (2021) SARS- CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nature Microbiology,* **6**, 821-823.

Koyama T., Platt D., Parida L. (2020) Variant analysis of SARS-CoV-2 genomes. *Bulletin of the World Health Organization,* **98**, 495-504.

Ksiazek T. G., Erdman D., Goldsmith C. S., Zaki S. R., Peret T., Emery S., Tong S.,Urbani C., Comer J. A., Lim W., Rollin P. E., Dowell S. F., Ling A.-E., Humphrey C. D., Shieh W.-J., Guarner J., Paddock C. D., Rota P., Fields B.,DeRisi J., Yang J.-Y., Cox N., Hughes J. M., LeDuc J. W., Bellini W. J., Anderson L. J. (2003) A novel coronavirus associated with severe acute respiratory syndrome. *New England Journal of Medicine,* **348**, 1953-1966.

Ma Y., Wu L., Shaw N., Gao Y., Wang J., Sun Y., Lou Z., Yan L., Zhang R., Rao Z. (2015) Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *Proceedings of the National Academy of Sciences, USA,* **112**, 9436.

Menachery V. D., Graham R. L., Baric R. S. (2017) Jumping species - a mechanism forcoronavirus persistence and survival. *Current Opinion in Virology,* **23**, 1-7.

Mercatelli D., Giorgi F. M. (2020) Geographic and genomic distribution of SARS-CoV-2 mutations. *Frontiers in Microbiology,* **11**.

Miller M. J., Loaiza J. R., Takyar A., Gilman R. H. (2020) COVID-19 in Latin America: novel transmission dynamics for a global pandemic? *PLOS Neglected Tropical Diseases,* **14**, e0008265.

Naveca F. G., Nascimento V., de Souza V. C., Corado A. d. L., Nascimento F., Silva G., Costa Á., Duarte D., Pessoa K., Mejía M., Brandão M. J., Jesus M., Gonçalves L., da Costa C. F., Sampaio V., Barros D., Silva M., Mattos T., Pontes G., Abdalla L., Santos J. H., Arantes I., Dezordi F. Z., Siqueira M. M., Wallau G. L., Resende P. C., Delatorre E., Gräf T., Bello G. (2021) COVID-19in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1emergence. *Nature Medicine,* **27**, 1230-1238.

Peiris J. S. M., Lai S. T., Poon L. L. M., Guan Y., Yam L. Y. C., Lim W., Nicholls J., Yee W. K. S., Yan W. W., Cheung M. T., Cheng V. C. C., Chan K. H., Tsang D. N. C., Yung R. W. H., Ng T. K., Yuen K. Y. (2003) Coronavirus as a possiblecause of severe acute respiratory syndrome. *The Lancet,* **361**, 1319-1325.

Sanjuan R. (2012) From molecular genetics to phylodynamics: Evolutionary relevance of mutation rates across viruses. *PLoS Pathogens,* **8**, e1002685.

Simmonds P., Aiewsakun P., Katzourakis A. (2019) Prisoners of war - host adaptationand its constraints on virus evolution. *Nature Reviews Microbiology,* **17**, 321- 328.

Wacharapluesadee S., Tan C. W., Maneeorn P., Duengkae P., Zhu F., Joyjinda Y., Kaewpom T., Chia W. N., Ampoot W., Lim B. L., Worachotsueptrakun K., Chen V. C.-W., Sirichan N., Ruchisrisarod C., Rodpan A., Noradechanon K., Phaichana T., Jantarat N., Thongnumchaima B., Tu C., Crameri G., Stokes M.M., Hemachudha T., Wang L.-F. (2021) Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nature Communications,* **12**, 972.

Wang C., Liu Z., Chen Z., Huang X., Xu M., He T., Zhang Z. (2020) The establishmentof reference sequence for SARS-CoV-2 and variation analysis. *Journal of Medical Virology,* **92**, 667-674.

World Health Organization (2020a) *Naming the coronavirus disease (COVID-19) andthe virus that causes it* [Online]. Available at: https://www.who.int/emergencies/ diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that- causes-it. Accessed on Feb 9, 2022.

World Health Organization (2020b) *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020* [Online]. Available at: https://www.who.int/director-general/speeches/detail/who-director-general-s- opening-remarks-at-the-media-briefing-on-covid-19-11-March-2020. Accessed on Feb 9, 2022.

Wu F., Zhao S., Yu B., Chen Y.-M., Wang W., Song Z.-G., Hu Y., Tao Z.-W., Tian J.-H., Pei Y.-Y., Yuan M.-L., Zhang Y.-L., Dai F.-H., Liu Y., Wang Q.-M., ZhengJ.-J., Xu L., Holmes E. C., Zhang Y.-Z. (2020) A new coronavirus associated with human respiratory disease in China. *Nature,* **579**, 265-269.

Yin C. (2020) Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics,* **112**, 3588-3596.

Zaki A. M., van Boheemen S., Bestebroer T. M., Osterhaus A. D. M. E., Fouchier R. A. M. (2012) Isolation of a novel coronavirus from a man with pneumonia in SaudiArabia. *New England Journal of Medicine,* **367**, 1814-1820.

Zhao Z., Li H., Wu X., Zhong Y., Zhang K., Zhang Y.-P., Boerwinkle E., Fu Y.-X.(2004) Moderate

mutation rate in the SARS coronavirus genome and its implications. *BMC Evolutionary Biology,* **4**, 21.

Zhou P., Yang X.-L., Wang X.-G., Hu B., Zhang L., Zhang W., Si H.-R., Zhu Y., Li B., Huang C.-L., Chen H.-D., Chen J., Luo Y., Guo H., Jiang R.-D., Liu M.-Q., Chen Y., Shen X.-R., Wang X., Zheng X.-S., Zhao K., Chen Q.-J., Deng F., Liu L.-L., Yan B., Zhan F.-X., Wang Y.-Y., Xiao G.-F., Shi Z.-L. (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature,* **579**, 270-273.