Identification of SARS-CoV-2 P.1-related lineages in Brazil provides new insights about the mechanisms of emergence of Variants of Concern

Tiago Gräf (✉ akograf@gmail.com)
Plataforma de Vigilância Molecular, Instituto Gonçalo Moniz, Fiocruz, Salvador, Bahia, Brazil
https://orcid.org/0000-0003-4921-7975

Gonzalo Bello
Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil.

Taina Moreira Martins Venas
Laboratório de Vírus Respiratórios e do Sarampo (LVRS), Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil

Elisa Cavalcante Pereira
Laboratório de Vírus Respiratórios e

Anna Carolina Dias Paixão
Laboratório de Vírus Respiratórios e

Luciana Reis Appolinario
Laboratório de Vírus Respiratórios e do Sarampo (LVRS), Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil

Renata Serrano Lopes
Laboratório de Vírus Respiratórios e do Sarampo (LVRS), Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil

Ana Carolina da Fonseca Mendonça
Laboratório de Vírus Respiratórios e do Sarampo (LVRS), Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil

Alice Sampaio Barreto da Rocha
Laboratório de Vírus Respiratórios e do Sarampo (LVRS), Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil

Fernando Couto Motta
Laboratório de Vírus Respiratórios e do Sarampo (LVRS), Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil

Tatiana Schäffer Gregianini
Laboratório Central de Saúde Pública do Estado do Rio Grande do Sul (LACEN-RS), Brazil

Richard Steiner Salvato
Laboratório Central de Saúde Pública do Estado do Rio Grande do Sul (LACEN-RS), Brazil
Research Article

Keywords: SARS-CoV-2; genomic surveillance; Brazil; variant of concern; lineage P.1

Posted Date: June 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-580195/v1

License: ☑ This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

One of the most remarkable features of the SARS-CoV-2 Variants of Concern (VOC) is the unusually large number of mutations they carry. However, the specific factors that drove the emergence of such variants since the second half of 2020 are not fully resolved. In this study, we described a new SARS-CoV-2 lineage provisionally designated as P.1-like-II that, as well as the previously described lineage P.1-like-I, shares several lineage-defining mutations with the VOC P.1 circulating in Brazil. Reconstructions of P.1 ancestor sequences demonstrate that the entire constellation of mutations that define the VOC P.1 did not accumulate within a single long-term infected individual, but was acquired by sequential addition during interhost transmissions. Our evolutionary analyses further estimate that P.1-ancestors strains carrying half of the P.1-lineage-defining mutations, including those at the receptor-binding domain of the Spike protein, circulated cryptically in the Amazonas state since August 2020. This evolutionary pattern is consistent with the hypothesis that partial human population immunity acquired from natural SARS-CoV-2 infections during the first half of 2020 might have been the major driving force behind natural selection that allowed VOCs' emergence and worldwide spread. These findings also support a long lag-time between the emergence of variants with key mutations of concern and expansion of the VOC P.1 in Brazil.

Introduction

The emergence of the SARS-CoV-2 variant of concern (VOC) P.1 in the Brazilian Amazonas state around November 2020 \(^1,2\) and its rapid dissemination to other regions was associated with a major COVID-19 epidemic wave that collapsed the Brazilian health system during early 2021. The lineage P.1, as the other described VOCs, harbors a large number of lineage-defining mutations, including ten non-synonymous substitutions in the Spike (S) protein (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I), five non-synonymous mutations distributed in the NSP3 (S370L and K977Q), NSP13 (E341D), NS8 (E92K) and N (P80R) proteins, one deletion in the NSP6 (S106del, G107del, F108del) and a four-nucleotide insertion at ORF8/N intergenic region (ins28263) \(^1,2\).

The most accepted hypothesis to explain such a high number of lineage-defining mutations is that VOCs result from selective pressures and adaptation of the virus during prolonged individual infections and subsequent transmission \(^3\). This hypothesis, however, was challenged by the early discovery of four P.1-like genomes, most of them sampled in the capital city of Amazonas state, that branched as a sister monophyletic clade concerning lineage P.1 \(^1,4\). The P.1-like clade also accumulated an unusually high number of genetic changes, including several P.1 lineage-defining mutations in the S (L18F, P26S, D138Y, K417T, E484K, N501Y), NSP3 (K977Q), and N (P80R) proteins and unique mutations in the NSP2 (K456R), NSP3 (T1189I), NSP6 (V149A), NSP13 (S74L), S (ins214 and D1139H) and NS8 (K2stop) proteins. This early finding supports the hypothesis that P.1 lineage-defining mutations did not accumulate in a unique long-term individual infection, but were acquired at sequential steps during the evolution of lineage B.1.1.28 in Amazonas.
In this study, we describe a second P.1-related virus variant that is spreading in several states from the different Brazilian regions and harbors 15 P.1 lineage-defining mutations and six unique mutations. The description of this new P.1-related variant allowed us to trace with more precision the evolutionary steps that resulted in the emergence of the VOC P.1. Moreover, these results confirm our previous hypothesis that some of the P.1 lineage-defining mutations were sequentially fixed over several months during the second half of 2020. Our analyses also revealed that despite sharing crucial mutations in the RBD of the S protein, the P.1-like variants displayed a much less efficient epidemic spread in Brazil than the VOC P.1.

**Materials And Methods**

**Ethics statement**

This study was approved by the FIOCRUZ-IOC (68118417.6.0000.5248 and CAAE 32333120.4.0000.5190) and the Amazonas State University Ethics Committee (CAAE: 25430719.6.0000.5016), and the Brazilian Ministry of the Environment (MMA) A1767C3.

**SARS-CoV-2 whole-genome sequencing**

Our genomic survey of SARS-CoV-2 positive samples sequenced by the Fiocruz COVID-19 Genomic Surveillance Network between 12\textsuperscript{th} March 2020 and 31\textsuperscript{st} March 2021 identified 44 sequences (EPI\_ISL\_2038926 to EPI\_ISL\_2038968, EPI\_ISL\_2102018, and EPI\_ISL\_2102063, Supplementary Table S1) with several overlapping mutations with the lineage P.1 (Table S1). The SARS-CoV-2 whole-genomes (>99% coverage) were recovered using Illumina sequencing protocols as previously described\textsuperscript{5,6}. The FASTQ reads obtained were imported into the CLC Genomics Workbench version 20.0.4 (Qiagen A/S, Denmark), trimmed, and mapped against the reference sequence EPI\_ISL\_402124 available in EpiCoV database in the GISAID (https://www.gisaid.org/). The alignment was refined using the InDels and Structural Variants module.

**Maximum likelihood phylogenetic analyses**

SARS-CoV-2 P.1-related sequences here obtained were aligned with high quality (<5% of N) and complete (>29 kb) sequences that were available in the EpiCoV database in the GISAID (https://www.gisaid.org/) on March 31\textsuperscript{st}, 2021 and belongs to three different clades: 1) B.1.1.28 sequences from Amazonas state, 2) P.1 sequences, and 3) previously described P.1-like sequences\textsuperscript{1,4}. This dataset was then aligned using MAFFT v7.475\textsuperscript{7} and subjected to maximum likelihood (ML) phylogenetic analysis using IQ-TREE v2.1.2\textsuperscript{8} under the GTR+F+R4 nucleotide substitution model, as selected by the ModelFinder application\textsuperscript{9}. Branch support was assessed by the approximate likelihood-ratio test based on the Shimodaira–Hasegawa procedure (SH-aLRT) with 1000 replicates. The sequence of ancestral nodes was reconstructed using Time-tree\textsuperscript{10}, and their mutational profile was investigated using the Nextclade tool (https://clades.nextstrain.org). The temporal signal was assessed by the regression analysis of the root-
to-tip genetic distance estimated from the ML phylogenetic tree against sampling dates using the program TempEst.  

**Bayesian phylogeographic analyses**

A time-scaled phylogenetic tree of the B.1.1.28 Amazonian diversity plus a subsampling of P1 genomes and P1-related sequences was reconstructed using the Bayesian Markov Chain Monte Carlo (MCMC) approach implemented in BEAST 1.10.4. A Bayesian tree was reconstructed using the GTR+F+G4 nucleotide substitution model, the Bayesian skyline (BSKL) coalescent model, and both strict and random local molecular clock models with a uniform substitution rate prior ($8 \times 10^{-4}$ – $10 \times 10^{-4}$ substitutions/site/year). Ancestral sampling locations were inferred using a reversible discrete phylogeographic model where transitions between Brazilian states were estimated in a continuous-time Markov chain (CTMC) rate reference prior. Convergence (effective sample size > 200) in parameter estimates was assessed using TRACER v1.7. The maximum clade credibility (MCC) tree was summarized with TreeAnnotator v1.10.4. ML and MCC trees were visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

**Results**

Mutation profile analysis of SARS-CoV-2 positive samples detected at different Brazilian states between 12th March 2020 and 31st March 2021 revealed 44 sequences (Table S1) that harbor 15 out of 22 P1 lineage-defining mutations, including the three mutations of concern at the receptor-binding domain (RBD) of the S protein (K417T, E484K, and N501Y), deletion in the NSP6 (S106del, G107del, F108del) and the four-nucleotide insertion at ORF8/N intergenic region (ins28263) (Figure 1). These P1-related sequences, here designated as P1-like-II, lack some of the P1 lineage-defining mutations at ORF1ab (C2749T, C12778T, and C13860T), NSP13 (E341D), S (T20N) and NS8 (E92K), and further displayed six unique substitutions at ORF1ab (C8905T, C16954T, and A20931G), NSP4 (D217H), E/M intergenic region (A26492T), and N (P383L). The P1-like-II sequences also share nine P1 lineage-defining mutations with the previously characterized P1-like clade (now designated as P1-like-I) (Figure 1).

ML phylogenetic analysis revealed that P1-like-II sequences branched in a highly supported (SH-aLRT = 96.6%) monophyletic clade together with seven sequences retrieved from the EpiCoV database (https://www.gisaid.org/) that displayed the same mutation profile and were classified as P1 in the EpiCoV database (Figure 2a). Clades P1-like-I and P1-like-II are not nested within the diversity of the VOC P1, but branch as sister monophyletic clades that evolved from a common ancestor. Although clades P1, P1-like-I, and II do not share the same set of lineage-defining mutations, they were designated as lineage P1 according to the PANGO rules. This classification is based on the mutations of concern (K417T, E484K, and N501Y) acquired in the same evolutionary event (https://github.com/cov-lineages/pango-designation/issues/77). We will then use lineage P1 to designate the entire clade comprising the original P1 and the new P1-like sub-lineages. VOC (or clade) P1 will be used to designate only the first P1 sub-lineage identified that dominated the Brazilian epidemic in 2021.
The P.1-like-II genomes were sampled at nine different Brazilian states, mainly from the South and Southeast regions (Figure 2b). The oldest one was detected in the Rio de Janeiro state on 19th January 2021, and the most recent one was identified in this study in the Amazonas state on 25th March 2021. The Brazilian state that comprises most P.1-like-II sequences identified so far was Santa Catarina (59%), followed by Rio de Janeiro (10%), Rio Grande do Sul (8%), and São Paulo (8%). Thus, unlike the clade P.1 that was efficiently disseminated both within and outside the Amazonas state, the clade P.1-like-II was more efficiently disseminated outside the Amazonas state. It is also important to note that while VOC P.1 comprises a substantial fraction (66%) (http://www.genomahcov.fiocruz.br) of SARS-CoV-2 sequences sampled at different Brazilian states during 2021, clades P.1-like-I and P.1-like-II comprises less than 1% of samples genotyped; supporting more successful dissemination of clade P.1 with respect to P.1-like clades in Brazil.

Analysis of the temporal structure revealed that clades P.1, P.1-like-I, and II accumulated a higher number of mutations when compared to B.1.1.28 sequences and evolved at a similar rate over time (Figure 2c). Reconstruction of sequences at ancestral nodes provides a clear picture of the evolutionary steps that resulted in the different P.1 and P.1-related variants (Figure 3). Three mutations were fixed in the basal B.1.1.28 Amazonian clade (previously named 28-AM-II) (1) from which all P.1 clades evolved. Nine mutations were fixed in the following evolutionary step that gave origin to the most recent common ancestor (MRCA) of lineage P.1 (designated as P.1$_{MRCA1}$). Six additional mutations were fixed in the evolutionary step that gave origin to the MRCA of clades P.1 and P.1-like-II (designated as P.1$_{MRCA2}$), and 6-12 mutations were fixed in the branches that originate the MRCA of each clade. Six out of the nine (67%) mutations in P.1$_{MRCA1}$ were in the S protein (including the three mutations of concern in the RBD), while only seven out of 32 (22%) mutations fixed in the subsequent steps were located in the S gene. It is also interesting to note that the total number of lineage-defining mutations accumulated by clades P.1 (n = 12), P.1-like-I (n = 14), and P.1-like-II (n = 12) since their divergence from P.1$_{MRCA1}$ was almost the same.

Bayesian phylogeographic analysis was next conducted combining all B.1.1.28 sequences from Amazonas (including clade 28-AM-II), early VOC P.1 viruses sampled in December 2020, and all P.1-like sequences. This analysis supports that most ancestors during the diversification of lineage P.1 were probably located in the state of Amazonas (Posterior State Probability [PSP] = 1). The only exception was the P.1-like-II ancestor whose posterior probability was divided between Amazonas (PSP = 0.40) and Santa Catarina (PSP = 0.31) (Figure 4). The great uncertainty in the location of the P.1-like-II ancestor probably reflects the low number of sequences from this clade detected in the Amazonas state so far, making it difficult to trace their origin to that Northern state. This analysis estimated that Santa Catarina was the most critical hub of dissemination of lineage P.1-like-II to other Brazilian states. It is also noteworthy that P.1-like-II genomes from Rio de Janeiro formed an independent basal cluster, supporting local transmission of this lineage in this state. The different molecular clock models used consistently traced the median time of the P.1$_{MRCA1}$ to mid-August 2020, the median time of the P.1$_{MRCA2}$ to late September 2020, and the emergence of clades P.1 and P.1-like to around late November and late December 2020, respectively (Table).
Discussion

Our genomic surveillance identified a new P.1-related genetic variant derived from the lineage B.1.1.28 Amazonian diversity designated as clade P.1-like-II. It shares a common ancestor and several lineage-defining mutations, including the mutations of concern in the RBD of the S protein (K417T, E484K, N501Y), with the VOC P.1 and the clade P.1-like-I previously identified by our group. The new clade P.1-like-II displayed an overall low prevalence (<1%), but is geographically dispersed in Brazil, particularly in the South and Southeast country regions.

The most widely accepted hypothesis suggests that mutations in VOCs arose during long-standing SARS-CoV-2 single infections, like those observed in immunosuppressed subjects. Our findings, however, revealed that the final constellation of mutations observed in the VOC P.1 was acquired through multiple interhost transmissions. During this evolutionary process that probably took several months, the stepwise acquisition of mutations was not uniformly distributed along the viral genome. Most VOC P.1 defining mutations located in the amino(N)-terminal domain (NTD; L18F, P26S, D138Y) and in the RBD (K417T, E484K, N501Y) of the S protein were fixed in the first evolutionary step; while most mutations located outside the S gene were fixed at subsequent steps. It is noteworthy that more intermediate evolutionary steps could exist between clade 28-AM-II and VOC P.1. However, the currently limited number of available genomes sampled in Amazonas between August and November (n = 87) limits the resolution of the evolutionary history reconstructed here.

The stepwise diversification of lineage P.1 in Brazil resembles the evolutionary pattern of the VOCs B.1.351 and B.1.617 that were first detected in South Africa and India, respectively. Similar to the P.1 family clades described in Brazil, the VOCs B.1.351 and B.1.617 also comprise a family of related clades with partial overlapping mutations. The mutation profile of lineage B.1.351 suggests that five non-synonymous mutations in the S protein (D80A, D215G, E484K, N501Y, and A701V) were fixed at the first progenitor and further S mutations (L18F, 242-244del, R246I, and K417N) were fixed at later steps in different descendent sub-lineages. Lineage B.1.617 was initially defined as a double S mutant (L452R and E484), but subsequent phylogenetic analysis revealed a high within lineage diversity with at least four different sub-clusters (PANGO lineages B.1.617, B.1.617.1, B.1.617.2, and B.1.617.3) that could be linked to partially overlapping constellations of S mutations.

Although this stepwise evolutionary pattern does not exclude the possibility that at least a subset of mutations could have originated in a long-term infected individual, sequential infections of such kind of patients are very unlikely. We propose that mutations of concern have been naturally selected during acute reinfections of partially protected immunocompetent individuals. According to this hypothesis, the partial immunity that human populations acquired through natural SARS-CoV-2 infections during early 2020 was a major selective force that drove the sequential emergence of mutations of concern in the second half of 2020. This model is consistent with a recent study that revealed a major change in selective pressures acting on SARS-CoV-2 variants circulating worldwide after October 2020, coinciding with the simultaneous expansion of different VOCs with convergent S mutations. This model is also
consistent with the ongoing evolution of the VOC P.1 in Brazil revealed by the recurrent acquisition of indels in the NTD of the S protein.

The presence of key mutations of concern in the RBD of S protein (K417T, E484K, N501Y) of the VOC P.1 can explain the higher transmissibility and successful dissemination of this VOC with respect to previous circulating B.1.1.28 lineage in Amazonas. Our analysis, however, suggests that RBD mutations were not the only driver of the P.1 expansion. First, our evolutionary reconstruction suggests that the ancestors P1_{MRCA1} and P1_{MRCA2}, which harbor the three key mutations of concern in the RBD, circulated cryptically in the Amazonas state since August-September 2020, without fueling a large outbreak. Second, despite all P.1 sub-lineages share the same key mutations of concern, the estimated prevalence of VOC P.1 (69%) in 2021 was much higher than that of clades P1-like-I and II (<1% each). These pieces of evidence suggest that viral mutations combined with human factors, such as lack of social distancing measures and mass gatherings events, may have contributed to the remarkable dissemination of the VOC P.1 in the Amazonas state and throughout Brazil afterward.

The time-lag between the emergence of variant progenitors carrying key mutations of concern and the start of epidemic waves observed in Amazonas was also observed in South Africa and India. The emergence of the B.1.351 progenitor, which harbors key RBD mutations (K417N, E484K, N501Y), was traced in South Africa around late August 2020, while the country’s second COVID-19 epidemic wave only began at the end of October 2020. Similarly, the B.1.617 progenitor with key RBD mutations (E484Q, L452R) probably dates back before October 2020 while the second COVID-19 epidemic wave in India only began in February 2021. It is also observed that sub-lineages B.1.167.1 (that dominates in India), B.1.167.2 (that is spreading in India and in the United Kingdom), and B.1.167.3 (that remained uncommon in India and elsewhere) displayed quite divergent epidemic trajectories, thus supporting a complex interplay between presence of mutations of concern and epidemic dynamics of SARS-CoV-2 lineages.

In summary, our findings reveal that VOC P.1 is part of a more diverse family of P1-related variants that evolved from a common ancestor, which carried key mutations of concern and circulated in Amazonas months before the abrupt resurgence of COVID-19 in the state in late 2020. The entire constellation of mutations that define the VOC P.1 was acquired in a stepwise process during multiple interhost transmissions. This stepwise interhost model, in opposition to the single long-term intrahost infection hypothesis, seems to be the most likely evolutionary mechanism to explain the emergence of VOCs in Brazil (P.1), South Africa (B.1.351), and India (B.1.617). The divergent epidemic trajectories of the different P.1 sub-lineages further suggest that mutations of concern combined with human-behavior factors were responsible for the successful spread of the VOC P.1 in Brazil.

Declarations

Acknowledgments
The authors wish to thank all the health care workers and scientists who have worked hard to deal with this pandemic threat, the GISAID team, and all the EpiCoV database's submitters, GISAID acknowledgment table containing sequences used in this study is available in Supplementary Table S2. We also appreciate the support of the Fiocruz COVID-19 Genomic Surveillance Network (http://www.genomahcov.fiocruz.br/; accessed in May 2021) members, the Respiratory Viruses Genomic Surveillance. General Coordination of the Laboratory Network (CGLab), Brazilian Ministry of Health (MoH), Brazilian States Central Laboratories (LACEN).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Financial support was provided by Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) (PCTI-EmergeSaude/AM call 005/2020 and Rede Genômica de Vigilância em Saúde-REGESAM); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grant 402457/2020–0); CNPq/Ministério da Ciência, Tecnologia, Inovações e Comunicações/Ministério da Saúde (MS/FNDCT/SCTIE/Decit) (grant 403276/2020-9); Inova Fiocruz/Fundação Oswaldo Cruz (Grants VPPCB-007-FIO-18–2–30 and VPPCB-005-FIO-20–2–87), INCT-FCx (465259/2014–6) and Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) (26/210.196/2020). F.G.N, G.L.W, G.B and M.M.S are supported by the CNPq through their productivity research fellowships (306146/2017–7, 303902/2019–1, 302317/2017–1 and 313403/2018-0, respectively). G.B. is also funded by FAPERJ (Grant number E-26/202.896/2018).

Data availability statement

SARS-CoV-2 genome sequences generated in this study have been deposited in the GISAID platform (https://www.gisaid.org/), accession numbers IDs EPI_ISL_2038926 to EPI_ISL_2038968, EPI_ISL_2102018, and EPI_ISL_2102063.

References

1 Naveca, F. G. et al. COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. Nat Med (2021). <https://www.nature.com/articles/s41591-021-01378-7>.

2 Faria, N. R. et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science, doi:10.1126/science.abh2644 (2021).

3 McCormick, K. D., Jacobs, J. L. & Mellors, J. W. The emerging plasticity of SARS-CoV-2. Science 371, 1306-1308, doi:10.1126/science.abg4493 (2021).
4 Resende, P. C. et al. The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the Spike protein. *medRxiv*, doi:10.1101/2021.03.19.21253946 (2021).

5 Resende, P. C. et al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. *bioRxiv*, doi:10.1101/2020.04.30.069039 (2020).

6 Nascimento, V. A. D. et al. Genomic and phylogenetic characterisation of an imported case of SARS-CoV-2 in Amazonas State, Brazil. *Mem Inst Oswaldo Cruz* **115**, e200310, doi:10.1590/0074-02760200310 (2020).

7 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/msa010 (2013).

8 Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).

9 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589, doi:10.1038/nmeth.4285 (2017).

10 Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**, 1812-1819, doi:10.1093/molbev/msx116 (2017).

11 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007, doi:10.1093/ve/vew007 (2016).

12 Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016, doi:10.1093/ve/vey016 (2018).

13 Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**, 1185-1192, doi:10.1093/molbev/msi103 (2005).

14 Ferreira, M. A. R. S., M.A. Bayesian analysis of elapsed times in continuous-time Markov chains. *The Canadian Journal of Statistics* **36**, 355-368, doi: [https://doi.org/10.1002/cjs.5550360302](https://doi.org/10.1002/cjs.5550360302) (2008).

15 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Comput Biol* **5**, e1000520, doi:10.1371/journal.pcbi.1000520 (2009).

16 Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Syst Biol*, doi:10.1093/sysbio/syy032 (2018).
Lamarca, A. et al. Genomic surveillance of SARS-CoV-2 tracks early interstate transmission of P.1 lineage and diversification within P.2 clade in Brazil. medRxiv doi:10.1101/2021.03.21.21253418 (2021).

Kemp, S. A. et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature 592, 277-282, doi:10.1038/s41586-021-03291-y (2021).

Tegally, H. et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 592, 438-443, doi:10.1038/s41586-021-03402-9 (2021).

Cherian, S. et al. Convergent evolution of SARS-CoV-2 spike mutations, L452R, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. bioRxiv, doi:https://doi.org/10.1101/2021.04.22.440932 (2021).

Ferreira, I. et al. SARS-CoV-2 B.1.617 emergence and sensitivity to vaccine-elicited antibodies. bioRxiv, doi:https://doi.org/10.1101/2021.05.08.443253 (2021).

Martin, D. P. et al. The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. medRxiv, doi:10.1101/2021.02.23.21252268 (2021).

CDC. SARS-CoV-2 Variant Classifications and Definitions, <https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html> (2021).

GISAID. Tracking of Variants, <https://www.gisaid.org/hcov19-variants/> (2021).

Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID’s innovative contribution to global health. Glob Chall 1, 33-46, doi:10.1002/gch2.1018 (2017).

Figures
Figure 1

Characteristic mutations of P.1 and P.1-related lineages. Schematic representation of the genomic organization of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) showing the open-reading frames (ORFs), structural, and accessory proteins. The names of the genomic regions were indicated only where lineage-defining mutations (circles with one-letter amino acid code and the mutation position) were found.
Figure 2

Genetic diversity and distribution of the B.1.1.28, P.1 and P.1-like lineages in Brazil. a) Maximum likelihood (ML) phylogenetic tree of the B.1.1.28, P.1, and P.1-like lineages identified in Brazil. Each lineage was highlighted with colored boxes as indicated in the legend. The SH-aLRT support values are indicated in key branches, and branch lengths are drawn to scale with the lateral bar indicating nucleotide substitutions per site. Nodes representing the most recent common ancestor (MRCA) of each lineage and the MRCA of all P.1 and P.1-related viruses (P.1MRCA1), and the MRCA of P.1 and P.1-like-II (P.1MRCA2) are highlighted with circles. b) Geographic distribution and frequency of the P.1-like-II lineage identified in
Brazil. Brazilian states' names follow the ISO 3166-2 standard. Color's gradient represents the number of sequences identified in this study. c) Correlation between the sampling date of B.1.1.28, P.1, P.1-like-I, and P.1-like-II and their genetic distance from the ML phylogenetic tree's root. Each lineage was colored following the legend. The slope of each regression is indicated.

Figure 3
Evolutionary steps associated with the emergence of P.1 and P.1-related lineages. Colored squares represent the node where the mutation emerged and was fixed during the diversification of the B.1.1.28 lineage in Brazil originating the P.1, P.1-like-I, and P.1-like-II lineages. Nodes' colors and topology are described in Figure 2a. The genomic position of the polymorphism is indicated at the top and the amino acid change at the bottom. Mutations of concern are in red. IGR: Intergenic region.

**Figure 4**

Bayesian phylogeographic analysis of the B.1.1.28, P1, and P1-related lineages. Tips and branches' colors indicate the Brazilian state (ISO 3166-2 standard) of sampling and the most probable inferred location of their descendent nodes, respectively, as indicated in the legend. Branch posterior probabilities are indicated in key nodes. Boxes with different colors highlight the 28-AM-II, P1, P1-like-I, and P1-like-II
lineages. All horizontal branch lengths are time-scaled, and the tree was automatically rooted under the assumption of the strict molecular clock model. Reconstructed ancestral key nodes representing the most recent common ancestor (MRCA) of each lineage and the MRCA of all P.1 and P.1-related viruses (labeled as P.1MRCA1) and the MRCA of P.1 and P.1-like-II (labeled as P.1MRCA2) are highlighted with circles.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- GISAIDAcknowledgementTableS2.pdf
- SupplementaryTable1.docx