Predominance of the SARS-CoV-2 lineage P.1 and its sublineage P.1.2 in patients from the metropolitan region of Porto Alegre, Southern Brazil in March 2021: a phylogenomic analysis

Vinícius Bonetti Franceschi¹,#, Gabriel Dickin Caldana²,#, Christiano Perin³, Alexandre Horn³, Camila Peter⁴, Gabriela Bettella Cybis⁵, Patrícia Aline Grôhs Ferrareze², Liane Nanci Rotta², Flávio Adsuara Cadegiani⁶, Ricardo Ariel Zimerman³,* , Claudia Elizabeth Thompson¹,2,7,*

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 Laboratório Exame, Novo Hamburgo, RS, Brazil
⁵ Department of Statistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil
⁶ Corpometria Institute, Brasília, DF, Brazil
⁷ Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, RS, Brazil

# V.B. Franceschi and G.D. Caldana equally contributed to this work, the order of authorship being arbitrary.
* R.A. Zimerman and C.E. Thompson are both responsible for the work coordination, the order of senior authorship being arbitrary.

* Corresponding authors

Address for correspondence:
Ricardo A. Zimerman, MD - Medical coordination
Department of Infection Control and Prevention, Hospital da Brigada Militar, 244 Doutor Castro de Menezes, Porto Alegre, RS, Brazil. ZIP code: 91900-590. Phone: +55 (51) 3288 3500
E-mail: ricardoarielzimerman@gmail.com

Claudia E. Thompson, Ph.D - Scientific Coordination
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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Running title: Phylogenomic analysis of lineage P.1 and sublineage P.1.2 of SARS-CoV-2 in Southern Brazil

Keywords: SARS-CoV-2, COVID-19, Severe acute respiratory syndrome coronavirus 2, Infectious Diseases, High-Throughput Nucleotide Sequencing, Molecular Evolution, Molecular Epidemiology, Phylogeny
Abstract

Almost a year after the COVID-19 pandemic had begun, The United Kingdom, South Africa, and Brazil became the epicenter of new lineages, the Variant of Concern (VOCs), B.1.1.7, B.1.351, and P.1, respectively. These VOCs are increasingly associated with enhanced transmissibility, immunity evasion, and mortality. The previous most prevalent lineages in the state of Rio Grande do South (Brazil), B.1.1.28 and B.1.1.33 were rapidly replaced by P.1 and P.2, two B.1.1.28-derived lineages harboring the E484K mutation. To perform a genomic characterization of SARS-CoV-2 samples from COVID-19 patients from the metropolitan region of Porto Alegre (Rio Grande do Sul, Southern Brazil), in this second pandemic wave, we sequenced viral samples from patients of this region to: (i) identify the prevalence of SARS-CoV-2 lineages in the region, the state and bordering countries/states, (ii) characterize the mutation spectra, and (iii) hypothesize possible viral dispersal routes by using phylogenetic and phylogeographic approaches. As results, we not only confirmed that 96.4% of the samples belonged to the P.1 lineage but also that approximately 20% of which could be assigned as the newer P.1.2 (a P.1 derived new sublineage harboring new signature substitutions recently described and present in other Brazilian states and foreign countries). Moreover, P.1 sequences from this study were allocated in several distinct branches (four clades and five clusters) of the P.1 phylogeny, suggesting multiple introductions of P.1 in Rio Grande do Sul still in 2020 and placing this state as a potential core of diffusion and emergence of P.1-derived clades. It is still uncertain if the emergence of P.1.2 and other P.1 clades are related to further virological, clinical, or epidemiological consequences. However, the clear signs of viral molecular diversification from recently introduced P.1 warrant further genomic surveillance.
Introduction

After its initial emergence in Wuhan, China, in late 2019, the Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spread rapidly around the world leading to the COVID-19 pandemic officially recognized in March 2020 (1). As of May 17, 2021, >163 million cases and >3.3 million deaths were confirmed. In Brazil, the third country most affected by COVID-19, >15.6 million cases and >435.000 deaths were reported. From a virological standpoint, this could be related to the continental magnitude of Brazil, leading to multiple viral introductions (2) and the recent emergence of novel “Variant of Concern” (VOC) presenting enhanced infectiousness.

Rio Grande do Sul (RS) is the southernmost state in Brazil. It is bordered southerly by Uruguay, westerly by Argentina, and northerly by the state of Santa Catarina, Brazil. With an estimated population of 11.5 million inhabitants and 39.79 inhabitants per square kilometer, RS is the sixth most populous and the 13th most densely populated state in the country (3). Since 2017, the Brazilian Institute of Geography and Statistics (IBGE) has divided RS into eight intermediate geographic regions: Porto Alegre, Pelotas, Uruguaiana, Santa Maria, Santa Cruz do Sul / Lajeado, Ijuí, Passo Fundo, and Caxias do Sul (4). The municipality of Porto Alegre is the state capital, and its metropolitan region comprises 34 municipalities aggregating >4 million inhabitants (~2% of the country’s population) and is characterized by intense transit of people. COVID-19 was firstly confirmed in RS on March 10, 2020, in a returning traveler from Italy (5). The state implemented on May 2020 the “controlled distancing system”, which divided the state into 21 regions and 26 areas (R01 to R26) and consisted of a flag system establishing restrictions and flexibilizations of non-essential activities based on the weekly occupation of intensive care unit beds and expected deaths. However, due to economic losses, the more amenable “shared management system” allowed mayors to appeal in court and adopt less restrictive flag protocols (6).

Important shifts of COVID-19 epicenters have occurred during 2020, starting with Asia and followed by Europe, North America, and South America. After months of relatively slow evolution, novel VOCs (e. g., B.1.1.7, B.1.351, and P.1) harboring a constellation of signature mutations in the Spike protein have emerged. This phenomenon independently arose in the United Kingdom (7), South Africa (8), and Brazil (9,10) and has fueled secondary outbreaks in places where they have emerged, despite previous rates of seroprevalence of up to 75% (11). The city of Manaus
(Amazonas, Brazil), the probable place of origin of P.1 lineage, faced a major second wave of COVID-19. An explosive resurgence of cases and deaths became evident in mid-December 2020. Since the P.1 variant carries multiple mutations of potential biological significance (especially E484K, K417T, and N501Y in the Receptor-Binding Domain [RBD] from spike protein), (i) many key substitutions may lead to the immunity evasion, (ii) higher transmissibility when compared with pre-existing lineages have been characterized, (iii) this VOC has been the focus of increased surveillance and has deserved being studied in greater detail (12). After this outbreak, almost all Brazilian states experienced increases in the number of cases, hospitalizations, intensive care unit (ICU) admissions, and deaths, resulting in a reemergence of the public health crisis previously experienced in the first wave of COVID-19 (13).

The diversity of SARS-CoV-2 during the first epidemic wave in Brazil was mainly composed of B.1.1.28 and B.1.1.33 lineages (2,14), although the very low sequencing rate across the country has limited these estimates (14). However, these previous lineages were rapidly replaced by P.1 and P.2, both derived from the common ancestor B.1.1.28 and harbor concerning mutations in the Spike protein (e.g., E484K and N501Y), from late 2020 and early 2021 (14,15). In the RS state, the most common lineages identified to May 2021 still are B.1.1.33 (n=290) and B.1.1.28 (n=238). Nevertheless, P.1 has emerged as the most prevalent lineage sequenced in more recent samples (16). Recently, newer mutations were detected in addition to the original set presented in P.1, giving rise to the sublineage P.1.2 (17). P.2 probably emerged in the Rio de Janeiro state (Southeast) (18), but was also found in several municipalities of the RS state as of October 2020 (19,20). The first P.1 infection in the state was once thought to be in a patient of Gramado city in February 2021 (21). However, in a more recent study, the actual first P.1 was detected on November 30. This happened in a patient with comorbidities from Campo Bom city, who were reinfected by the P.2 lineage on March 11 (22).

Even though RS was one of the least affected Brazilian states in the first epidemic wave, it suffered a pronounced increase in cases in late 2020 (13). In February 2021, the progressive increases in cases and hospitalizations (3.8-fold) led to the collapse of the local state healthcare system. Since recent findings of the widespread dissemination of the SARS-CoV-2 lineage P.1 in Brazil have been confirmed, we sequenced samples from patients from the metropolitan region of
Porto Alegre to (i) identify the prevalence of SARS-CoV-2 lineages in the region, the state and bordering countries/states, (ii) characterize the mutation spectra, and (iii) hypothesize possible viral dispersal routes by using phylogenetic and phylogeographic approaches.
Methods

Ethics approval and consent to participate

Ethical approval was obtained from the Brazilian’s National Ethics Committee (Comissão Nacional de Ética em Pesquisa — CONEP) under process number CAAE 41909121.0.0000.5553 and Comitê de Ética em Pesquisa em Seres Humanos da Universidade Federal de Ciências da Saúde de Porto Alegre (CEP - UFCSPA) under process number CAAE 35083220.2.0000.5345. The study was performed in accordance with the Declaration of Helsinki. Patients included in the clinical trial approved by CONEP were informed in detail about the study and gave written informed consent to participate. All samples belonging to the Hospital da Brigada Militar patients that yielded positive RT-qPCR had their laboratory electronic records reviewed to compile metadata such as date of collection, sex, age, symptoms, exposure history, and clinical status, when available. Samples were anonymized before being received by the study investigators, following Brazilian and international ethical standards.

Sample collection and clinical testing

Samples were obtained from Hospital da Brigada Militar patients, both admitted or visiting the emergency ward, from Porto Alegre, RS, Brazil. Nasopharyngeal swabs were collected and placed in saline solution. Samples were transported to the clinical laboratory (Laboratório Exame) and tested on the same day for SARS-CoV-2 using Real Time Reverse-transcriptase Polymerase Chain Reaction (Charité RT-qPCR assays). The RTq-PCR assay used primers and probes recommended by the World Health Organization targeting the Nucleocapsid (N1 and N2) genes (23). Remnant samples were stored at -20°C.

Between March 9th to March 17th, all the routinely tested samples of the clinical laboratory provenient of the Hospital da Brigada Militar patients and yielded positive RT-qPCR were selected. Subsequently, those positive clinical samples were submitted to a second RT-qPCR performed by BiomeHub (Florianópolis, Santa Catarina, Brazil), using the same protocol (charite-berlin). Only samples with quantification cycle (Cq) below 30 for at least one primer were submitted to the SARS-CoV-2 genome sequencing. In total, 56 patients who presented symptoms such as fever,
cough, sore throat, dyspnea, anosmia, fatigue, diarrhea, and vomiting (moderate and severe clinical status) (24) were included in the study.

**RNA extraction, library preparation, and sequencing**

Total RNAs were prepared as in the reference protocol (25) using SuperScript IV (Invitrogen, Carlsbad, USA) for cDNA synthesis and Platinum Taq High Fidelity (Invitrogen, Carlsbad, USA) for specific viral amplicons. Subsequently, cDNA was used for the library preparation with Nextera Flex (Illumina, San Diego, USA) and quantified with Picogreen and Collibri Library Quantification Kit (Invitrogen, Carlsbad, USA). The sequencing was performed on the Illumina MiSeq (Illumina, San Diego, USA) 150x150 runs with 500xSARS-CoV-2 coverage (50-100 thousand reads/per sample).

**Quality control and consensus calling**

Quality control, reference mapping, and consensus calling were performed using an in-house pipeline developed by BiomeHub (Florianópolis, Santa Catarina, Brazil). Briefly, adapters were removed and reads were trimmed by size=150. Reads were mapped to the reference SARS-CoV-2 genome (GenBank accession number NC_045512.2) using Bowtie v2.4.2 (end-to-end and very-sensitive parameters) (26). Mapping coverage and depth were retrieved using samtools v1.11 (27) (minimum base quality per base (Q) ≥ 30). Consensus sequences were generated using bcftools mpileup (Q ≥ 30; depth (d) ≤ 1000) combined with bcftools filter (DP>50) and bcftools consensus v1.11 (28). Coverage values for each genome were plotted using the karyoploterR v1.12.4 R package (29). Finally, we assessed the consensus sequences quality using Nextclade v0.14.2 (https://clades.nextstrain.org/).

**Mutation analysis**

Single Nucleotide Polymorphisms (SNPs) and insertions/deletions in each sample were identified using snippy variant calling and core genome alignment pipeline v4.6.0 (https://github.com/tseemann/snippy), which uses FreeBayes v1.3.2 (30) to call variants and snpEff v5.0 (31) to annotate and predict their effects on genes and proteins. Genome map and SNP histogram were generated after running MAFFT v7.475 (32) alignment using the msastats.py script, and plotAlignment and plotSNPHist functions (33). Sequence positions refer to GenBank
RefSeq sequence (NC_045512.2), isolated and sequenced from an early case from Wuhan (China) in 2019.

We identified global virus lineages using the dynamic nomenclature implemented in Pangolin v2.3.8 (26; https://github.com/cov-lineages/pangolin) and global clades and mutations using Nextclade v0.14.2 (https://clades.nextstrain.org/). We also used the Pathogenwatch (https://pathogen.watch/) and Microreact (35) to explore mutations and lineages across time and geography initially.

**Maximum likelihood phylogenomic analysis**

All available SARS-CoV-2 genomes (1,048,519 sequences) were obtained from GISAID on April 26, 2021 and combined with our 56 sequences to obtain a global representative dataset. These sequences were subjected to analysis inside the NextStrain ncov pipeline (27; https://github.com/nextstrain/ncov). In this workflow, sequences were aligned using nextalign v0.1.6 (https://github.com/neherlab/nextalign). In the initial filtering step, short and low-quality sequences or those with incomplete sampling dates were excluded. Uninformative sites and ends (100 positions in the beginning, 50 in the end) were also masked from the alignment. Genetically closely related genomes to our focal subset were selected, prioritizing sequences geographically closer to Brazil's state RS. The maximum likelihood (ML) phylogenetic tree was built using IQ-TREE v2.1.2 (37), employing the General time-reversible (GTR) model with unequal rates and base frequencies (38). The tree’s root was placed between lineage A and B (Wuhan/WH01/2019 and Wuhan/Hu-1/2019 representatives), and sequences that deviate more than four interquartile ranges from the root-to-tip regression of genetic distances against sampling dates were excluded from the analysis.

A time-scaled ML tree was generated with TreeTime v0.8.1 (39) under a strict clock under a skyline coalescent prior with a mean rate of $8 \times 10^{-4}$ substitutions per site per year. Finally, clades and mutations were assigned and geographic movements inferred. The results were exported to JSON format to enable interactive visualization through Auspice. Additionally, as P.1 sequences mostly represent our dataset, we downloaded all complete and high-quality global genomes assigned to P.1 PANGO lineage (4,499 sequences) submitted until April 26th, 2021. These
sequences were aligned using MAFFT v7.475, the ends of the alignment (300 in the beginning and 500 in the end) were masked, and the ML tree was built with IQ-TREE v2.0.3 using the GTR+F+R3 nucleotide substitution model as selected by the ModelFinder (40). Branch support was calculated using the Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT) (41) with 1,000 replicates.

Local sequences were classified according to the following scheme: monophyletic clades composed by one local genome were classified as “isolated”, while clades composed by 2 < genomes < 4 were considered “clusters” and if ≥ 4 local genomes represented, we assigned a “clade” designation.

ML trees were inspected in TempEst v1.5.3 (42) to investigate the temporal signal through regression of root-to-tip genetic divergence against sampling dates. For the P.1 ML tree, samples with missing days of the collection were filled with the 15th day of the month. ML and time-resolved trees were visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) and ggtree R package v2.0.4 (43).

**Discrete Bayesian phylogeographic and phylodynamic analysis**

Considering the four identified clades composed by ≥ 4 sequences from this study, we extracted the clade members using the caper R package v1.0.1 (44). Clade-specific ML trees and root-to-tip regression assignments were generated as described above. Evolutionary parameter estimates and spatial diffusion were estimated separately for each clade using a Bayesian Markov Chain Monte Carlo (MCMC) approach implemented in BEAST v10.4 (45) using the BEAGLE library (46) to enhance computational time. Time-stamped Bayesian trees were generated using the HKY+Γ nucleotide model (47), a strict molecular clock model with a Continuous Time Markov Chain (CTMC) rate reference prior (48) (mean rate = 8×10^{-4}) and a non-parametric skygrid tree prior (49) with grid points defined by the approximate number of weeks spanned by the duration of the phylogeny.

The MCMC chains were run in duplicates for at least 50 million generations, and convergence was checked using Tracer v1.7.1 (50). Log and tree files were combined using
LogCombiner v1.10.4 to ensure stationarity and good mixing (43) after removing 10% as burn-in. Maximum clade credibility (MCC) was generated using TreeAnnotator v1.10.4 (45). Viral migrations were reconstructed using a reversible discrete asymmetric phylogeographic model (51) to infer the locations of the internal nodes of the tree. A discretization scheme with a resolution of different Brazilian states and other countries was applied. Location diffusion rates were estimated using the Bayesian stochastic search variable selection (BSSVS) (51) procedure employing Bayes factors to identify well-supported rates.

**Geoplotting**

Geographical maps and other plots were generated using R v3.6.1 (52), and the ggplot2 v3.3.2 (53), geobr v.1.4 (54), and sf v0.9.8 (55) packages. For the discrete phylogeographic analysis, the SpreaD3 v0.9.7.1 (56) was used to map spatiotemporal information embedded in MCC trees.
Results

Epidemiological information

From the 56 samples of hospitalized patients between March 09-17 2021, 75.0% \( (n=42) \) of them were male, and the mean age was 37.2 years (interquartile range (IQR): 13.5 years). The mean cycle threshold (Ct) value for the first RT-qPCR conducted at Laboratório Exame was 19.12 cycles (median: 18.00, IQR: 6.00 cycles). Forty-seven (83.9%) had contact with a confirmed or suspected case. The majority of them were from the RS state capital, Porto Alegre \( (n=32; 57.1\%) \). In total, 51 (91.1%) were from the intermediate geographic region of Porto Alegre and 5 (8.9%) from the intermediate region of Santa Maria (Table 1, Figure S2C).

SARS-CoV-2 mutations and lineages

Consensus SARS-CoV-2 genomes were obtained with an average coverage depth of 813.2× (median: 820.6×, IQR: 184.7×) (Supplementary File 1). We detected 175 different mutations comprising all samples (Figure 1A). The ORF1ab carried 102 (58.3%) replacements followed by spike \( (n=24; 13.7\%) \), nucleocapsid \( (n=18; 10.3\%) \) ORF3a \( (n=14, 8.0\%) \), ORF7a \( (n=6; 3.43\%) \), ORF8 \( (n=5; 2.86\%) \), and membrane \( (n=3; 1.7\%) \) genes. Remarkably, 50% of the spike substitutions occurred in only one genome, and of these, 9 (75.0%) were missense (Supplementary file 2). Fifty-nine (33.7%) mutations were identified in two or more sequences. From these, 36 (61.0%) are non-synonymous (missense), 21 (35.6%) are synonymous, one (1.7%) intergenic at 5' Untranslated Region (UTR), and one (1.7%) inframe deletion. Highly frequent (≥ 10 genomes) mutations were found in 34 genomic positions, 24 (70.5%) being missense and 9 (26.5%) synonymous. Fifteen substitutions (10 in the spike protein: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I) are P.1 lineage-defining mutations (Figure 1B, Table 2). The only P.1 defining replacement not found at high frequency in our study was the deletion in ORF1ab (del:11288:9), called in only four genomes. This result is due to the stringent coverage depth filter applied (DP>50) for calling the genomic positions in the consensus sequences.

After comparing the frequency of mutations from the recently sequenced samples and the Brazilian P.1 genomes, we observed a combination of mutations that stood out in a significant
proportion (n=11; 19.6%) compared with previous P.1 sequences from Brazil. This combination was previously described (17) and gave rise to the P.1.2 lineage, which harbors three ORF1ab replacements (synC1912T, D762G, T1820I), one in ORF3a (D155Y), and one in N protein (synC28789T) (Table 2). Additionally, two of these genomes (18.2%) carry T11296G (ORF1ab nsp6: F3677L) and 8 (72.7%) harbor G25641T (ORF3a: L83F) substitutions. Another cluster, made of four genomes and subsequently named Clade 2, possesses three defining mutations (ORF1ab nsp4: V2862L, synC10507T, ORF3a: M260K) was also detected. This cluster does not fall into a lineage designation at this moment but deserves further monitoring (Table 2, Figure S1).

Considering PANGO lineages, 54 genomes (96.4%) were designated as P.1, one (1.8%) as P.2, and one (1.8%) as B.1.1.28. Even without being classified according to the Pango-designation's most updated version, the P.1.2 lineage was present in 11/54 (20.4%) of the P.1 sequences (https://github.com/cov-lineages/pango-designation/issues/56) (Figure S1).

Lineage distribution in neighboring countries and Brazilian regions

The RS state shares borders with Argentina in its west (Figure S2), leading to the transit of people at the frontiers. From March to April 2020, B.1 was the most prevalent lineage in this bordering country. B.1.499 and N.3 were abundant from May to July when the N.5 started to rise and surpassed B.1.499 in November 2020 (Figure 2A). Importantly, N.3 and N.5 are derived from the B.1.1.33 lineage widespread in Brazil (https://cov-lineages.org/lineages.html). The P.2 lineage, which initially emerged in Brazil (18) and derived from another Brazilian disseminated lineage (B.1.1.28), was firstly found in November 2020, and by January and February 2021 it had already outnumbered the other lineages in Argentina (Figure 2A).

In the entire Brazil, despite early introductions of B.1 and B.1.1, lineages B.1.1.28 and B.1.1.33 were most abundant from March to October 2020. In October, P.2 already represented an important portion of the sequences, and by November it had already surpassed B.1.1.33. In December 2020 and January 2021, with the emergence of P.1, this lineage and P.2 already became the most prevalent, while between February and April P.1 replaced all other lineages (Figure 2A and 2B). Some fluctuations evidently occurred in different Brazilian regions, such as a prevalence of more local lineages (e. g. B.1.195 and B.1.1.378 in the Northern region, B.1.1 and
N.9 in the Northeast, and B.1.1.7 in the Southeast and Centre-West (Figure S3). In RS, a similar landscape was observed compared to the Brazilian scenario. B.1.1.28 and B.1.1.33 were most prevalent until October 2020, when P.2 emerged and remained until January 2021 along with B.1.1.28 as the most prevalent. After the introduction of P.1 (January 2021), this lineage practically supplanted the others in February and March 2021 (Figure 2A and 2B).

After dividing RS into the intermediate regions proposed by IBGE (Figure S2), it was possible to gain insights into the dynamics of the lineages in the state, despite the low sample size of some regions (Figure 2C). In most regions, the lineages B.1.1.28 and B.1.1.33 were more prevalent, but P.2 was also detected. In fact, in the Caxias do Sul region, more P.2 (n=31; 49.2%) were sequenced in relation to other lineages. Since the Porto Alegre region has a larger sample size, we divided its results by year to check the most recent (2021) evolutionary abundance. In 2020, B.1.1.33 (n=229; 49.3%) and B.1.1.28 (n=137, 29.5%) were the most abundant, followed by P.2 (n=37; 8.0%). In 2021, P.1 (n=26, 68.4%) and P.2 (n=6, 15.8%) have already outperformed the other lineages (Figure 2C). In our study from March 2021, 96.4% of the samples were classified as P.1. We were able to identify a new P.1 sublineage (P.1.2) in 11 (20.4%) genomes from four different municipalities (Porto Alegre, Canoas, São Sebastião do Caí, and Santa Maria), demonstrating the possible diversification of P.1 and its spread within RS (Figure 2C, Figure S1).

**Maximum likelihood phylogenomic analysis**

After running the Nextstrain workflow using quality control and subsampling approaches, we obtained a dataset of 8,635 time- and geographical-representative genomes. From these, 861 were from Africa, 1,370 from Asia, 2,219 from Europe, 481 from North America, 218 from Oceania, and 3,486 from South America. Brazil was represented by 2,608 sequences and the RS state by 730 sequences (56 from this study and 674 available in GISAID) (Supplementary file 3).

The time-resolved ML phylogenetic tree confirmed the PANGO lineages assigned since 54 genomes (96.4%) grouped with P.1 representatives, one (1.8%) with B.1.1.28, and one (1.8%) with P.2 sequences. We also observed a strong correlation between genetic distances and sampling dates (R² = 0.71). The P.1 sequences were grouped above the regression line, showing higher evolutionary rates than the other lineages in the SARS-CoV-2 phylogeny as observed in other
studies (10,11). We highlighted the most abundant global lineages present in the RS state that passed the quality control criteria (B.1.1 [n=32], P.1 [n=83], P.2 [n=83], B.1.1.28 [n=203], and B.1.1.33 [n=286]). We also noticed the high abundance of B.1.1.28 and B.1.1.33 lineages until October-November 2020, followed by the rise and establishment of P.2 and P.1, respectively (Figure 3A).

To get a more detailed understanding of the P.1 diffusion throughout Rio Grande do Sul, other Brazilian regions, and worldwide countries, we built a ML tree of 4,499 genomes belonging to this lineage (Supplementary file 4). P.1 sequences from this study were allocated in several distinct branches, suggesting multiple introductions and the formation of different P.1-derived clades and clusters.

We identified four clades, five clusters, and 13 isolated sequences (Figure 3B, Supplementary file 5). Most importantly, clade 1 was composed of 11 sequences originated in this study that shared five lineage-defining mutations as previously described (Table 2) and were recently attributed to the P.1.2 sublineage (https://github.com/cov-lineages/pango-designation/issues/56). As of April 26, 2021, this sublineage is already distributed worldwide in 93 sequences (the Netherlands, Spain, England, and the USA) and in other Brazilian states (Rio de Janeiro [RJ] and São Paulo [SP]) (17). Clade 2 sequences harbored two mutations in the ORF1ab:V2862L (nsp4) and synC10507T, one in the ORF3a:M260K, and comprised 81 genomes. Four samples are from this study. The majority is from Amazonas (n=15), São Paulo (n=11), RS (n=8), Bahia (BA) (n=4), and worldwide sequences are mainly represented by French Guiana, USA, Spain, Japan, and Jordan. Clade 3 is represented by three ORF1ab mutations (synC1420T, D1600N [nsp3], and synT8392A) in 3 of 7 local genomes. It is composed of sequences from RS (n=25), SP (n=15), Maranhão (n=10), RJ (n=8), as well as other countries (mainly Spain, French Guiana, and the USA). Clade 4 is characterized by two ORF1ab substitutions (G400S [nsp2] and S6822I [2’-O-ribose methyltransferase]), one N:synT26861C in three genomes and carries other additional mutations (ORF1ab: synG10096A, G3676S [nsp6], F3677L [nsp6]) and M:synT26861C. This clade is mainly found in SP (n=11), RS (n=7), Santa Catarina (n=5), BA (n=4), Goiás (n=4), and other countries (mainly USA, Chile, and England).

Clusters 1 and 3 have respectively one (ORF1ab: G3676S [nsp6]) and two (ORF1ab:
synC1471T, A1049V [nsp3]) shared mutations. Among all identified clusters, the most diverse was cluster 5, which contains three samples from this study and has five defining mutations: four in ORF1ab (synT4705C, synC11095T, syn11518, T5541I [helicase]) and one in ORF7a: E16D. Moreover, two sequences share one distinct mutation (ORF1ab: F3677L [nsp6]).

**Bayesian molecular clock and phylogeographic analysis**

To date, the time of the most recent common ancestor (TMRCA) and the diffusion of the four P.1 clades identified in our ML analysis, we used coalescent and phylogeographic methods. For clade 1, which is correspondent to the recently labeled P.1.2 lineage, sequences showed a strong correlation of genetic distances and sampling dates (correlation coefficient: 0.59, \( R^2 = 0.34 \)) (Figure 4A). We estimated a median evolutionary rate of \( 7.68 \times 10^{-4} \) (95% Highest Posterior Density interval [HPD]: \( 4.18 \times 10^{-4} \) to \( 1.14 \times 10^{-3} \) subst/site/year) and the TMRCA on December 18, 2020 (95% HPD: October 29, 2020 to January 31, 2021). Interestingly, the tree’s root was placed in RS, between a sequence from RS (the oldest sequence from this clade: EPI_ISL_983865) and a subclade from the USA. The divergence of these subclades was dated on January 15, 2021 (95% HPD: January 15 to March 26, 2021). The subclade composed by the RS sequences formed two separate clusters, one with three sequences from this study and one Australian genome and another composed of sequences from RS, SP, UK, Portugal, USA, and transmission clusters from RJ and Netherlands (Figure 4B). The emergence of an important cluster in RJ carrying additional mutations (17) was dated here on March 11, 2021 (95% HPD: March 11 to April 6, 2021). As USA sequences formed a separate subclade, local transmission is probably occurring in the country. The divergence of the USA subclade was dated February 7, 2021 (95% HPD: February 1 to May 11, 2021). Although the BSSVS procedure identified well-supported rates of diffusion from Rio Grande do Sul to other Brazilian states such as São Paulo (Bayes Factor [BF]: 6.82, posterior probability [PP]: 0.52), Rio de Janeiro (BF: 39.18, PP: 0.86), and other countries such as USA (BF: 31.94, PP: 0.84) and Netherlands (BF: 80.16; PP: 0.93), it is possible that this lineage emerged in another Brazilian state but its earlier representatives were not sampled. This is a strong hypothesis since this sequence is associated with community transmission after contact with tourists in a city of RS (Gramado) that receives numerous visitors annually (21).
For clade 2, we estimated a median evolutionary rate of \(5.0 \times 10^{-4}\) (95% HPD: \(4.18 \times 10^{-4}\) to \(7.71 \times 10^{-4}\) subst/site/year), and the TMRCA was dated November 30, 2020 (95% HPD: November 2 to December 21, 2020). This clade includes sequences from 11 Brazilian states from all five regions and 9 other countries. We were able to detect at least five introductions from Amazonas, where this clade probably emerged. These introductions ranged from December 28, 2020 (95% HPD: December 28, 2020 to January 5, 2021) to January 28, 2021 (95% HPD: January 28 to March 7, 2021). Importantly, we identified a well-supported subclade (PP = 1) of four genomes from this study (Figure 5A).

For clade 3, the TMRCA was estimated on December 20, 2020 (95% HPD: November 25 to December 29, 2020) and the median evolutionary rate was \(7.85 \times 10^{-4}\) (95% HPD: \(6.06 \times 10^{-4}\) to \(1.02 \times 10^{-3}\) subst/site/year). This clade harbors sequences from nine Brazilian states and 10 other countries. Amazonas is the most probable source of its emergence. From then onwards, multiple transmission clusters were established in foreign countries (e.g. Spain, Portugal, USA) and Brazilian states (especially Maranhão, SP, and RS). This clade was introduced at least five times in RS, leading to two major subclades represented by 18 and four sequences, respectively. The major subclade (n=18, PP = 0.98) was dated January 11, 2021 (95% HPD: January 11 to February 1, 2021) (Figure 5B).

For clade 4, the TMRCA was dated on December 02, 2020 (95% HPD: October 7, 2020 to January 3, 2021), and the median evolutionary rate was \(6.26 \times 10^{-4}\) (95% HPD: \(3.51 \times 10^{-4}\) to \(1.01 \times 10^{-3}\)). This clade comprises nine Brazilian states and five foreign countries. After its initial emergence and spread in Amazonas, it had already formed transmission clusters in SP, BA, United Kingdom, and the USA. Most important, a subclade containing sequences from two neighboring states from Southern Brazil (seven from RS and five from Santa Catarina [SC]) indicates its diffusion from RS to SC, probably leading to two separate introductions. The divergence of this subclade was estimated on December 16, 2020 (95% HPD: December 16, 2020 to January 19, 2021) (Figure 5C).

Phylogenetic and molecular clock approaches suggest the wide circulation of the VOC P.1 both nationally and internationally between late 2020 and early 2021. This lineage has already diversified into some clades that bear characteristic mutations, although they exhibit similar
evolutionary rates. We have inferred that P.1 (and its derived clades) was introduced multiple times in the southernmost Brazilian state (RS) still in 2020, probably in December. Remarkably, this date is close to the first P.1 detection in Manaus, which is located ~4 thousand kilometers away. These early introductions led to the formation of local subclades that could be identified even using a reduced set of sequenced samples.

Discussion

In this study, the analysis of 56 samples from the state of Rio Grande do Sul (RS), Southern Brazil, confirmed that the P.1 lineage was already highly prevalent. Interestingly, we demonstrated that P.1 is already showing signs of diversification and has originated a new sublineage (P.1.2) by indicating the likely origin and the first clusters of this novel lineage. This sublineage was detected in three Brazilian states, and other countries, and its most recent common ancestor was dated on December 18th, 2020 (95% HPD: October 29th, 2020 to January 31th, 2021). In accordance with the majority of the states from Brazil, this state experienced significant increases in hospitalizations in early 2021. This scenario was related to the emergence and rapid spread of the P.1 variant across the country.

In most RNA viruses, the accuracy of RNA replication is low, leading to the emergence of frequent nucleotide substitutions. However, since its very long RNA genomic strand requires higher fidelity replication in order to keep up genome integrity, coronaviruses behave differently in this regard. In SARS-CoV-2, for example, a proofreading mechanism increases 100 to 1,000 fold the fidelity of RNA synthesis through the activity of an exonuclease present in NSP14. This enzyme corrects nucleotide misincorporation during RNA duplication by the error-prone RNA-dependent RNA polymerase (57–59). As a result, coronaviruses have lower rates of mutation than other RNA viruses that lack proofreading activity (58).

After almost one year of relatively slow SARS-CoV-2 evolution, the emergence of multiple and convergent lineages harboring a constellation of mutations in the spike protein raised concern in the scientific community. This protein is present on the viral surface of SARS-CoV-2 and is composed of two subunits, S1 and S2. S1 contains the receptor-binding domain (RBD), responsible for virus attachment at the cell surface, the N-Terminal Domain (NTD), also critical for
binding properties and the more conserved CTS1. S2 is an alpha-helix-rich subunit that contains HR1 and HR2, critical for viral-cell membrane fusion (58). Spike protein, therefore, is responsible for mediating interaction with the human Angiotensin-Converting Enzyme 2 receptor (hACE2) and is a primary target of neutralizing antibodies and vaccine development (60). The variants harboring different mutational signatures, including spike protein substitutions, were classified as VOCs and “Variant of interest” (VOI), depending on their growing relevance in the current pandemic. The first three VOC emerged in England (B.1.1.7) (7), South Africa (B.1.351) (8), and Brazil (P.1) (9). More recently, B.1.427/429 (California) also was characterized as VOC (61). As far as May 2020, B.1.526 (New York), B.1.617 (India), and P.2 (Brazil) (18) were still categorized as VOIs. B.1.1.7, B.1.351 and P.1, the most studied VOCs, have the D614G and N501Y mutations in common. B.1.351 and P.1 share a mutation in the K417 site (K417N and K417T, respectively) and the E484K replacement, which is also observed in the P.2 lineage. Additionally, B.1.1.7 carries the P681H substitution in the furin-cleavage site and multiple VOIs bear the L452R substitution (62). The presence of common substitutions in different SARS-CoV-2 lineages suggests co-evolutionary and convergent mutational processes (7–9,63).

D614G is a substitution of aspartic acid to glycine at amino acid position 614 at the CTS1 segment of the S1 subunit. It emerged at the end of January 2020, first reported in April 2020, and quickly became prevailing in several B.1-derived lineages, including most VOCs and VOIs described thus far (64–67). This mutation is interesting, since it does not occur near RBD and thus does not directly modify the binding affinity for hACE-2. Instead, it disrupts important hydrogen bonds with neighbor S2, resulting in altered interprotomeric configurations. As a consequence, the active “one RBD up and two RBD down” is favored. This allows binding to ACE2 more effectively, leading to higher replication rates and viral loads (65). However, G614 mutants are similarly (or even more) susceptible to immune neutralization than the original D614 variant (65,68).

The substitution from asparagine to a tyrosine at position 501 (N501Y) has first appeared in the UK in September 2020. It is one of the six key contact residues within RBD interacting with ACE2 and has been associated with increased binding affinity to human and murine ACE2 due to the formation of an extra hydrogen bond with the ACE2 receptor (69–71). Studies suggest that this mutation could enhance SARS-CoV-2 transmissibility and mortality (72–74).
The substitution of glutamic acid for lysine at the 484 RBD position (E484K) creates a strong ion interaction with the amino acid 75 of ACE2 due to the electrostatic change from a negatively charged to a positively charged amino acid (75). E484K can also lead to immune evasion since it is located at a flexible loop previously irrelevant for receptor binding when the original glutamic acid is in place. This may explain the enhanced dissemination and improved infectivity and dissemination of SARS-CoV-2 (76,77). In fact, as of May 17th 2021, 66,405 sequences with the E484K mutation have been detected (62). Recent reinfection cases of E484K-containing SARS-CoV-2 (78,79) and the recent proof of its fixation in different lineages (80) are suggestive that this mutation must be investigated in more detail.

Despite the residue 452 does not directly contact the ACE2 receptor, L452 with the residues F490 and L492 form a hydrophobic patch on the surface of the spike RBD. This stabilizes the interaction between the spike protein and the ACE2 receptor and promotes an increased virus entry into the cell Moreover, lineages carrying this mutation seem to present a moderate resistance to neutralization by antibodies elicited by prior infection or vaccination (61). The P681H mutation, in turn, has not its significance completely elucidated. The adjacent position to the furin cleavage site (five sites upstream of the arginine residues), which is needed for the cell membrane fusion, can potentially affect the viral infectivity. Furthermore, the independent appearance in different lineages suggests convergent evolution and a possible adaptive advantage since the acquisition of a multibasic S1/S2 cleavage site was essential for SARS-CoV-2 infection in humans (81,82).

In the present study, we noticed that B.1.1.33 and B.1.1.28 lineages, detected at the beginning of the pandemic in Brazil (2), had been similarly prevalent in different regions until September 2020, before the appearance of P.2 (in October) and P.1 (in December 2020). The B.1.1.33 lineage shows variable abundance in different Brazilian states (ranging from 2% in Pernambuco to 80% in Rio de Janeiro), with moderate prevalence in South American countries (5-18%). Surprisingly, this lineage was firstly detected in early March 2020 in other American countries (e.g. Argentina and the USA). Apparently, an intermediate strain probably emerged in Europe and subsequently spread to Brazil, where its spread gave rise to B.1.1.33 (83) and possibly triggered secondary outbreaks in Argentina and Uruguay (83,84). We had found that N.3 and N.5, both derived from B.1.1.33, represented an important proportion of the sequences from Argentina.
from May to December 2020, when it was replaced by the P.2 lineage, which probably emerged in Rio de Janeiro (Southeastern Brazil). The B.1.1.28 lineage, despite apparently less abundant than B.1.1.33 in several Brazilian regions, quickly diversified into two variants: VOC P.1 and VOI P.2 (85). Since the end of 2020, these two lineages lead the diversity of SARS-CoV-2 in Brazil (14) and have caused concern in other countries after several introductions.

The emergence of a B.1.1.28 derived lineage carrying the S:E484K mutation (P.2) was dated, in a retrospective study, late February 2020 in the Southeast (São Paulo and Rio de Janeiro), followed by transmission to the South (especially RS). Since then, multiple dispersion routes were observed between Brazilian states, especially in late 2020 and early 2021 (15). However, this lineage went unreported until October 2020, when it was first detected in the state of Rio de Janeiro (18) and in the small municipality of Esteio in RS (19). The increased frequency of B.1.1.28 and derived lineages was corroborated by another study that included samples from several municipalities of RS in November 2020. This study found that 86% of the genomes could be classified as B.1.1.28 and ~50% of these, in fact, belong to the new lineage P.2 (20). Nonetheless, our current study suggests that P.2 has already been nearly entirely replaced by the P.1 lineage or is not particularly well represented among the analyzed patients seeking emergency consultation or requiring hospitalization.

Between June and October 2020, an extremely high seroprevalence (44-76%) was observed in Manaus (Amazonas, Brazil) in a study from blood donors (11). However, despite these numbers, Manaus faced a resurgence of cases and a 6-fold increase in hospitalizations between December 2020 and January 2021. The most plausible hypotheses that would justify this condition are: (i) the previous overestimation of seroprevalence in Manaus, (ii) the immune evasion property of some SARS-CoV-2 mutations found in VOCs, and (iii) higher transmissibility and pathogenicity of SARS-CoV-2 lineages circulating in the second wave compared with pre-existing lineages (12).

A genomic epidemiology study that used 250 SARS-CoV-2 genomes from 25 different municipalities from Amazonas sampled between March 2020, and January 2021 shows that the first exponential phase in the state was driven mainly by the spread of lineage B.1.195, which was gradually replaced by B.1.1.28. The second wave coincided with the emergence of P.1 in November, which rapidly replaced the parental lineage (<2 months) (10) and whose emergence
was preceded by a period of rapid molecular evolution (9). Importantly, rapid accumulation of mutations over short timeframes have been reported in chronically infected or immunocompromised hosts (86,87). However, preliminary findings pointed to the existence of P.1 intermediate lineages, suggesting that the constellation of mutations defining P.1 were acquired at sequential steps during multiple rounds of infections instead of within a single long-term infected individual (88). The VOC P.1 carries three deletions, four synonymous substitutions, a four base-pair nucleotide insertion, and at least 17 other lineage-defining replacements, including 10 missense mutations in the Spike protein (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I), eight of which are knowingly subjected to positive selection (9).

Regarding infectiousness, transmissibility, and case fatality, the viral load was ~10-fold higher in P.1 infections than in non-P.1 infections (10). Although another study points to uncertainties regarding viral load and duration of infection after accounting for confounding effects (9). Moreover, it was estimated to be 1.7–2.4-fold more transmissible, raising the probability that reinfections would be caused more frequently in hosts infected by P.1 rather than by older lineages. Remarkably, infections were 1.2–1.9 times more likely to result in death in the period following the emergence of P.1 compared to previous time frames (9). These findings support that successive lineage replacements in Amazonas were driven by a complex combination of factors, including the emergence of the more transmissible VOC P.1 virus (10).

A study conducted in RS described a P.1 lineage infection on November 30th followed by a P.2 lineage reinfection on March 11th in a patient with comorbidities. This report was the first detected P.1 in the state (22). Other analyses suggest that the P.1 lineage presumably emerged in Manaus, Brazil, in mid-November 2020 (9,10). Therefore, the P.1 lineage was present in Southern Brazil about just 15 days after the P.1 emergence in Brazil. Our molecular clock analysis supported this scenario. Another study, once thought to be the first P.1 report in RS, documented local transmission of P.1 from a person who had close contact with tourists and was positive to COVID-19 in early February 2021 (21). This happened in the city of Gramado, a town on the mountains that receives around 6.5 million tourists every year and belongs to the Caxias do Sul intermediate region. Interestingly, this sample from Gramado was the earliest representative of a new P.1-derived lineage (P.1.2), described in 11 patients from this study and found in transmission clusters.
from the RJ state in Southeastern Brazil, the USA, and the Netherlands. Remarkably, these local sequences are more similar to genomes from other countries compared to the RJ cluster, which acquired at least four additional mutations (including S:A262S) (17).

Whether P.1.2 has worse clinical outcomes than its prior variant (P.1) is unknown. However, as described above, the missense mutations characteristic of the new sublineage are located at nsp2 and nsp3 (ORF1ab), ORF3a, and Nucleocapsid. These sites are known for their interaction with human proteome, potentially influencing the immunological and inflammatory response against SARS-CoV-2 infection (89). The ORF3a:D155Y substitution is located near SARS-CoV caveolin-binding Domain IV. The binding interaction of viral ORF3a protein to host caveolin-1 is essential for entry and endomembrane trafficking of SARS-CoV-2. Since this mutation breaks the salt bridge formation between Asp155-Arg134, it can interfere with the binding affinity of ORF3a to host caveolin-1 and change virulence properties. Most importantly, this disrupted interaction may be associated with improved viral fitness, since it can avoid the induction of host cell apoptosis or extend the asymptomatic phase of infection (90). We hypothesize that these new substitutions could, therefore, influence epidemiological and clinical outcomes favouring P.1.2 evolution. This is elusive at best at this time, however, and further sublineage characterization is needed to further exploit its real relevance.

Some limitations should be considered. Firstly, the sample size is low and not necessarily representative of the RS state. Furthermore, publicly available genomes are a result of episodic sequencing efforts, especially in Brazil. This scenario restricts more precise inferences about introductions and diffusion processes in regional and worldwide contexts since samples are not geographical and temporally well distributed. Therefore, more research and surveillance are essential to unravel a more precise genomic characterization of SARS-CoV-2 in Brazil, identifying novel variants promptly to better respond and control its spread.

In summary, our study corroborates the total virtual substitution of previous lineages by P.1 in Southern Brazil in COVID-19 cases sequenced in March 2020. Moreover, we confirmed various cases caused by the novel P.1.2 sublineage and placed its origin in the State of Rio Grande do Sul. The continuous evolution of the VOC P.1 is worrisome, considering its clinical and epidemiological impact, and warrants enhanced genomic surveillance.
Availability of data and materials

Full tables acknowledging the authors and corresponding labs submitting sequencing data used in this study can be found in Supplementary Files 3 and 4. Consensus genomes generated in this study were deposited in the GISAID database under Accession IDs: EPI_ISL_2139494 to EPI_ISL_2139549. Additional information used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Funding

This study was supported by donations from Florense Brands, Beppler & Puppi Advogados, Smellbox Produtos de Higiene Ltda., and Dr. Leonardo Mestre Negri. 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 (UFCSPA). The funders had no role in the study design, data generation and analysis, decision to publish or the preparation of the manuscript.

Author’s contributions

Formal analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration.

All authors have read and approved the manuscript.

Acknowledgements

We thank the administrators of the GISAID database and research groups across the world for supporting the rapid and transparent sharing of genomic data during the COVID-19 pandemic. We also thank the staff of Hospital da Brigada Militar, Laboratório Exame, Beppler & Puppi Advogados, Smellbox Produtos de Higiene Ltda., Dr. Leonardo Mestre Negri, Florense Brands, and Biome that directly contributed to this study.
References


### Table 1. Epidemiological characteristics of the 56 sequenced samples from Rio Grande do Sul, Southern Brazil

<table>
<thead>
<tr>
<th>Study ID (HBM-RS)</th>
<th>GISAID ID (EPI_ISL_)</th>
<th>Cycle threshold</th>
<th>Municipality of residence</th>
<th>Gender</th>
<th>Age group</th>
<th>Lineage</th>
<th>Contact with confirmed case</th>
</tr>
</thead>
<tbody>
<tr>
<td>39468</td>
<td>2139494</td>
<td>16</td>
<td>São Leopoldo</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39469</td>
<td>2139495</td>
<td>19</td>
<td>Porto Alegre</td>
<td>Female</td>
<td>20-29</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39470</td>
<td>2139496</td>
<td>19</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>60-69</td>
<td>P.1</td>
<td>No</td>
</tr>
<tr>
<td>39471</td>
<td>2139497</td>
<td>18</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39472</td>
<td>2139498</td>
<td>17</td>
<td>Gravatai</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>No</td>
</tr>
<tr>
<td>39473</td>
<td>2139499</td>
<td>26</td>
<td>Cachoeira do Sul</td>
<td>Female</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39474</td>
<td>2139500</td>
<td>18</td>
<td>Gravatai</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39475</td>
<td>2139501</td>
<td>18</td>
<td>Porto Alegre</td>
<td>Female</td>
<td>20-29</td>
<td>P.1</td>
<td>No</td>
</tr>
<tr>
<td>39476</td>
<td>2139502</td>
<td>15</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39477</td>
<td>2139503</td>
<td>21</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39478</td>
<td>2139504</td>
<td>15</td>
<td>Cachoeira do Sul</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39479</td>
<td>2139505</td>
<td>22</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>50-59</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39480</td>
<td>2139506</td>
<td>17</td>
<td>Novo Hamburgo</td>
<td>Male</td>
<td>40-49</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39481</td>
<td>2139507</td>
<td>14</td>
<td>Porto Alegre</td>
<td>Female</td>
<td>70-79</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39482</td>
<td>2139508</td>
<td>14</td>
<td>Porto Alegre</td>
<td>Female</td>
<td>80-89</td>
<td>P.1.2</td>
<td>No</td>
</tr>
<tr>
<td>39483</td>
<td>2139509</td>
<td>13</td>
<td>Gravatai</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39484</td>
<td>2139510</td>
<td>20</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39485</td>
<td>2139511</td>
<td>16</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>50-59</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39486</td>
<td>2139512</td>
<td>27</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>30-39</td>
<td>P.2</td>
<td>No</td>
</tr>
<tr>
<td>39487</td>
<td>2139513</td>
<td>14</td>
<td>São Sebastião do Cai</td>
<td>Male</td>
<td>40-49</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39488</td>
<td>2139514</td>
<td>28</td>
<td>Santo Antônio da Patrulha</td>
<td>Male</td>
<td>70-79</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39489</td>
<td>2139515</td>
<td>27</td>
<td>Porto Alegre</td>
<td>Female</td>
<td>20-29</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39490</td>
<td>2139516</td>
<td>18</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39491</td>
<td>2139517</td>
<td>15</td>
<td>Alvorada</td>
<td>Female</td>
<td>20-29</td>
<td>B.1.1.28</td>
<td>Yes</td>
</tr>
<tr>
<td>39492</td>
<td>2139518</td>
<td>17</td>
<td>Gravatai</td>
<td>Female</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39493</td>
<td>2139519</td>
<td>22</td>
<td>Canoas</td>
<td>Male</td>
<td>30-39</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39494</td>
<td>2139520</td>
<td>17</td>
<td>Porto Alegre</td>
<td>Female</td>
<td>30-39</td>
<td>P.1</td>
<td>No</td>
</tr>
<tr>
<td>39495</td>
<td>2139521</td>
<td>17</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39496</td>
<td>2139522</td>
<td>17</td>
<td>Canoas</td>
<td>Female</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39497</td>
<td>2139523</td>
<td>21</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>40-49</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39498</td>
<td>2139524</td>
<td>20</td>
<td>Porto Alegre</td>
<td>Female</td>
<td>40-49</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39499</td>
<td>2139525</td>
<td>22</td>
<td>Portão</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39500</td>
<td>2139526</td>
<td>11</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>20-29</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39501</td>
<td>2139527</td>
<td>14</td>
<td>Santa Maria</td>
<td>Male</td>
<td>20-29</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39502</td>
<td>2139528</td>
<td>21</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39503</td>
<td>2139529</td>
<td>16</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>30-39</td>
<td>P.1</td>
<td>No</td>
</tr>
<tr>
<td>39504</td>
<td>2139530</td>
<td>21</td>
<td>Gravatai</td>
<td>Male</td>
<td>40-49</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39505</td>
<td>2139531</td>
<td>13</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>30-39</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39506</td>
<td>2139532</td>
<td>23</td>
<td>Porto Alegre</td>
<td>Female</td>
<td>40-49</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39507</td>
<td>2139533</td>
<td>28</td>
<td>Canoas</td>
<td>Female</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39508</td>
<td>2139534</td>
<td>22</td>
<td>Porto Alegre</td>
<td>Male</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39509</td>
<td>2139535</td>
<td>23</td>
<td>Alvorada</td>
<td>Male</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39510</td>
<td>2139536</td>
<td>19</td>
<td>Canoas</td>
<td>Male</td>
<td>50-59</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>ID</td>
<td>Sample Code</td>
<td>Age</td>
<td>Gender</td>
<td>Location</td>
<td>Age Group</td>
<td>Study ID</td>
<td>Result</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>-----</td>
<td>--------</td>
<td>------------------</td>
<td>-----------</td>
<td>----------</td>
<td>--------</td>
</tr>
<tr>
<td>39511</td>
<td>2139537</td>
<td>22</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>30-39</td>
<td>P.1</td>
<td>No</td>
</tr>
<tr>
<td>39512</td>
<td>2139538</td>
<td>25</td>
<td>Female</td>
<td>Cachoeira do Sul</td>
<td>40-49</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39513</td>
<td>2139539</td>
<td>23</td>
<td>Male</td>
<td>Santa Maria</td>
<td>40-49</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39514</td>
<td>2139540</td>
<td>15</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>30-39</td>
<td>P.1</td>
<td>No</td>
</tr>
<tr>
<td>39515</td>
<td>2139541</td>
<td>21</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39516</td>
<td>2139542</td>
<td>28</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>50-59</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39517</td>
<td>2139543</td>
<td>17</td>
<td>Male</td>
<td>Sapiranga</td>
<td>30-39</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39518</td>
<td>2139544</td>
<td>17</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>30-39</td>
<td>P.1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>39519</td>
<td>2139545</td>
<td>23</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39520</td>
<td>2139546</td>
<td>15</td>
<td>Male</td>
<td>Campo Bom</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39521</td>
<td>2139547</td>
<td>15</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39522</td>
<td>2139548</td>
<td>21</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>50-59</td>
<td>P.1</td>
<td>Yes</td>
</tr>
<tr>
<td>39523</td>
<td>2139549</td>
<td>18</td>
<td>Male</td>
<td>Porto Alegre</td>
<td>20-29</td>
<td>P.1</td>
<td>Yes</td>
</tr>
</tbody>
</table>

All samples were nasopharyngeal swabs collected between March 9—17, 2021, from residents of the RS state. Study ID: Study identifier only known by study investigators. The Ct values are related to the first RT-qPCR conducted at Laboratório Exame.
Table 2. Detailed description and frequency of mutations found in our 56 sequences compared with all Brazilian P.1 sequences until April 26, 2021.

<table>
<thead>
<tr>
<th>Genomic position</th>
<th>Effect</th>
<th>Amino acid change</th>
<th>Gene/region</th>
<th>Product</th>
<th>Frequency our study (%)</th>
<th>Frequency in Brazilian's P.1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C241T</td>
<td>Intergenic</td>
<td>NA</td>
<td>5' UTR</td>
<td>NA</td>
<td>100.0</td>
<td>97.2</td>
</tr>
<tr>
<td>T733C</td>
<td>Synonymous</td>
<td>D156D</td>
<td>Leader Protein</td>
<td>96.4</td>
<td>99.9</td>
<td></td>
</tr>
<tr>
<td>C1912T</td>
<td>Synonymous</td>
<td>S549S</td>
<td>nsp2</td>
<td>19.6</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>A2550G</td>
<td>Missense</td>
<td>D762G</td>
<td></td>
<td>19.6</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>C2749T</td>
<td>Synonymous</td>
<td>D828D</td>
<td></td>
<td>94.6</td>
<td>99.7</td>
<td></td>
</tr>
<tr>
<td>C3037T</td>
<td>Synonymous</td>
<td>F924F</td>
<td></td>
<td>100.0</td>
<td>99.9</td>
<td></td>
</tr>
<tr>
<td>C3828T</td>
<td>Missense</td>
<td>S1188L</td>
<td></td>
<td>96.4</td>
<td>95.3</td>
<td></td>
</tr>
<tr>
<td>A5648C</td>
<td>Missense</td>
<td>K1795Q</td>
<td>nsp3</td>
<td>85.7</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>C5724T</td>
<td>Missense</td>
<td>T1820I</td>
<td>ORF1ab</td>
<td>17.9</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>A6319G</td>
<td>Synonymous</td>
<td>P2018P</td>
<td></td>
<td>87.5</td>
<td>99.5</td>
<td></td>
</tr>
<tr>
<td>A6613G</td>
<td>Synonymous</td>
<td>V2116V</td>
<td></td>
<td>96.4</td>
<td>99.8</td>
<td></td>
</tr>
<tr>
<td>T11296G</td>
<td>Missense</td>
<td>F3677L</td>
<td>nsp6</td>
<td>30.4</td>
<td>8.2</td>
<td></td>
</tr>
<tr>
<td>C12778T</td>
<td>Synonymous</td>
<td>Y4171Y</td>
<td>nsp9</td>
<td>94.6</td>
<td>98.9</td>
<td></td>
</tr>
<tr>
<td>C13860T</td>
<td>Missense</td>
<td>T4532I</td>
<td>RdRp</td>
<td>94.6</td>
<td>99.8</td>
<td></td>
</tr>
<tr>
<td>C14408T</td>
<td>Synonymous</td>
<td>L4715L</td>
<td></td>
<td>98.2</td>
<td>96.8</td>
<td></td>
</tr>
<tr>
<td>G17259T</td>
<td>Missense</td>
<td>S5665I</td>
<td>Helicase</td>
<td>94.6</td>
<td>99.7</td>
<td></td>
</tr>
<tr>
<td>C21614T</td>
<td>Missense</td>
<td>L18F</td>
<td></td>
<td>96.4</td>
<td>99.9</td>
<td></td>
</tr>
<tr>
<td>C21621A</td>
<td>Missense</td>
<td>T20N</td>
<td></td>
<td>94.6</td>
<td>99.8</td>
<td></td>
</tr>
<tr>
<td>C21638T</td>
<td>Missense</td>
<td>P26S</td>
<td></td>
<td>96.4</td>
<td>99.1</td>
<td></td>
</tr>
<tr>
<td>G21974T</td>
<td>Missense</td>
<td>D138Y</td>
<td>S Surface Glycoprotein</td>
<td>96.4</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>G22132T</td>
<td>Missense</td>
<td>R190S</td>
<td></td>
<td>96.4</td>
<td>98.4</td>
<td></td>
</tr>
<tr>
<td>A22812C</td>
<td>Missense</td>
<td>K417T</td>
<td></td>
<td>96.4</td>
<td>83.4</td>
<td></td>
</tr>
<tr>
<td>G23012A</td>
<td>Missense</td>
<td>E484K</td>
<td></td>
<td>96.4</td>
<td>99.9</td>
<td></td>
</tr>
<tr>
<td>Genomic Coordinate</td>
<td>Type</td>
<td>Amino Acid Change</td>
<td>Frequency</td>
<td>Coverage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>---------</td>
<td>-------------------</td>
<td>-----------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A23063T</td>
<td>Missense</td>
<td>N501Y</td>
<td>96.4</td>
<td>99.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A23403G</td>
<td>Missense</td>
<td>D614G</td>
<td>100.0</td>
<td>97.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C23525T</td>
<td>Missense</td>
<td>H655Y</td>
<td>96.4</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C24642T</td>
<td>Missense</td>
<td>T1027I</td>
<td>96.4</td>
<td>99.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G25088T</td>
<td>Missense</td>
<td>V1176F</td>
<td>100.0</td>
<td>99.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G25855T</td>
<td>Missense</td>
<td>D155Y</td>
<td>19.6</td>
<td>1.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T26149C</td>
<td>Missense</td>
<td>S253P</td>
<td>94.6</td>
<td>98.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G28167A</td>
<td>Missense</td>
<td>E92K</td>
<td>94.6</td>
<td>99.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C28512G</td>
<td>Missense</td>
<td>P80R</td>
<td>96.4</td>
<td>98.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C28789T</td>
<td>Synonymous</td>
<td>Y172Y</td>
<td>19.6</td>
<td>1.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGTAGGG 28877-28883TCTAAAC</td>
<td>Missense</td>
<td>RG203-204KR</td>
<td>96.4</td>
<td>99.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Original bases or amino acids are represented before the genomic coordinate, while the mutated ones are presented after. Only mutations observed in more than 10 genomes from this study are shown. P.1 lineage-defining mutations are highlighted in bold. P.1.2 (new lineage) defining replacements are underlined.

**Figures**

**Figure 1.** Mutations of the SARS-CoV-2 genomes from the RS state, Southern Brazil sampled in March 2021.

(A) Genome map for the 56 genomes sequenced. Nucleotide substitutions are colored in red and blank regions represent low sequencing coverage.

(B) Frequency of SNPs per SARS-CoV-2 genome position along the 56 genomes. These mutations are corresponding to the red lines in (A), and only missense substitutions represented by >10 sequences have their respective amino acid changes indicated above the bars. Main Open Reading Frames (ORFs) and SARS-CoV-2 proteins are indicated at the bottom to allow a rapid visualization of the viral proteins affected.
Figure 2. Distribution of SARS-CoV-2 lineages across time in Argentina, other Brazilian regions and in the RS state.

(A) Per month lineage distribution in Argentina, entire Brazil and the RS state. Only the 10 most prevalent lineages were considered.

(B) Timeline showing the distribution of the most prevalent lineages until the end of 2020 (B.1.1.28 and B.1.1.33) and from the end of 2020 onward (P.1 and P.2) for the fifth Brazilian regions, the RS state and considering only this study (March 2021).

(C) Abundance of different lineages in the eight intermediate regions of the RS state defined by IBGE. The region of Porto Alegre was amplified and the frequencies from 2020, 2021 and only considering the present study are presented on the right.

NA=Other lineages.
Figure 3. Phylogenetics analysis of genomes sequences in the RS state in global context.
(A) Time resolved ML tree of 8,635 global representative SARS-CoV-2 genomes. Circles represent global sequences belonging to the five most abundant lineages in the RS state that passed quality control criteria: B.1.1 (n=32), P.1 (n=83), P.2 (n=83), B.1.1.28 (n=203), B.1.1.33 (n=286), and diamonds represent RS genomes (available in GISAID and sequenced in this study). Root-to-tip regression is represented on the right of the tree.
(B) ML tree of 4,499 SARS-CoV-2 genomes belonging to the P.1 lineage. Tips are colored by Brazilian regions, South America or Other continents. Introductions, clusters and clades are annotated in the tree (see Methods). Root-to-tip regression is depicted on the left of the tree and sequences from “Other continents” were dropped to improve visualization.
Figure 4. Bayesian discrete asymmetric phylogeographic analysis of the identified Clade 1 (lineage P.1.2).

(A) Root-to-tip regression of genetic distances and sampling dates for Clade 1. Correlation coefficient and R squared are depicted above the graph.
(B) MCC tree of the 93 sequences included in this analysis up to 26 April, 2021 (82 from GISAID and 11 from this study). Numbers above branches represent the posterior probability of each branch. Only posteriors > 0.5 are shown.

Circles indicate countries outside Brazil and Brazilian states (BR suffix).
Figure 5. Bayesian discrete asymmetric phylogeographic analysis of the identified clades 2, 3 and 4.

(A) MCC tree of the 71 sequences included in this analysis up to 26 April, 2021.
(B) MCC tree of the 122 genomes included in this analysis up to 26 April, 2021.
(C) MCC tree of the 50 sequences included in this analysis up to 26 April, 2021.

For all MCC trees, numbers above branches represent the posterior probability of each branch. Only posteriors > 0.5 are shown. Asterisks represent potential introductions in the RS state and subclades cited in the text are indicated. Circles indicate countries outside Brazil and Brazilian states (BR suffix).
Supplementary files

Supplementary file 1. Coverage depth plots for each genome sequenced.

Supplementary file 2. Collection of all mutations (n=175) found in the 56 sequenced genomes, including effects on SARS-CoV-2 genes and proteins. The table is ordered by genomic position.

Supplementary file 3. GISAID acknowledgment table of the 8,635 global sequences used in the Nextstrain workflow.

Supplementary file 4. GISAID acknowledgment table of the 4,499 P.1 sequences analyzed.

Supplementary File 5. Expanded ML tree built using 4,499 P.1 sequences deposited in GISAID until 26 April, 2021. Tips represented by sequences sequenced in this study are augmented to enable visualization of the different introductions, clusters and clades reported.