medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Mutation hotspots, geographical and temporal distribution of SARS-CoV-2 lineages in Brazil, February 2020 to February 2021: insights and limitations from uneven sequencing efforts

Vinícius Bonetti Franceschi[1], Patrícia Aline Gröhs Ferrareze[2], Ricardo Ariel Zimerman[3], Gabriela Bettella Cybis[4], Claudia Elizabeth Thompson[1,2,5]*

[1] Center of Biotechnology, Graduate Program in Cell and Molecular Biology (PPGBCM), Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

[2] Graduate Program in Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil

[3] Department of Infection Control and Prevention, Hospital da Brigada Militar, Porto Alegre, RS, Brazil

[4] Department of Statistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

[5] Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil

*** Corresponding author**

Address for correspondence:

Claudia Elizabeth Thompson

Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), 245/200C Sarmento Leite St, Porto Alegre, RS, Brazil. ZIP code: 90050-170. Phone: +55 (51) 3303 8889.

E-mail: cthompson@ufcspa.edu.br, thompson.ufcspa@gmail.com

Running title: SARS-CoV-2 diversity and distribution in Brazil

Keywords: COVID-19, Severe acute respiratory syndrome coronavirus 2, Infectious Diseases, Sequencing, Molecular Evolution.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

The COVID-19 pandemic has already reached approximately 110 million people and it is associated with 2.5 million deaths worldwide. Brazil is the third worst-hit country, with approximately 10.2 million cases and 250 thousand deaths. Unprecedented international efforts have been established in order to share information about epidemiology, viral evolution and transmission dynamics. However, sequencing facilities and research investments are very heterogeneous across different regions and countries across the globe. The understanding of the SARS-CoV-2 biology is a vital part for the development of effective strategies for public health care and disease management. This work aims to analyze the available genomes sequenced in Brazil between February 2020 and February 2021, in order to identify mutation hotspots, geographical and temporal distribution of SARS-CoV-2 lineages in the Brazilian territory by using phylogenetics and phylodynamics analyses from high-quality genomes. We describe heterogeneous and episodic sequencing efforts, the progression of the different lineages along time, evaluating mutational spectra and frequency oscillations derived from the prevalence of novel and specific lineages across different Brazilian regions. We found at least seven major (1-7) and two minor clades (4.2 and 5.3) related to the six most prevalent Brazilian lineages and described its distribution across the Brazilian territory. The emergence and recent frequency shift of lineages (P.1 and P.2) containing mutations of concern in the spike protein (*e. g.*, E484K, N501Y) draws attention due to their association with immune evasion and enhanced receptor binding affinity. Improvements in genomic surveillance are of paramount importance and should be extended in Brazil to better inform policy makers and enable precise evidence-based decisions to fight the COVID-19 pandemic.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

After its initial emergence in China in late 2019 (Huang et al. 2020), *Severe Acute Respiratory Syndrome 2 Virus* (SARS-CoV-2) has spread rapidly around the world causing the COVID-19 pandemic (World Health Organization 2020). New epicenters of the disease have been established throughout 2020, mainly in Europe, USA, and South America (Ruiu 2020; Pei et al. 2020). As of February 22, 2021, more than 110 million cases and approximately 2.5 million deaths worldwide have been confirmed (Johns Hopkins Coronavirus Resource Center, 2021). Several countries are currently experiencing second waves of infections, decreasing optimism regarding a brief solution to the pandemic.

Since the sequencing of the first SARS-CoV-2 genome (Zhou et al. 2020), international efforts of open science have been established through data sharing in the GISAID database (Shu and McCauley 2017). These sequencing and metadata information were made public and have enabled the study of the viral spread pattern through space and time (Pybus and Rambaut 2009). However, sequencing facilities and research investments are very heterogeneous across the world. Asian, European, North American and Oceanian countries have contributed with more data proportionally to the number of cases (Furuse 2021), while African and South American genomic surveillance have been more limited. Disparities are even deeper on an individual country basis. For instance, while the United Kingdom leads sequencing efforts (5.85% of cases sequenced), Brazil has sequenced only 0.03% of all its cases, despite being the third worst-hit country, with approximately 10.2 million cases and 250 thousand deaths.

Many studies were performed to characterize early viral introductions and transmission dynamics in several countries (*e. g.* China (Lu et al. 2020), USA (Deng et al. 2020; Worobey et al. 2020; Maurano et al. 2020), Australia (Seemann et al. 2020), Italy (Bartolini et al. 2020), United Kingdom (da Silva Filipe et al. 2021; du Plessis et al. 2021). In Brazil, SARS-CoV-2 arrived officially on February 25, 2020, in a returning traveller from Italy, and early efforts were made both at the national (Candido et al. 2020b) and regional levels (Xavier et al. 2020; Paiva et al. 2020) to further characterize viral introduction and spread. B.1 and derived lineages were prevalent in the country at the beginning of the pandemic and significant movements between state borders after

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

international travel restrictions have been demonstrated (Candido et al. 2020b). Unfortunately, little is known about the viral evolution in the entire Brazilian territory after these earliest studies.

More recently, new SARS-CoV-2 lineages have been emerged and are considered as "Variants Of Concern" (VOC) , mainly those carrying mutations in the spike (S) glycoprotein due to its role in binding to the human ACE2 receptor (hACE2), allowing the virus to invade the host cell. Up to February 2021, there are three VOCs described worldwide, namely: B.1.1.7, B.1.351, and P.1. The former emerged in England in mid-September 2020 and it is characterized by 14 lineage-specific amino acid substitutions, especially N501Y (a key contact residue interacting with hACE2) and P681H (one of four amino acids comprising the insertion that creates a novel furin cleavage site between S1 and S2) in the S protein (Rambaut et al. 2020b). The second emerged in South Africa in October 2020 and harbor a constellation of mutations in Receptor Binding Domain (RBD) of the S protein (especially K417N, E484K and N501Y) (Tegally et al. 2020). The last and more recent lineage is P.1, derived from B.1.1.28, a widespread lineage from Brazil. It was recently reported in returning travelers from Manaus (Amazonas, Brazil), after arriving in Japan. It has the same three mutations (except for K417T instead of 417N) in the RBD as the South African lineage, but it arose independently (Faria et al. 2021a). Importantly, B.1.351 and P.1 carry the E484K mutation associated with escape from neutralizing antibodies, which may be passively acquired by convalescent plasma or actively induced by vaccination (Baum et al. 2020; Weisblum et al. 2020; Greaney et al. 2020). Recently, a E484K harboring virus was identified in a reinfected patient (Nonaka et al. 2021) from Brazil, confirming the ability to evade naturally developed antibodies from previous infection as well. Moreover, all three VOC lineages harbor N501Y mutation, already associated with enhanced receptor binding affinity (Starr et al. 2020), which could lead to increased infectiousness.

It is believed that after more than a year of its emergence, some mutations of SARS-CoV-2 (*e. g.* E484K and N501Y) have been positively selected, since they may confer adaptive advantages leading to convergent evolution in different lineages spreading across multiple countries. Despite the low sequencing rate, a deeper analysis of mutations and lineages throughout the Brazilian states would allow a better understanding of the relative importance of different mutations and their contribution to global viral diversity. Thus, we aimed to identify

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

mutation hotspots, geographical and temporal distribution of SARS-CoV-2 lineages in the Brazilian territory by using phylogenetics and phylodynamics analyses from high-quality SARS-CoV-2 genome sequences.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

SARS-CoV-2 genomes and epidemiological data retrieval

Complete SARS-CoV-2 genomes (>29,000 bp) and the associated metadata were obtained from the GISAID database. Considering 2,751 available sequences from Brazil submitted until February 16 2021, 2732 were retrieved applying filters for human host and complete collection date. Number of cases per state per day and across Brazil were downloaded from https://covid19br.wcota.me/en/ (Cota 2020) on the same date. This initiative aggregates data from the Brazilian Ministry of Health and epidemiological bulletins of each federative unit.

Mutation analysis

The GenBank RefSeq sequence NC_045512.2 from Wuhan (China) was used as the reference for our analysis. Single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) were assessed by using snippy variant calling pipeline v4.6.0 (https://github.com/tseemann/snippy), which uses FreeBayes v1.3.2 as variant caller and snpEff v5.0 (Cingolani et al. 2012) to annotate and predict the effects of variants on genes and proteins. Mutations and lineages were concatenated with associated metadata and counted by Brazilian states using custom Python and R scripts. Histogram of SNPs were generated after running MAFFT v7.471 alignment using a modified code from Lu et al. 2020 (https://github.com/laduplessis/SARS-CoV-2_Guangdong_genomic_epidemiology/).

Phylogenetics, phylogeographic and phylodynamics analysis

All available SARS-CoV-2 genomes (537,360 sequences) were retrieved from GISAID on February 16, 2021. These sequences were then subjected to analysis inside NextStrain ncov pipeline (https://github.com/nextstrain/ncov; Hadfield et al. 2018) through a Brazilian-focused subsampling scheme using time- and worldwide-representative contextual samples.

In this workflow, sequences were filtered out based on high divergence, incompleteness and sampling date availability. Next, filtered genomes were aligned using nextalign v0.1.6 (https://github.com/neherlab/nextalign) and their ends were masked (100 positions in the beginning, 50 in the end). Maximum likelihood (ML) phylogenetics tree was built using IQ-TREE

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

v2.0.3 (Nguyen et al. 2015), employing the best-fit model of nucleotide substitution as selected by ModelFinder (Kalyaanamoorthy et al. 2017). The root of the tree was placed between lineage A and B (Wuhan/Hu-1/2019 and Wuhan/WH01/2019). The clock-like behavior of the inferred tree was inspected using TempEst v1.5.3 (Rambaut et al. 2016) to generate the root-to-tip regression against sampling dates (correlation coefficient = 0.83, R^2 = 0.68). Sequences that deviate more than four interquartile ranges from the root-to-tip regression were removed from the analysis. Time-scaled ML phylogenetics tree was generated using TreeTime v0.8.1 under a strict clock and a skyline coalescent prior with a rate of $8x10^{-4}$ substitutions per site per year (Sagulenko et al. 2018). Results were then exported to JSON format to enable interactive genetic and geographical visualization using Auspice. Additionally, ML and time-stamped trees were visualized using FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/) and ggtree R package v2.0.4 (Yu et al. 2017). We identified global lineages using the dynamic nomenclature implemented in Pangolin v2.2.2 (https://github.com/cov-lineages/pangolin; Rambaut et al. 2020a).

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Results

Distribution of Brazilian sequences through time and space

Sequencing efforts from Brazil were concentrated mainly in the first epidemic wave (March to April, 2020) (Figure 1A). In March, 503 genomes (8.64% of the confirmed cases) and in April, 942 sequences (1.16% of the cases) were sampled. All following months fell below 1% of sequencing rate (Table S1). All Brazilian states sequenced less than 0.1% of the confirmed cases through the first year of the pandemic (Figure 1B). From the Southeast region, the states of Rio de Janeiro (0.09%) and São Paulo (0.06%) have led the country's sequencing initiatives, followed by Rio Grande do Sul (0.045%) from the South region, Amazonas (0.041%) from the North region and Pernambuco (0.040%) from the Northeast region. The Centre-West was the region with the lowest sequencing rate (Figure 1B, Table S2).

The Southeast region contributed with 1,704 genomes (62.53%), followed by Northeast (n=359; 13.17%), South (n=319; 11.71%), North (n=310; 11.38%), and Centre-West (1.21%) regions (Figure 1C, Table S3). In total, 59 different lineages were detected in Brazil, and the states sequencing more genomes (São Paulo, Rio de Janeiro and Rio Grande do Sul) detected higher numbers of circulating lineages (33, 17 and 16, respectively) (Figure 1D, Table S3). Importantly, all States that sequenced at least 100 genomes identified ≥ 10 lineages (Table S3).

High frequency mutations

A total of 3,919 mutations were detected across the 2,731 Brazilian genomes and only 354 (12.96%) occured in >5 sequences, 44 (1.61%) in >50 genomes, and 38 (1.39%) in >100 sequences (Figure 2, Table 1). Twenty-five (65.79%) of these 38 mutations were non-synonymous. Of these, 11 (44.0%) were in the spike protein, 5 (20.0%) in the nucleocapsid protein, and 5 (20.0%) in the ORF1ab polyprotein (Table 1).

Three mutations were found in >95% of the genomes: A23403G (S:D614G), C14408T (ORF1ab:L4715), and C3037T (ORF1ab:F924), which are signatures of the B.1 and derived lineages that spread early in the pandemic. The adjacent replacement GGG28881AAC (N:RG203-204KR) was found in 85.17%, representing a clear signature of B.1.1 lineage. The defining-mutations of the Brazilian most widespread lineages (B.1.1.28 and B.1.1.33) were also

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

found in high abundance. The G25088T (S:V1176F) replacement from B.1.1.28 occured in 47.56% of all sequences, while T29148C (N:I292T) and T27299C (ORF6:I33T) from B.1.1.33 in ≈32.5%. The G23012A (S:E484K) mutation in the Receptor Binding Domain (RBD) of spike that recently emerged independently in three Brazilian lineages (B.1.1.33, P.1 and P.2) is already among the most frequent detected up to February, 2021 (11.42%). The E484K viruses have been spreading mostly between mid-2020 up to early-2021. Additionally, the multiple lineage-defining mutations found in emergent P.1 and P.2 lineages from Brazil were observed in >100 and >200 genomes, respectively (Figure 2, Table 1).

Lineage dynamics of Brazilian epidemic

In March 2020, the majority of Brazilian sequences belonged to 3 lineages: B.1 (n=101; 20.08%), B.1.1.28 (n=156; 31.01%) and B.1.1.33 (n=131; 26.04%). The first was probably introduced in Brazil through multiple imports from other continents, and the others probably emerged from community transmission inside the country (Candido et al. 2020b). Between April and August, >75% of all sequences per month were classified as B.1.1.28 or B.1.1.33. During October and November, the B.1.1.28 derived lineage P.2 was the most prevalent (n=32; 37.65% and n=92; 40.71%), while B.1.1.28 and B.1.1.33 represented together <50% of the genomes. From December 2020 onward, the VOC P.1 emerged in Manaus and has been established, together with P.2, as the most prevalent lineages represented by sequencing data (Figure 3A).

Regarding distribution between the five different Brazilian regions, the Southeast and Southern regions sequenced a larger proportion of B.1.1.28 and B.1.1.33 viruses. The Northeast apparently has a slightly different dynamics, since B.1.1.74 (n=138; 38.44%) is the most prevalent lineage followed by B.1.1.33 (n=91; 25.35%). In the Northern region, P.2 is already the second most prevalent lineage (n=81; 26.13%) (Figure 3B and 3C), but this may be related to the low quantity of sequences from the beginning of the pandemic, and the higher surveillance in the region at present, due to enhanced sequencing efforts after the emergence of P.1. In the Centre-West, the extremely low sequencing rate prevents us from making any assumptions about the genetic diversity of the circulating lineages (Figure 3B).

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Statewide view shows the wide distribution of lineages B.1.1.28, B.1.1.33 and P.2 across almost all Brazilian states. Although the P.1 lineage has been represented by genomes from only 7 states (AM, PA, RS, RO, RR, SC, and SP) up to February 16 2021, other states have already reported its detection after this date. States reporting a higher proportion of B.1 lineage (CE, GO, MG, SC, and SE) have concentrated their sequencing efforts mainly in the early phase of the pandemic. B.1.1.74 is overrepresented (>50%) in two states from the Northeast (PE and PB), suggesting a more geographically restricted distribution. Additionally, the contribution of other lineages that are not among the 10 most frequent in the country are apparently limited (Figure 3C and Figure S1).

The time between the first and last detection of the 18 lineages represented by more than five Brazilian genomes varied markedly, with a mean of 186.28 days, median of 179.0 days and standard deviation of 119.58 days. Eleven of the 18 lineages were sampled in different states between the first and last detection, suggesting its spread through the Brazilian territory (Table S4). B and B.1 lineages have presented a relatively low time of spread, while several B.1.1 derived lineages probably have spreaded for a longer period (*e. g,* B.1.1.143, B.1.1.28, B.1.1.314, B.1.1.33, B.1.1.74, B.1.1.94) (Figure S2). B.1.1.7 (UK-associated lineage) was firstly detected in the country in mid-December (Claro et al. 2021), while P.1 was found in November in Manaus (Amazonas, Brazil) and Japan (Faria et al. 2021a). P.2 was discovered in mid-April 2020, but has been increasing rapidly in frequency since August 2020 (Figure 3A, Figure S2).

Phylogenetics, phylogeographic and phylodynamic patterns of SARS-CoV-2 spread in Brazil

We obtained 10,573 time-, geographical- and genetic-representative genomes to proceed phylogenetic inferences. Of these, 1,135 were from Africa, 1,963 were from Asia, 3,248 from Europe, 612 from North America, 285 from Oceania, and 3,330 from South America. Among the latter, 2,350 were from Brazil.

The maximum likelihood tree showed the evolutionary diversity of SARS-CoV-2 sequences across Brazilian territory (Figure 4A). Almost all sequences from Brazil harbor the S:D614G and ORF1b:P314L, which were imported from other continents to Brazil (mainly to Southeastern states)

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

in the early epidemic wave of COVID-19. After the importation of B.1.1 lineages characterized by N:R203K and N:G204R mutations, community transmission massively occured and gave rise to B.1.1.28 (S:V1176F) and B.1.1.33 lineages (ORF6:I33T and N:I292T), widely distributed in Brazilian regions along the first year of the pandemic (Figure 4A and 4B). B.1.1.28 have diversified in two lineages, P.1 (20J/501Y.V3) and P.2, which are already widely represented by many sequences and distributed across all regions. The larger branch length leading to P.1 lineage draws attention and its emergence is probably driven by an accelerated molecular evolution (Figure 4B and 5A).

We estimated an evolutionary rate of 7.76×10^{-4} substitutions per site per year (23.22 mutations per year) for the Brazilian-focused subsampling (Figure 5A). Since the end of 2020, two evolutionary patterns are observed in the root-to-tip regression of sampling dates. Despite the majority of sequences follow the expected substitution rate for SARS-CoV-2, we observe a rise in P.1 (20J/501Y.V3) and B.1.1.7 (20I/501Y.V1) abundance, both characterized by an abnormal clock rate (Figure 5A) and a constellation of mutations in the spike protein. Time-resolved maximum likelihood tree highlighted the importation of lineages in the early phase of the pandemic and its rapid diversification through community transmission inside the country, leading the spread of two main lineages (B.1.1.28 and B.1.1.33) within and between state borders (Figure 5B) and a more restricted circulation of B.1.1.74 lineage in the Northeast. A time-resolved phylogeny considering only Brazilian sequences reinforces the nationwide distribution of multiple SARS-CoV-2 lineages and clades (Figure 5C and Figure S3).

We found at least seven major clades (1-7) (Figure 6) and two minor (4.2 and 5.3) (Figure S4) related to the six most prevalent Brazilian lineages. Clade 1 (B.1 lineage) is represented by 33 genomes, >50% from the Southeast with a few introductions to other regions and a restricted time of spread until around May 2020 (Figure 6B). Clade 2 includes 98 genomes and is associated with B.1 and B.1.212 lineages. B.1 (Clade 2.1) sequences from this clade are mostly restricted to Southern and Southeast Brazil, while B.1.212 (Clade 2.2) has spread mainly in the Northeast and Northern regions (Figure 6C). Clade 3 is represented by 804 genomes, mostly from B.1.1.33 lineage. It reaches an unprecedented dissemination through Brazil since March, whose introductions in different regions and states were mostly driven by the Southeast followed by

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

massive community transmission and the establishment of local clusters (Figure 6D). Clade 4 (n=107) is characterized by B.1.1.74 sequences and it is more geographically restricted to the Northeast, especially Pernambuco (n=41) and Paraiba (n=23). However, it accounts for occasional introductions to Southern and Southeastern regions (Figure 6E). Clade 5 harbors B.1.1.28 (Clade 5.1) and P.1 (Clade 5.2) sequences. B.1.1.28 sequences are mostly widespread in Northern and Southeastern regions (Figure 6F), and P.1 genomes (n=96) are mostly found in Northern Brazil, specially Amazonas (n=50) and Rondônia (n=15), and Southeast (São Paulo; n=17) (Figure 6H). Clade 6 is also represented by genomes that fall into the B.1.1.28 lineage (n=529). However, it is more widespread in the Southeast (n=436), especially in São Paulo (n=412) and in the Southern region (n=72), especially Rio Grande do Sul (n=65) (Figure 6G). Clade 7 is highly supported by sequences of P.2 lineage (n=204), which harbor B.1.1.28 defining mutations and its proper ones. This clade is also widely distributed throughout Brazilian regions, especially the Southeast (n=83), South (n=63), and North (n=42), even giving rise to new local clusters in the end of 2020 and early-2021 (Figure 6H).

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

Viral sequencing is essential to track viral evolution and spread patterns. Despite the initial efforts to obtain a representative genomic dataset of the Brazilian first epidemic wave to better characterize viral introductions and early spread (Candido et al. 2020b), the initiatives across different regions have been limited and non-uniform after on. Interestingly, the previously described clades (Candido et al. 2020b) have spread and diversified through the country. Clade 1 (after named B.1.1.28) were mostly restricted to the Southeast (São Paulo) and Clade 2 (after named B.1.1.33) were already present in 16 states in this early phase. In this work, we showed the emergence of at least four clades derived from B.1.1.28. The first was widely distributed in Brazilian regions (Clade 5.1) and evolved to P.1 lineage (Clade 5.2). The second (Minor Clade 5.3) and third (Clade 6) are most widespread in the Southeast, accounting for a few introductions in other regions. The fourth gave rise to P.2 lineage (Clade 7), which is distributed in all Brazilian regions. B.1.1.33 continues to be composed of a larger clade (Clade 3) with a wide distribution among all Brazilian regions.

The B.1.1.33 lineage was studied in further details using 190 genomes from 13 Brazilian states, showing its variable abundance in different states (ranging from 2% in Pernambuco to 80% in Rio de Janeiro), and its moderate prevalence in South American countries (5-18%). Surprisingly, this lineage was firstly detected in early-March in other American countries (*e. g.,* Argentina, Canada, and USA), and additional analysis suggest that an intermediate lineage (B.1.1.33-like) most probably arose in Europe and was later disseminated to Brazil, where its spread gave origin to lineage B.1.1.33 (Resende et al. 2021) and possibly seeded secondary outbreaks in Argentina and Uruguay (Resende et al. 2021; Mir et al. 2021).

The states of Pernambuco (Northeast) and Minas Gerais (Southeast) presented more restricted viral dynamics. In Pernambuco, 88% of 101 early sequences were classified as lineage B.1.1 and six local B.1.1 clades were identified seeded through both national and international traveling (Paiva et al. 2020). This finding is consistent with the prevalence of B.1.1.74 lineage found in the Northeast, especially in Pernambuco. In Minas Gerais, 92.5% of the 40 genomes from March 2020 belong to the B lineage (mostly B.1.1) and epidemiological analysis revealed that the

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

distribution of cases and deaths was more spatially uniform, while in other Southeastern states it was more centralized around capital cities (Xavier et al. 2020).

Studies from Rio de Janeiro (Southeast) and Rio Grande do Sul (South) identified B.1.1.28 and B.1.1.33 in higher proportion from April to December 2020 (Voloch et al. 2020; Franceschi et al. 2021; Francisco Jr et al. 2021). The emergence of a B.1.1.28-derived lineage carrying the E484K mutation (after named P.2) was dated in July 2020, however it began to appear more frequently and almost simultaneously in October 2020 in the Rio de Janeiro state (Voloch et al. 2020) and in the small municipality of Esteio, Rio Grande do Sul (Franceschi et al. 2021), suggesting its wide distribution in the Southern and Southeastern regions of Brazil and uncertainty regarding its origin. This assumption and the frequency increase of B.1.1.28 and derived lineages were corroborated by another study from several municipalities of Rio Grande do Sul, which found that 86% of the sequenced genomes were classified as B.1.1.28 and ~50% of these belong to the new lineage P.2 (Francisco Jr et al. 2021). Here, we found that P.2 is already distributed in all Brazilian regions up to mid-February 2021.

Recent findings using 250 genomes (March 2020 to January 2021) from Manaus, showed that the first exponential growth phase was driven mostly by the dissemination of B.1.195 which was gradually replaced by B.1.1.28, and the second wave coincides with the emergence of the VOC P.1. This variant probably evolved from a local B.1.1.28 clade in late November and replaced the parental lineage in less than two months. An evolutionary intermediate between B.1.1.28 and P.1 (named P.1-like) was identified, suggesting that the diversity of SARS-CoV-2 variants harboring spike mutations in Manaus could be larger than initially expected and that those variants probably circulated for some time before the emergence and expansion of P.1 (Naveca et al. 2021).

Another study following the circulation of P.1 estimated its emergence to November 2020 preceded by a period of faster molecular evolution. Additionally, virus exchanges between Amazonas and the urban metropolises in Southeast Brazil follow patterns in national air travel mobility, since states reporting P.1 until end-February 2021 received around 100,000 air passengers from Manaus in November. Of the 10 new amino acid mutations in the spike protein (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I) compared to its

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

immediate ancestor (B.1.1.28), molecular selection analyses found evidence that 9 of these 10 mutations are under diversifying positive selection (Faria et al. 2021b).

Detecting mutations that are subjected to positive pressure is of paramount importance in order to predict the SARS-CoV-2 pandemic future. By correlating amino acid replacements with expected structural changes, it is possible to anticipate risk of immune evasion with consequent infection recurrence and or vaccine mismatching. Various specific mutations were encountered in different lineages that are potentially associated with selective advantages. RBD and its hACE-2 interacting core, the Receptor Binding Motif (RBM), is of evident importance, since substitutions in this motif were associated with increased receptor binding forces (*e. g.*, N501Y) or immune evasion (*e.g.*, E484K). The E484K seems to be of particular relevance, as its presence shifts the main interaction residue to this site. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces a higher number of conformational changes than N501Y mutant alone, potentially resulting in an escape mutant (Nelson et al. 2021). Since this site is not involved in interaction with hACE-2 when the original glutamate is in place, its occurrence has been linked to reinfection, convalescent plasma activity abolishment and decreased post-vaccination neutralizing activity (Nelson et al. 2021; Li et al. 2021).

Other mutations likely play important roles by allosteric mechanisms and have been positively selected early during the SARS-CoV-2 pandemic. Almost all Brazilian sequences harbor S:D614G, a hallmark of the ancestral B.1 lineage. Although this mutation is outside the RBD, it is speculated that it abolishes the hydrogen bond between the 614 position in S1 and a threonine residue located at S2 from the neighbour protomer. In consequence, RBD would be locked in its activated “up” position, thus increasing viral infectivity (Ozono et al. 2021). Therefore, the establishment of this mutation in Brazilian sequences seems to be related to either a founder effect based on the importation of primarily G614 variants to Brazil and an evolutionary advantage in comparison with D614.

The precise forces that drive the appearance of complex mutational signatures characteristic of different lineages over short time periods remain largely unknown. Under specific circumstances, the combination of prolonged viral shedding with high selective pressure could lead

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

to major evolutionary leaps. Critically ill and immunosuppressed patients chronically infected with SARS-CoV-2 and treated with convalescent plasma have been linked to viral breakthroughs caused by mutant viruses (Kemp et al. 2021). Whichever phenomena allow for rapid viral evolution, they probably have to permit multiple substitutions to occur almost simultaneously, since our phylodynamics analysis shows constant and relatively slow rates of mutation accumulation, except for VOC viruses.

Although the proximal origins of the most important VOCs remain to be determined, some conclusions about their nature could be already drawn. First, the mutations shared between P.1 and B.1.1.351 seem to be associated with a rapid increase in cases even in locations where previous attack rates were thought to be very high (Buss et al. 2020). Lineage P.1, emerged from the Brazilian state of Amazonas between November and December 2020, have accumulated a high number of non-synonymous mutations and is now dispersed across novel Brazilian regions, representing one of the most frequent lineages up to February 2021 (Faria et al. 2021a; Naveca et al. 2021). Second, the fact that the set of mutations shared by P.1, B.1.1.7 and B.1.351 seem to have arisen independently, as we have previously demonstrated with emergence of E484K in others Brazilian lineages (P.2, B.1.1.28 and B.1.1.33), is suggestive of convergent molecular evolution (Ferrareze et al. 2021; Faria et al. 2021b).

Our study shows that the Brazilian territory was affected by at least 59 different lineages during the COVID-19 pandemic. This is not completely unexpected, considering the size of the country and its touristic and economical relevance. South African, the original source of B.1.351 lineage, similarly had multiple and diverse viral introductions. Of note, a recent genomic study detected 42 different circulating lineages in the country, between the first epidemic wave (March) and mid-September, 2020. Moreover, the three main lineages (B.1.1.54, B.1.1.56 and C.1), which represented the majority of cases in the first wave, were responsible for ~42% total of the infections by the end of 2020, since B.1.351 had emerged in an explosive fashion in mid-October (Tegally et al. 2021). This later lineage is of major concern and South Africa is, up to February 2021, the leader country in COVID-19 related deaths in Africa (Li et al. 2021; Johns Hopkins Coronavirus Resource Center, 2021). In Brazil, the prevalence of lineages B.1.1.28 and its derivatives P.1 and P.2 have been representing progressively more cases of the sequenced

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

genomes by the time of writing. In March, 2020, B.1.1.28 was one of the 28 circulating lineages present in the country, with more than 30% of the sequenced genomes. At its side, B.1.1.33 was the identified lineage for approximately 26% of analyzed genomes. Ten months later (January, 2021), the VOC P.1 appeared as the prevalent one among 12 different lineages, reaching more than 45% of the sequenced cases. Next, followed the P.2 lineage (~27%). At that point, B.1.1.28 and B.1.1.33, which achieved 85.5% of the sequenced genomes together in June, matched for, in January, less than 10% of the cases. Therefore, in both countries, despite potentially different initial founding effects that could have led to diverse lineage dissemination patterns, eventually complex VOCs harboring advantageous mutations were selected for viral spread, indicating an increased evolutionary fitness for these viruses.

Unfortunately, incomplete and erratic sequencing efforts have limited a better SARS-CoV-2 characterization in Brazil, since genomes are not equally distributed in geographical or temporal scales due to episodic sampling efforts prompted by resource availability. This is reflected by a very small fraction of SARS-CoV-2 cases being sequenced. Unequal temporal distribution implies that some of the conclusions are disproportionately affected by events in heavily sampled periods (March-May). Therefore, different lineage distributions could be an artefact of distinct sequencing coverage among states and across different time-frames. Additionally, the majority of the sequences come from the Southeastern region of Brazil. While this is in fact an economic and travel hub for the country, accounting for >70% of the international passengers arriving in Brazil in the beginning of the pandemic (Candido et al. 2020a), inferences regarding this region can be inflated in relation to undersampled regions. However, this is, to the best of our knowledge, the first attempt to characterize SARS-CoV-2 mutations, phylogenetics, phylogeography and phylodynamics in the entire Brazil after the study that characterized the first epidemic wave in Brazil using 490 representative genomes (Candido et al. 2020b).

In summary, by systematic analysis of viral genomes distributed across Brazil over time, we were able to confirm the early introductions of multiple lineages, its rapid diversification to constitute new lineages, probable convergent evolution of important mutations (*e. g.,* E484K, N501Y), and the emergence of P.1, arguably one of the most potentially dangerous lineages identified worldwide up to February 2021. The occurrence of this lineage and the emergence of

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

newer variants could jeopardize the efficacy of vaccines and immunotherapies and may lead the health care system to overload. We concluded that enhanced genomic surveillance is, therefore, of paramount importance and should be extended as soon as possible as a means to better inform policy makers and enable precise evidence-based decisions to fight the COVID-19 pandemic.

Data availability statement

A full table acknowledging the authors and corresponding labs submitting sequencing data used in this study can be found in Supplementary File 1. Additional information used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interest statement

The authors declare no competing interests.

Funding

Scholarships and Fellowships were supplied by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and Universidade Federal de Ciências da Saúde de Porto Alegre. The funders had no role in the study design, data generation and analysis, decision to publish or the preparation of the manuscript.

Acknowledgements

We thank the administrators of the GISAID database and research groups across the world (especially Brazilians) for supporting the rapid and transparent sharing of genomic data during the COVID-19 pandemic. We also thank the Mayor's Office, Health Department and São Camilo Hospital (Esteio, RS, Brazil), Leonardo Duarte Pascoal and Ana Regina Boll for their work in combating COVID-19 and for supporting the work developed by our research group.

Author contributions

Vinicius B. Franceschi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Patrícia A. G. Ferrareze**: Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. **Ricardo A. Zimerman**: Formal analysis, Investigation, Writing -

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Original Draft, Writing - Review & Editing. **Gabriela B. Cybis**: Methodology, Formal analysis, Investigation, Writing - Review & Editing. **Claudia E. Thompson:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

References

Bartolini B, Rueca M, Gruber CEM, Messina F, Carletti F, Giombini E, Lalle E, Bordi L, Matusali G, Colavita F et al. SARS-CoV-2 Phylogenetic Analysis, Lazio Region, Italy, February–March 2020 - Volume 26, Number 8—August 2020 - Emerging Infectious Diseases journal - CDC. doi: 10.3201/eid2608.201525

Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, Giordano S, Lanza K, Negron N, Ni M et al. (2020) Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science 369:1014–1018. doi: 10.1126/science.abd0831

Buss LF, Prete CA, Abrahim CMM, Mendrone A, Salomon T, Almeida-Neto C de, França RFO, Belotti MC, Carvalho MPSS, Costa AG et al. (2020) Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. Science. doi: 10.1126/science.abe9728

Candido DDS, Watts A, Abade L, Kraemer MUG, Pybus OG, Croda J, de Oliveira W, Khan K, Sabino EC and Faria NR (2020a) Routes for COVID-19 importation in Brazil. J Travel Med. doi: 10.1093/jtm/taaa042

Candido DS, Claro IM, Jesus JG de, Souza WM, Moreira FRR, Dellicour S, Mellan TA, Plessis L du, Pereira RHM, Sales FCS et al. (2020b) Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science 369:1255–1260. doi: 10.1126/science.abd2161

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X and Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6:80–92. doi: 10.4161/fly.19695

Claro IM, Sales FC da S, Ramundo MS, Candido DS, Silva CAM, Jesus JG de, Manuli ER, Oliveira CM de, Scarpelli L, Campana G et al. Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020 - Volume 27, Number 3—March 2021 - Emerging Infectious Diseases journal - CDC. doi: 10.3201/eid2703.210038

Cota W (2020) Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level. doi: 10.1590/SciELOPreprints.362

da Silva Filipe A, Shepherd JG, Williams T, Hughes J, Aranday-Cortes E, Asamaphan P, Ashraf S, Balcazar C, Brunker K, Campbell A et al. (2021) Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. Nat Microbiol 6:112–122. doi: 10.1038/s41564-020-00838-z

Deng X, Gu W, Federman S, Plessis L du, Pybus OG, Faria N, Wang C, Yu G, Bushnell B, Pan C-Y et al. (2020) Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. doi: 10.1126/science.abb9263

du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, Raghwani J, Ashworth J, Colquhoun R, Connor TR et al. (2021) Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. doi: 10.1126/science.abf2946

Faria N, Claro IM, Candido D, Franco LAM, Andrade PS, Coletti TM, Silva CAM, Fraiji NA, Esashika Crispim MA, Carvalho M do PSS et al. (2021a) Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. In: Virological. https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586. Accessed 14 Jan 2021

Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, Crispim MAE, Sales FC, Hawryluk I, McCrone JT et al. (2021b) Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil. medRxiv 2021.02.26.21252554. doi: 10.1101/2021.02.26.21252554

Ferrareze PAG, Franceschi VB, Mayer A de M, Caldana GD, Zimerman RA and Thompson CE (2021) E484K as an innovative phylogenetic event for viral evolution: Genomic analysis of the E484K spike mutation in SARS-CoV-2 lineages from Brazil. bioRxiv 2021.01.27.426895. doi: 10.1101/2021.01.27.426895

Franceschi VB, Caldana GD, Mayer A de M, Cybis GB, Neves CAM, Ferrareze PAG, Demoliner M, Almeida PR de, Gularte JS, Hansen AW et al. (2021) Genomic Epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. medRxiv 2021.01.21.21249906. doi:

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

10.1101/2021.01.21.21249906

Francisco Jr R da S, Benites LF, Lamarca AP, de Almeida LGP, Hansen AW, Gularte JS, Demoliner M, Gerber AL, de C Guimarães AP, Antunes AKE et al. (2021) Pervasive transmission of E484K and emergence of VUI-NP13L with evidence of SARS-CoV-2 co-infection events by two different lineages in Rio Grande do Sul, Brazil. Virus Res 296:198345. doi: 10.1016/j.virusres.2021.198345

Furuse Y (2021) Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. Int J Infect Dis 103:305–307. doi: 10.1016/j.ijid.2020.12.034

Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, Hilton SK, Huddleston J, Eguia R, Crawford KHD et al. (2020) Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe. doi: 10.1016/j.chom.2020.11.007

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T and Neher RA (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123. doi: 10.1093/bioinformatics/bty407

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X et al. (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 395:497–506. doi: 10.1016/S0140-6736(20)30183-5

Johns Hopkins Coronavirus Resource Center COVID-19 Map. In: Johns Hopkins Coronavirus Resour. Cent. https://coronavirus.jhu.edu/map.html. Accessed 10 Nov 2020

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A and Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589. doi: 10.1038/nmeth.4285

Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, Hosmillo M, Rees-Spear C, Mlcochova P, Lumb IU et al. (2021) SARS-CoV-2 evolution during treatment of chronic infection. Nature 1–10. doi: 10.1038/s41586-021-03291-y

Li Q, Nie J, Wu J, Zhang L, Ding R, Wang H, Zhang Y, Li T, Liu S, Zhang M et al. (2021) No higher infectivity but immune escape of SARS-CoV-2 501Y.V2 variants. Cell. doi: 10.1016/j.cell.2021.02.042

Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H, Sun J, François S, Kraemer MUG, Faria NR et al. (2020) Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell 181:997-1003.e9. doi: 10.1016/j.cell.2020.04.023

Maurano MT, Ramaswami S, Zappile P, Dimartino D, Boytard L, Ribeiro-dos-Santos AM, Vulpescu NA, Westby G, Shen G, Feng X et al. (2020) Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City Region. Genome Res gr.266676.120. doi: 10.1101/gr.266676.120

Mir D, Rego N, Resende PC, López-Tort F, Fernandez-Calero T, Noya V, Brandes M, Possi T, Arleo M, Reyes N et al. (2021) Recurrent dissemination of SARS-CoV-2 through the Uruguayan-Brazilian border. medRxiv 2021.01.06.20249026. doi: 10.1101/2021.01.06.20249026

Naveca F, Nascimento V, Souza V, Corado A, Nascimento F, Silva G, Costa Á, Duarte D, Pessoa K, Mejía M et al. (2021) COVID-19 epidemic in the Brazilian state of Amazonas was driven by long-term persistence of endemic SARS-CoV-2 lineages and the recent emergence of the new Variant of Concern P.1. doi: 10.21203/rs.3.rs-275494/v1

Nelson G, Buzko O, Spilman P, Niazi K, Rabizadeh S and Soon-Shiong P (2021) Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. bioRxiv 2021.01.13.426558. doi: 10.1101/2021.01.13.426558

Nguyen L-T, Schmidt HA, von Haeseler A and Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 32:268–274. doi: 10.1093/molbev/msu300

Nonaka CKV, Franco MM, Gräf T, Mendes AVA, Aguiar RS de, Giovanetti M and Souza BS de F (2021) Genomic Evidence of a Sars-Cov-2 Reinfection Case With E484K Spike Mutation in Brazil.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

doi: 10.20944/preprints202101.0132.v1

Ozono S, Zhang Y, Ode H, Sano K, Tan TS, Imai K, Miyoshi K, Kishigami S, Ueno T, Iwatani Y et al. (2021) SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. Nat Commun 12:848. doi: 10.1038/s41467-021-21118-2

Paiva MHS, Guedes DRD, Docena C, Bezerra MF, Dezordi FZ, Machado LC, Krokovsky L, Helvecio E, da Silva AF, Vasconcelos LRS et al. (2020) Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil. Viruses 12:1414. doi: 10.3390/v12121414

Pei S, Kandula S and Shaman J (2020) Differential effects of intervention timing on COVID-19 spread in the United States. Sci Adv 6:eabd6370. doi: 10.1126/sciadv.abd6370

Pybus OG and Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 10:540–550. doi: 10.1038/nrg2583

Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L and Pybus OG (2020a) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5:1403–1407. doi: 10.1038/s41564-020-0770-5

Rambaut A, Lam TT, Max Carvalho L and Pybus OG (2016) Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. doi: 10.1093/ve/vew007

Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, Connor T, Peacock T, Robertson D and Volz E (2020b) Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. In: Virological. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563. Accessed 4 Jan 2021

Resende PC, Delatorre E, Gräf T, Mir D, Motta FC, Appolinario LR, Paixão ACD da, Mendonça AC da F, Ogrzewalska M, Caetano B et al. (2021) Evolutionary Dynamics and Dissemination Pattern of the SARS-CoV-2 Lineage B.1.1.33 During the Early Pandemic Phase in Brazil. Front Microbiol. doi: 10.3389/fmicb.2020.615280

Ruiu ML (2020) Mismanagement of Covid-19: lessons learned from Italy. J Risk Res 23:1007–1020. doi: 10.1080/13669877.2020.1758755

Sagulenko P, Puller V and Neher RA (2018) TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. doi: 10.1093/ve/vex042

Seemann T, Lane CR, Sherry NL, Duchene S, Gonçalves da Silva A, Caly L, Sait M, Ballard SA, Horan K, Schultz MB et al. (2020) Tracking the COVID-19 pandemic in Australia using genomics. Nat Commun 11:4376. doi: 10.1038/s41467-020-18314-x

Shu Y and McCauley J (2017) GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. doi: 10.2807/1560-7917.ES.2017.22.13.30494

Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC et al. (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell 182:1295-1310.e20. doi: 10.1016/j.cell.2020.08.012

Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N et al. (2020) Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medRxiv 2020.12.21.20248640. doi: 10.1101/2020.12.21.20248640

Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, Mlisana K, Bhiman JN, von Gottberg A, Walaza S et al. (2021) Sixteen novel lineages of SARS-CoV-2 in South Africa. Nat Med 1–7. doi: 10.1038/s41591-021-01255-3

Voloch CM, F R da S, Almeida LGP de, Cardoso CC, Brustolini OJ, Gerber AL, Guimarães AP de C, Mariani D, Costa RM da, Ferreira OC et al. (2020) Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. medRxiv 2020.12.23.20248598. doi: 10.1101/2020.12.23.20248598

Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, Muecksch F, Rutkowska M, Hoffmann H-H, Michailidis E et al. (2020) Escape from neutralizing antibodies by SARS-CoV-2

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

spike protein variants. eLife 9:e61312. doi: 10.7554/eLife.61312

World Health Organization WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. 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 10 Nov 2020

Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, Rambaut A, Suchard MA, Wertheim JO and Lemey P (2020) The emergence of SARS-CoV-2 in Europe and North America. Science. doi: 10.1126/science.abc8169

Xavier J, Giovanetti M, Adelino T, Fonseca V, Costa AVB da, Ribeiro AA, Felicio KN, Duarte CG, Silva MVF, Salgado Á et al. (2020) The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. Emerg Microbes Infect 9:1824–1834. doi: 10.1080/22221751.2020.1803146

Yu G, Smith DK, Zhu H, Guan Y and Lam TT-Y (2017) ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. doi: https://doi.org/10.1111/2041-210X.12628

Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L et al. (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. doi: 10.1038/s41586-020-2012-7

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Tables

Table 1. Mutations of Brazilian genomes found in > 100 genomes, associated effects on encoded proteins and major associated lineages.

REF	Genomic position	ALT	Effect	Amino acid	Gene	Product	Number of sequences	% in Brazilian genomes	Major associated lineages
A	23403	G	Missense	D614G	S	Surface glycoprotein	2651	97.07	B.1 and derived
C	14408	T	Synonymous	L4715L	ORF1ab	RdRp	2648	96.96	B.1 and derived
C	3037	T	Synonymous	F924F	ORF1ab	nsp3	2635	96.48	B.1 and derived
C	241	T	Intergenic	-	-	-	2374	86.93	B.1 and derived
GGG	28881	AAC	Missense	RG203-204KR	N	N phosphoprotein	2326	85.17	B.1.1 and derived
G	25088	T	Missense	V1176F	S	Surface glycoprotein	1299	47.56	B.1.1.28, B.1.1.143, B.1.1.94, P.1, and P.2
T	29148	C	Missense	I292T	N	N phosphoprotein	888	32.52	B.1.1.33, B.1.1.314, N.1, N.4
T	27299	C	Missense	I33T	ORF6	ORF6 protein	885	32.41	B.1.1.33, B.1.1.314
C	12053	T	Missense	L3930F	ORF1ab	nsp7	761	27.87	B.1.1.28, B.1.1.74, B.1.1.143, P.2
G	23012	A	Missense	E484K	S	Surface glycoprotein	312	11.42	B.1.1.33, P.1, P.2
T	26149	C	Missense	S253P	ORF3a	ORF3a protein	249	9.12	B.1.1.28, P.1
A	6319	G	Synonymous	P2018P	ORF1ab	nsp3	244	8.93	B.1.1.28, P.1
C	28253	T	Synonymous	F120F	ORF8	ORF8 protein	222	8.13	B.1.1.28, B.1.1.33, P.2
G	28975	T	Missense	M234I	N	N phosphoprotein	217	7.95	B.1.1.28, P.2
G	28628	T	Missense	A119S	N	N phosphoprotein	211	7.73	P.2
C	11824	T	Synonymous	I3853I	ORF1ab	nsp6	207	7.58	P.2
T	10667	G	Missense	L3468V	ORF1ab	3C-like proteinase	206	7.54	P.2
A	12964	G	Synonymous	G4233G	ORF1ab	nsp9	176	6.44	P.2
A	6613	G	Synonymous	V2116V	ORF1ab	nsp3	137	5.02	B.1.1.28, P.1
GTCTGGTTTT	11287	G	Deletion	3675-3677 SGF	ORF1ab	nsp6	130	4.76	B.1.1.7, P.1
C	29754	T	Intergenic	-	-	-	117	4.28	P.2
AGTAGGG	28877	TCTAAAC	Missense	RG203-204KR	N	N phosphoprotein	116	4.25	B.1.1.28, P.1
C	21614	T	Missense	L18F	S	Surface glycoprotein	116	4.25	B.1.1.28, P.1
G	17259	T	Missense	S5665I	ORF1ab	Helicase	116	4.25	P.1
A	23063	T	Missense	N501Y	S	Surface glycoprotein	115	4.21	B.1.1.7, P.1
C	21638	T	Missense	P26S	S	Surface glycoprotein	113	4.14	P.1
T	733	C	Synonymous	D156D	ORF1ab	Leader protein	111	4.06	P.1

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

C	13860	T	Missense	T4532I	ORF1ab	RdRp	111	4.06	P.1
C	24642	T	Missense	T1027I	S	Surface glycoprotein	111	4.06	P.1
C	12778	T	Synonymous	Y4171Y	ORF1ab	nsp9	111	4.06	P.1
C	21621	A	Missense	T20N	S	Surface glycoprotein	111	4.06	P.1
C	28512	G	Missense	P80R	N	N phosphoprotein	110	4.03	P.1
A	5648	C	Missense	K1795Q	ORF1ab	nsp3	109	3.99	P.1
C	23525	T	Missense	H655Y	S	Surface glycoprotein	109	3.99	P.1
G	28167	A	Missense	E92K	ORF8	ORF8 protein	108	3.95	P.1
G	21974	T	Missense	D138Y	S	Surface glycoprotein	107	3.92	B.1.1.33, P.1
G	22132	T	Missense	R190S	S	Surface glycoprotein	105	3.84	P.1
C	2749	T	Synonymous	D828D	ORF1ab	nsp3	105	3.84	P.1

REF: Reference nucleotide(s); ALT: Replaced nucleotide(s); UTR= Untranslated region; ORF=Open reading frame; S: Spike; N: Nucleocapsid; nsp: nonstructural protein; RdRp: RNA-dependent RNA polymerase.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figures

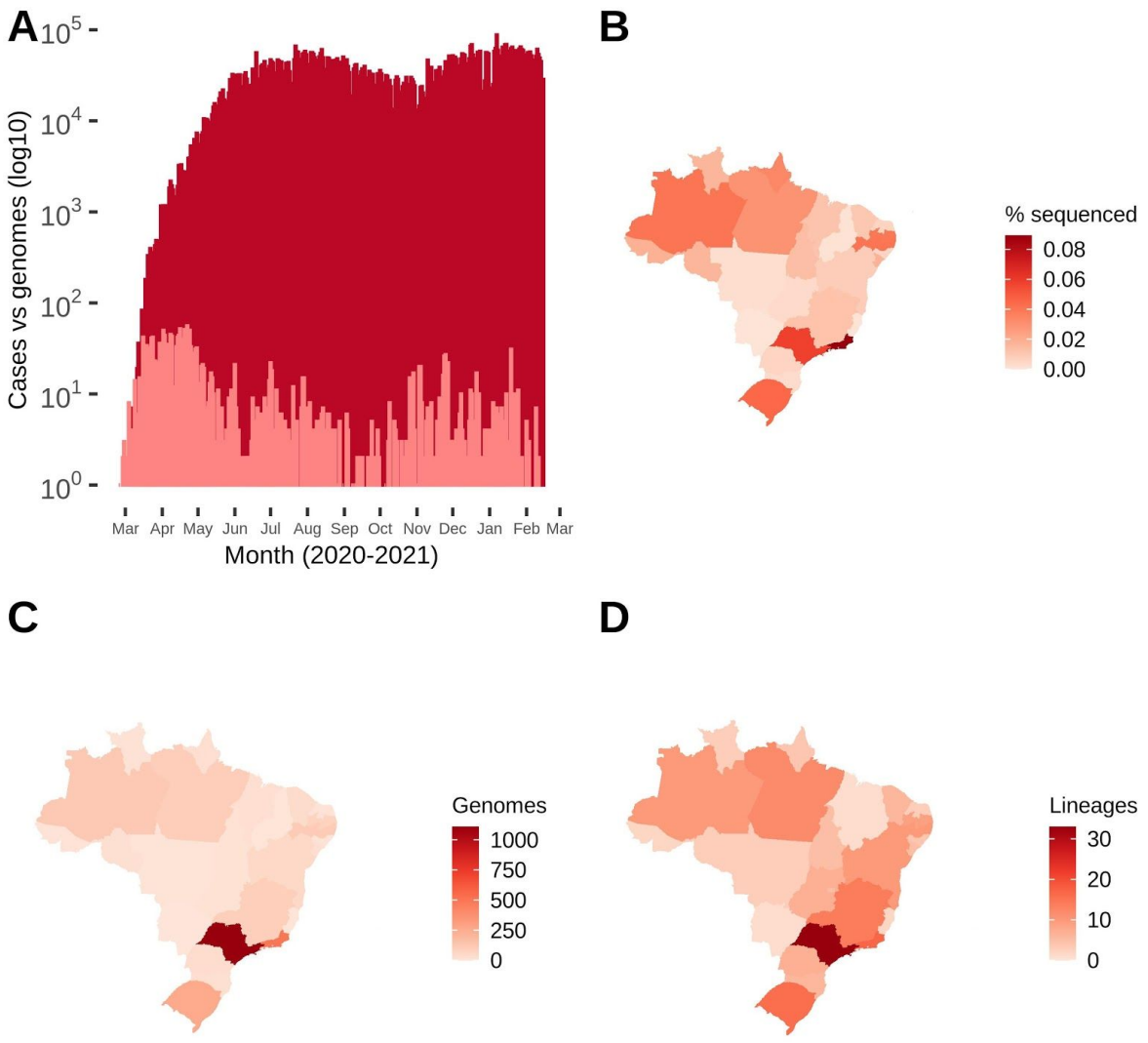

Figure 1. Distribution of Brazilian genomes through time (end-February 2020 to mid-February 2021) and space (Brazilian states). (A) Number of new cases (strong red) compared with number of collected genomes (light red) per day (log10 scale). (B) Fraction of genomes sequenced related to number of cases per Brazilian state. (C) Total number of genomes deposited per Brazilian state. (D) Total number of different SARS-CoV-2 lineages detected per Brazilian state.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

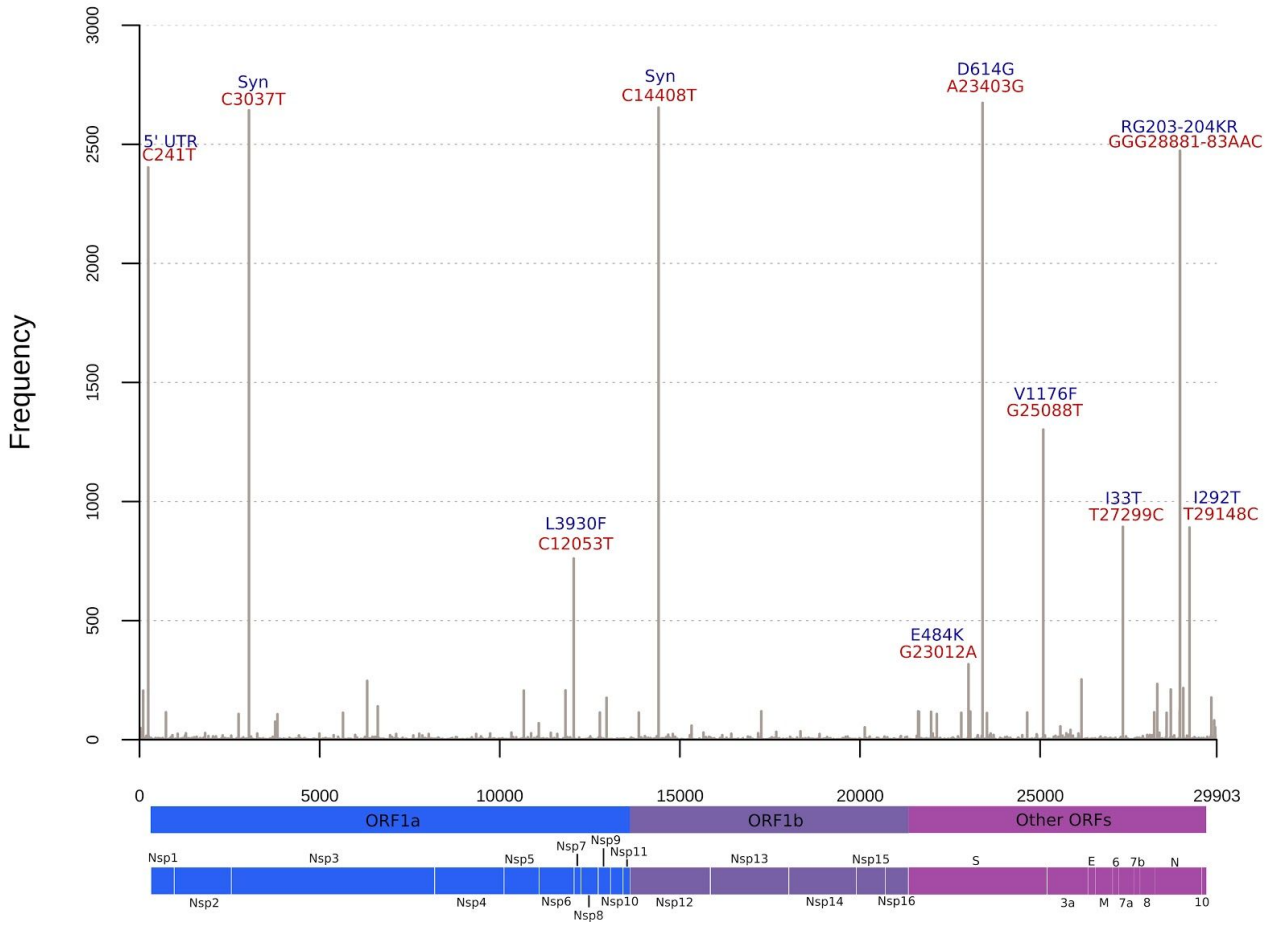

Figure 2. High frequent mutations across Brazilian sequences. Nucleotide replacements occurring in >250 genomes are indicated in red and associated amino acid substitutions in blue. Other mutations occurring in less than 250 genomes but more than 100 are indicated in Table 1.

Syn: Synonymous.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

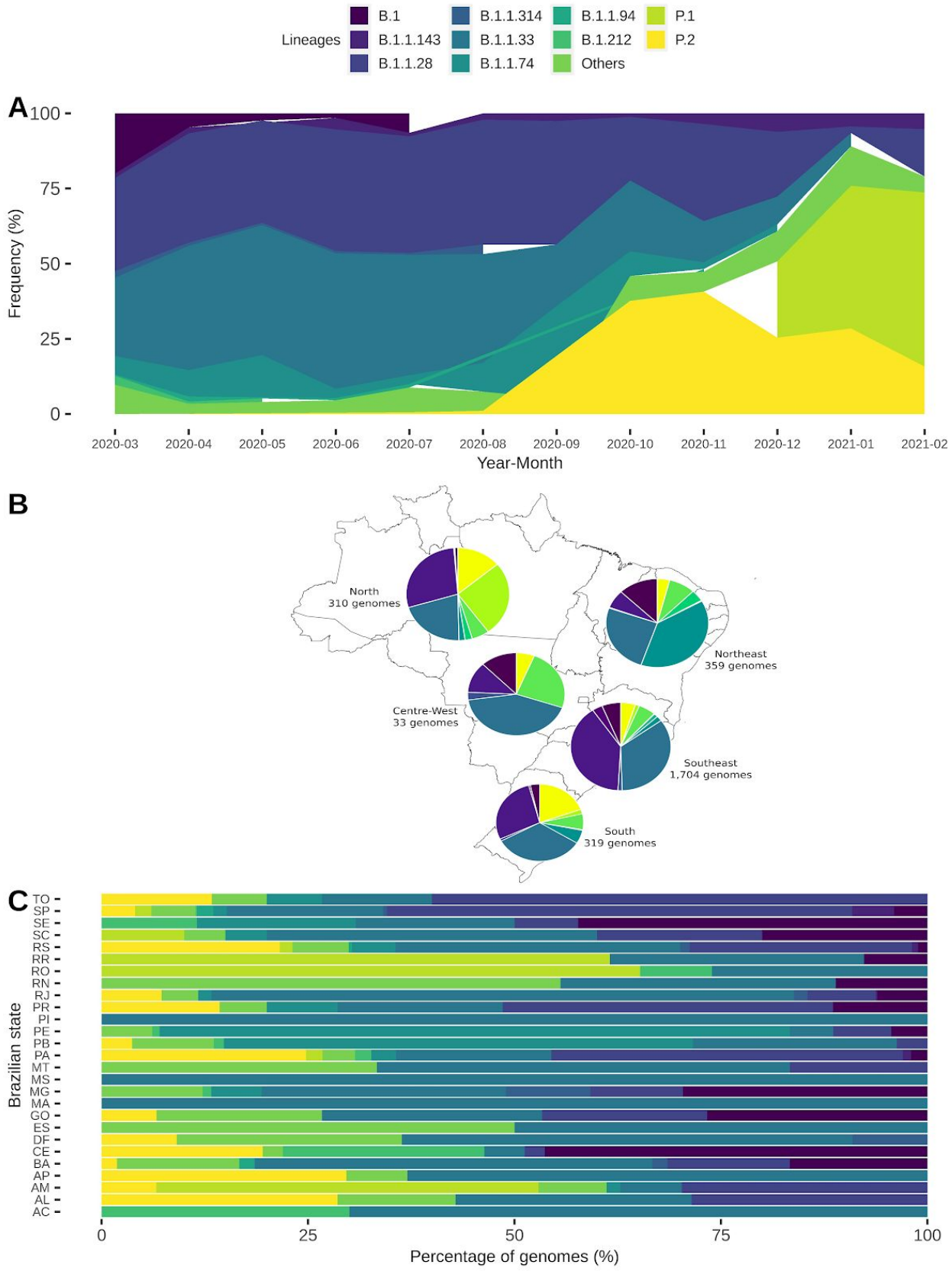

Figure 3. Distribution of the 10 most prevalent SARS-CoV-2 lineages (n>25) inside Brazil from end-February 2020 to mid-February 2021. (A) Frequency of these lineages through time in the entire Brazil. (B) Map showing the fraction of each of these lineages across all five Brazilian regions. (C) Distribution of these lineages across all Brazilian states, proportional to the number of

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

sequenced genomes.

AC: Acre; AL: Alagoas; AM: Amazonas; AP: Amapá; BA: Bahia; CE: Ceará; DF: Distrito Federal; ES: Espírito Santo; GO: Goiás; MA: Maranhão; MG: Minas Gerais; MS: Mato Grosso do Sul; MT: Mato Grosso; PA: Pará; PB: Paraíba; PE: Pernambuco; PI: Piauí; PR: Paraná; RJ: Rio de Janeiro; RN: Rio Grande do Norte; RO: Rondônia; Roraima: RR; RS: Rio Grande do Sul; SC: Santa Catarina; SE: Sergipe; SP: São Paulo; TO: Tocantins.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

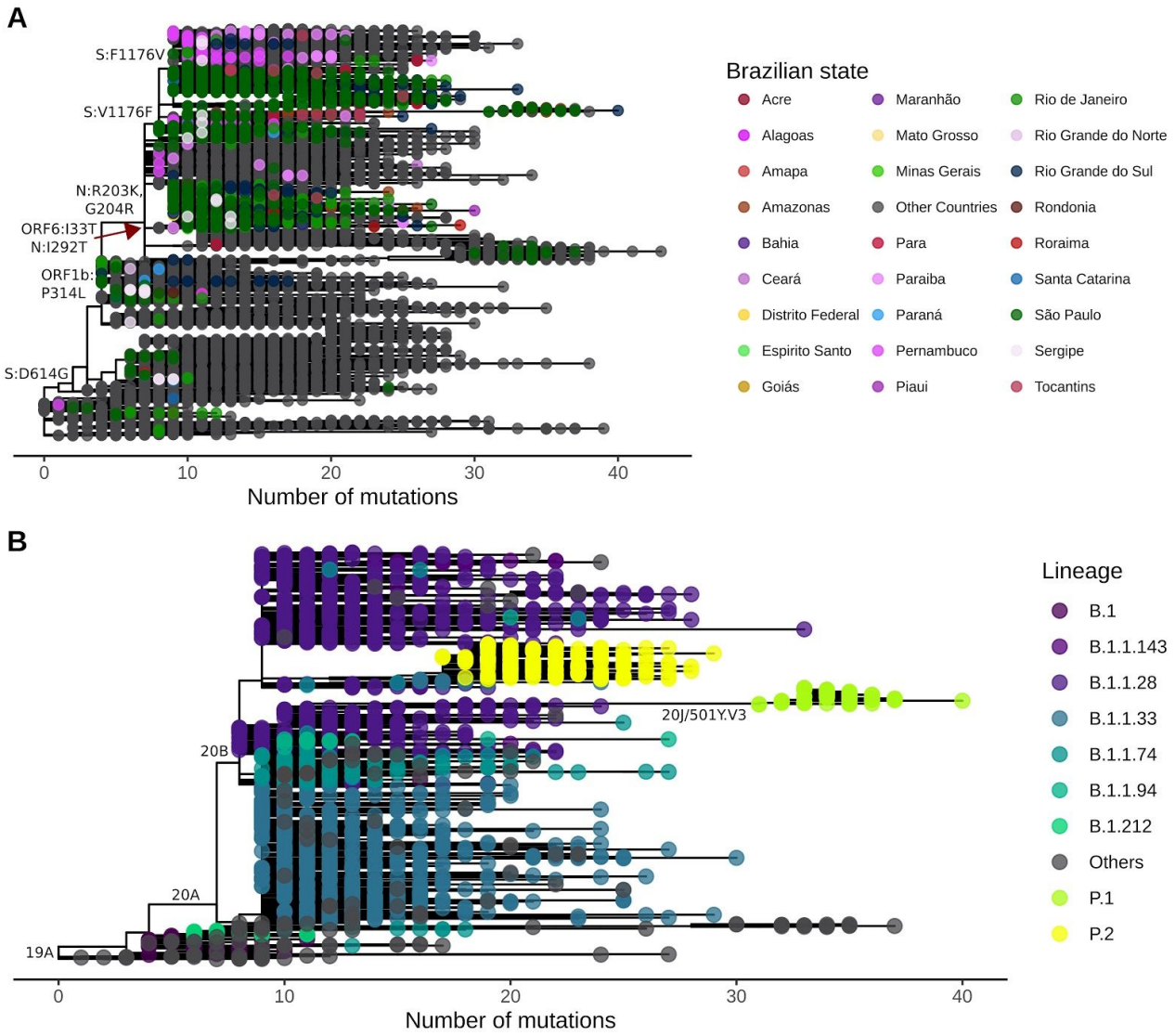

Figure 4. Evolutionary distribution of SARS-CoV-2 genomes from Brazil. (A) Maximum likelihood phylogenetics tree of 2,346 Brazilian sequences and 8,227 additional global representative genomes. States belonging to each specific Brazilian region are colored using similar colors (Centre-West: yellow; North: red; Northeast: purple and pink; South: blue; Southeast: green). Key mutations are represented in the respective branches. (B) Maximum likelihood phylogenetics tree dropping sequences from other countries and highlighting the most frequent Brazilian lineages presented in Figure 3. Nextstrain clades are represented in key branches.

Evolutionary rate is represented by the number of mutations (divergences) related to SARS-CoV-2 reference sequence (NC_045512.2).

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

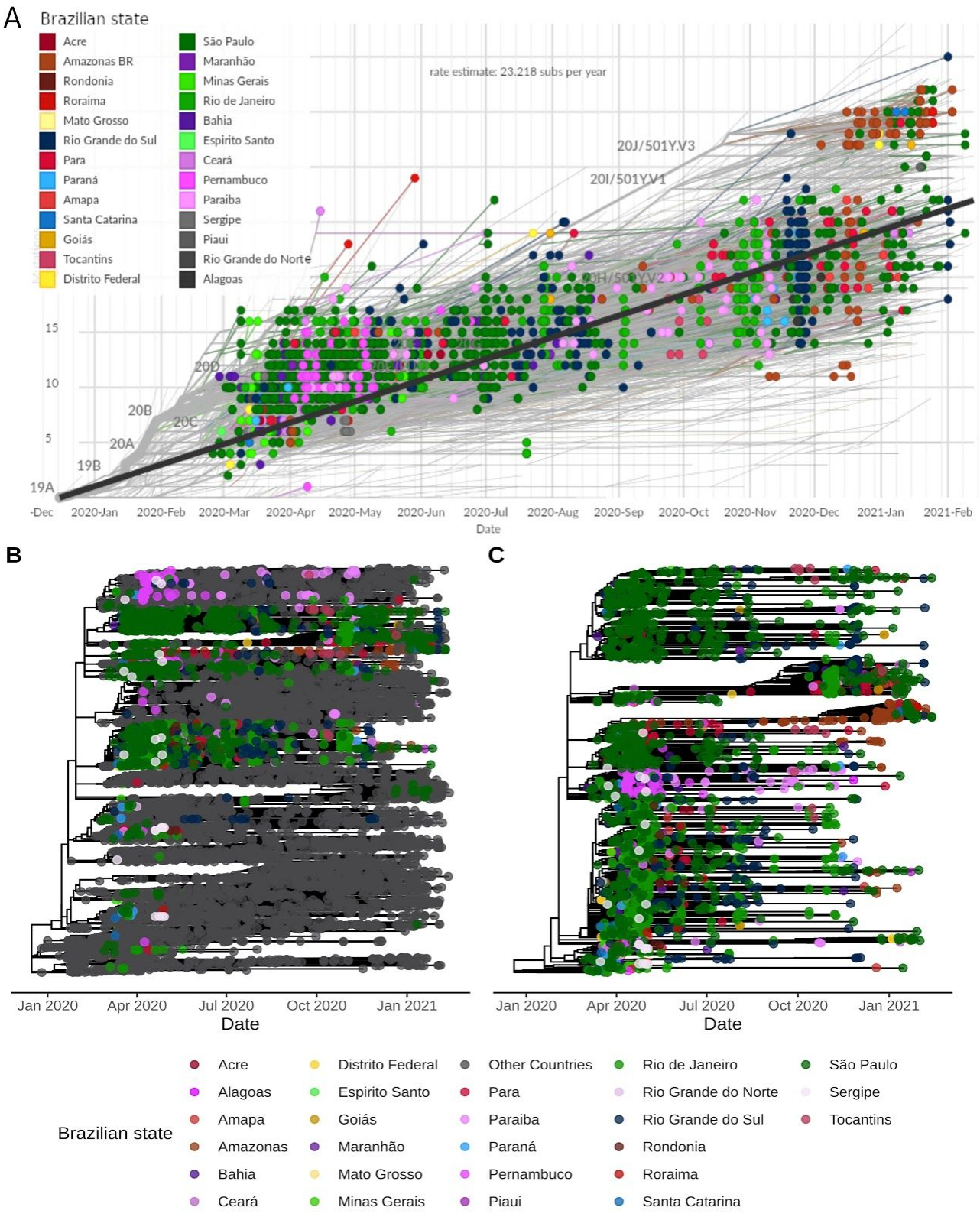

Figure 5. Molecular clock estimates of SARS-CoV-2 genomes from Brazil. (A) Root-to-tip regression of genetic distances (in number of mutations) against sampling dates filtered by Brazilian sequences. States belonging to each specific Brazilian region are colored using similar colors (Centre-West: yellow; North: red; Northeast: pink and gray; South: blue; Southeast: green). Nextstrain clades are represented in key branches. (B) Time-resolved Maximum Likelihood phylogenetics tree of the 10,573 genomes included colored by Brazilian states and other countries.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

(C) Maximum likelihood phylogenetics tree considering only the dynamics inside Brazil.

In (B) and (C), state colorings follow the same scheme as (A), except for Northeast, where purple and pink colors define sequences from this region. Tree topology remains the same for these two trees, but node ordering is slightly different.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.08.21253152; this version posted March 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

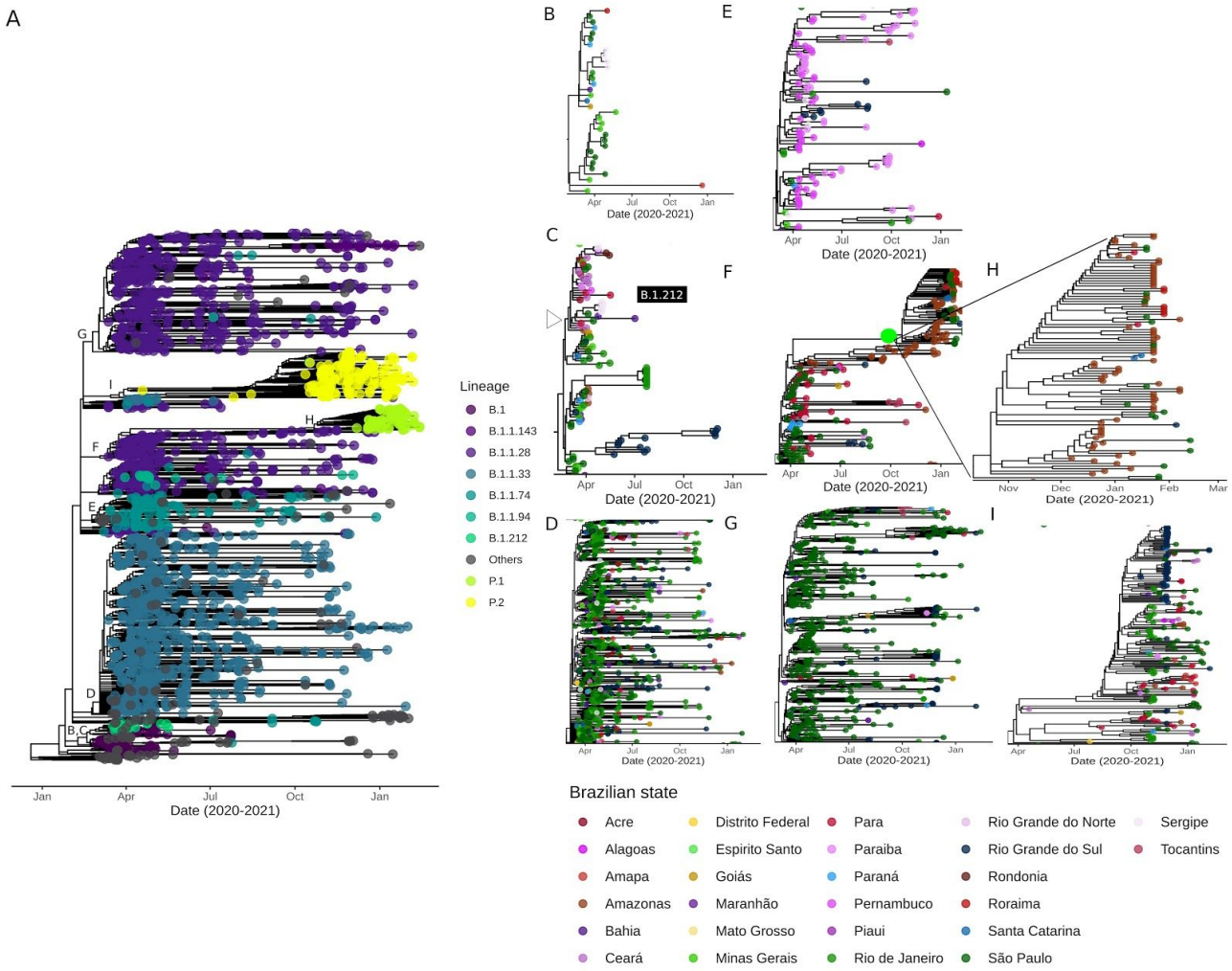

Figure 6. Zoom-in on clades corresponding to the major six Brazilian lineages. (A) Time-resolved ML tree of Brazilian sequences colored by PANGO lineages. Letters around clades are augmented in the respective figures and colored by Brazilian states. (B) Clade 1 is represented by sequences from the B.1 lineage. (C) Clade 2 has sequences from the B.1 and B.1.212 lineages (D) Clade 3 corresponds to the B.1.1.33 lineage. (E) Clade 4 is represented by B.1.1.74 genomes. (F) Clade 5 harbor sequences from B.1.1.28 and P.1. (G) Clade 6 has sequences from the B.1.1.28 lineage. (H) Zoom-in on P.1 sequences from Clade 5. (I) Clade 7 corresponds to the P.2 lineage.
From (B) to (I), states belonging to each specific Brazilian region are colored using similar colors (Centre-West: yellow; North: red; Northeast: purple and pink; South: blue; Southeast: green).