



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



Survey of SARS-CoV-2 genetic diversity in two major Brazilian cities using a fast and affordable Sanger sequencing strategy

Erick Gustavo Dorlass^{a,1}, Karine Lima Lourenço^{b,1}, Rubens Daniel Miserani Magalhães^{b,1}, Hugo Sato^b, Alex Fiorini^b, Renata Peixoto^b, Helena Perez Coelho^b, Bruna Larotonda Telezynski^a, Guilherme Pereira Scagion^a, Tatiana Ometto^a, Luciano Matsumiya Thomazelli^a, Danielle Bruna Leal Oliveira^c, Ana Paula Fernandes^b, Edison Luiz Durigon^a, Flavio Guimarães Fonseca^{b,*}, Santuza Maria Ribeiro Teixeira^{b,*}

^a Universidade de São Paulo, São Paulo, Brazil

^b Centro de Tecnologia de Vacinas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

^c Hospital Israelita Albert Einstein, São Paulo, Brazil

ARTICLE INFO

Keywords:

Sars-Cov-2
genetic variants
Spike
Sanger sequencing

ABSTRACT

Genetic variants of SARS-CoV-2 have been emerging and circulating in many places across the world. Rapid detection of these variants is essential since their dissemination can impact transmission rates, diagnostic procedures, disease severity, response to vaccines or patient management. Sanger sequencing has been used as the preferred approach for variant detection among circulating human immunodeficiency and measles virus genotypes. Using primers to amplify a fragment of the SARS-CoV-2 genome encoding part of the Spike protein, we showed that Sanger sequencing allowed us to rapidly detect the introduction and spread of three distinct SARS-CoV-2 variants in two major Brazilian cities. In both cities, after the predominance of variants closely related to the virus first identified in China, the emergence of the P.2 variant was quickly followed by the detection of the P1 variant, which became dominant in less than one month after it was first detected.

1. Introduction

After the announcement of the SARS-CoV-2 pandemic in March of 2020, different virus variants have emerged and spread fast, many of them becoming dominant over previous ones. These variants have been categorized as Variants of Interest (VOI) or Variants of Concern (VOC) [1]. Although the appearance of new variants is expected, as SARS-CoV2 continues to evolve and differentiate in response to different host-derived selective pressures, only four of the currently identified variants are considered as VOC: B.1.1.77 (UK/Alpha variant), P.1 (Brazil/Gamma variant), B.1.351 (South Africa/Beta variant) and B.1.617.2 (India/Delta variant). These four variants have specific amino acid substitutions in the Spike protein, two of them known to be associated with increased infection and mortality: E484K and N501Y [2–4]. Also, some studies have shown a decreased in the immune response when convalescent serum and immunized serum from vaccinated individuals

were tested against variants with these mutations [5–8].

The P.1 variant, which is thought to have evolved from the B.1.1.28 lineage in a period of fast molecular evolution dated around mid-November of 2020, is of particular interest in this study [9]. The P.1 emergence has been unequivocally traced to the city of Manaus, and it is thought to have been one of the main factors responsible for the upsurge of COVID-19 cases and deaths in this city during December 2021/January 2021 period, leading to the complete collapse of the local health system. The P.1 genetic background includes 21 missense, lineage-specific mutations, 10 of which are located within the Spike protein (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y and T1027I) [10]. Circulating undetected at least until mid-February 2021, the mutant rapidly spread to other Brazilian cities, most likely due to infections carried by people fleeing the collapsed Amazon city. Different from P.1, the P.2 variant bears the E484K mutation only, but not the other two mutations of concern, N501Y and K417T. Four other lineage-

* Corresponding authors at: Centro de Tecnologia de Vacinas – CTVacinas/UFMG, Rua Professor José Vieira de Mendonça, 770, Engenho Nogueira, Belo Horizonte, MG 31310-260, Brazil.

E-mail addresses: fdafonseca@icb.ufmg.br (F.G. Fonseca), santuzat@ufmg.br (S.M.R. Teixeira).

¹ These authors contributed equally to this work.

<https://doi.org/10.1016/j.ygeno.2021.10.015>

Received 29 June 2021; Received in revised form 28 September 2021; Accepted 24 October 2021

Available online 28 October 2021

0888-7543/© 2021 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

defining mutations were also found in other parts of the genome, including UTR, Orf1ab and N. Genomic surveillance of SARS-CoV-2 is an essential component of the epidemiological profile of the COVID-19 pandemic since variant identification can be associated with new outbreaks and re-infection. Fast and precise detection of these variants is crucial for public health officials to monitor the impact caused by their introduction in the population of circulating viruses so that control measures can be adjusted accordingly. The high cost of complete genomic sequencing still prevents low/medium-income countries to properly perform such surveillance using Next-Generation Sequencing (NGS) platforms. The more traditional Sanger Sequencing may be an alternative for identification of circulating variants since it can provide results very rapidly and be performed using a sequencing infrastructure already in place in many laboratories. Developed in 1977 by Frederick Sanger, also known as the “chain-termination” method, it is based on selective incorporation of dideoxynucleotides by DNA polymerase which results in fragments of various lengths that terminated with a fluorescent labelled dideoxynucleotides at the 3' end. Although NGS methods have unquestionable advantages since it allows the determination of the complete sequence of a fragmented genome in few hours, in contrast to Sanger sequencing, which generates sequences from only a small number of fragments, implementation of NGS requires a large investment in both equipment and technical training of lab personnel.

Using the Sanger method to determine the sequence of a 1006 bp region of the SARS-CoV-2 genome we were able to detect circulating VOI and VOCs from samples collected in two major cities in Brazil during a period of 12 months. São Paulo, with an estimated population of 12,325,232 [11], and Belo Horizonte, with 2,521,564 citizens [12] are the first and sixth largest cities in Brazil and have reported a total of 1,048,061 [13] and 177,432 [14] cases of COVID-19 as of April 30.

2. Methods

2.1. Ethical statement

Part of study involving samples from São Paulo city was approved by Instituto de Ciências Biomédicas ethical committee. CAAE number: 32712620.9.0000.5467. The study involving samples from Belo Horizonte was approved by the UFMG Ethics Committee, CAAE-35074720.3.0000.5149.

2.2. Sample collection and SARS-CoV-2 identification

Nasopharyngeal swabs from patients with airway infection symptoms were collected from seven hospitals in the city of São Paulo, SP as well as from a specialized COVID-19 Health Center (UPA Centro-Sul) in Belo Horizonte, MG, Brazil. Part of the samples from the Belo Horizonte metropolitan area, which includes the cities of Betim, Contagem, Ribeirão das Neves e Santa Luzia, were obtained from Fundação Ezequiel Dias (Funed-MG). RNA was extracted either with Viral Pathogens kit (ThermoFisher) in MagMax (São Paulo samples) or with the QIAamp Viral RNA Mini Kit (Qiagen, USA) (Belo Horizonte samples), according to protocols provided by the manufacturers. RT-qPCR was performed using primers and probes described in the Berlin (Charité/Berlin) protocol [15], targeting the gene E from SARS-CoV-2 and the human RNase P mRNA, used as endogenous amplification reaction control. All samples tested positive were selected for gene amplification with PCR primers targeting the Spike gene and sequencing using the Sanger method.

2.3. Primer selection and PCR amplification

Primers specific for the amplification of a 1006 bp region, codifying part of subunit one of the Spike protein of SARS-CoV-2, was selected in the ARTIC protocol [16]. The primers are shown in **Supplementary Table 1**. The 1006 bp amplified region encodes amino acids K417, L452, T478, E484, N501 and A570, which have been associated with the

characterization of several VOI and the VOCs. The same pair of primers together with a third primer were used for all sequencing reactions. PCR amplifications were performed after Reverse Transcription reactions using random primers and the High-Capacity cDNA Reverse Transcription Kit (thermoFisher) following manufacturer's instructions. PCR amplifications were performed with the Invitrogen Platinum Taq DNA Polymerase kit and the following amplification protocol: 5 min at 95°C followed by 40 cycles of 30 s at 94°C, 30 s at 56°C and 80 s at 72°C with a final extension time of 5 min at 72°C. The amplicons were analyzed by 1% agarose gel electrophoresis. After treatment with Exosap-it (ThermoFisher), sequencing reactions were performed with the BigDye™ Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and either the ABI 3730 (Belo Horizonte samples) or the ABI PRISM 3130XL Genetic Analyzer (São Paulo samples) according to manufacturer's instructions.

2.4. Sequence analyses

Assembling of Sanger contigs were performed with the GeneStudio software [17]. Each primer corresponding sequence was trimmed in a fixed position (40 and 450) followed by manual evaluation of aligned chromatograms for each sample. By this approach 228 contigs were constructed. These sequences were aligned by the MAFFT tool [18,19] with default parameters: 0 “Offset”, 1.53 “gap open penalty”, 0 “Maximum number of iterative refinements”. Benchling platform [20] and AliView software [21] were used for visualization of alignments. The samples were manually genotyped by evaluation of the modifications of interest in the contigs (K417N/T, L452R, T478K, E484K, N501Y, A570D), after *in silico* translation of the consensus nucleotide sequences using standard genetic code. Stacked Bar Plots were constructed with the R ggplot2 package.

Next-Generation Sequencing (NGS) was performed with 16 samples using the Ion Torrent Platform with SARS-CoV-2 Ion Ampliseq panel. Sample libraries were constructed on the Ion Chef system with the Ion Ampliseq Chef DL8 kit. Sequencing was performed in two 530 Ion Chip with a capacity of 20 million reads of each chip. Sequencing analysis was performed on Torrent server with IRMA plugin for read assembly and SnpEff for variant annotation. All final sequences were submitted to Pangolin [22] for lineage assignment.

To construct the dendrograms, the aligned Sanger contigs were trimmed in 5' and 3' end to eliminate all gaps and divergent sequencing quality. Nine short contigs (PAC1346B, COV958, HIDV3706, PAC1705, PAC1803, ECO207, PAC721, COV701, HIDV4996) were excluded to maximize the size of multiple alignment after trimming. An alignment of 770 nt was obtained after these steps. The genomic sequences were aligned using the MAFFT tool using default parameters. The phylogenetic trees were constructed using the iqtree tool [23], with the “Maximum likelihood” statistical approach and substitution model TPM3 + F + I to Sanger and GTR + F + I NGS dataset. Lineage B (GISAID_ID: EPI_ISL_402123, 2019) was defined as “outgroup” and the R gtree package [24] was used for the tree visualization.

3. Results

A total of 228 SARS-CoV-2 positive samples were successfully amplified using the primers shown in **Fig. 1**. The amplicons correspond to viruses present in positive samples from 130 infected individuals from the city of São Paulo collected between March/2020 and Abril/2021 and 98 positive samples from individuals from the Belo Horizonte metropolitan area collected between August/2020 and Abril/2021. As shown in **Table 1**, mutations in the 1006 bp amplified region corresponding to amino acids positions 417, 484, 501 and 570 of the Spike protein allowed the detection of several VOIs and VOCs. Because it was not possible to discern between the P.2 and N.9 variants, based on the sequencing of the 1006 bp amplicon, we also amplified an additional region of 969 bp, containing the V1176F mutation, allowing us to

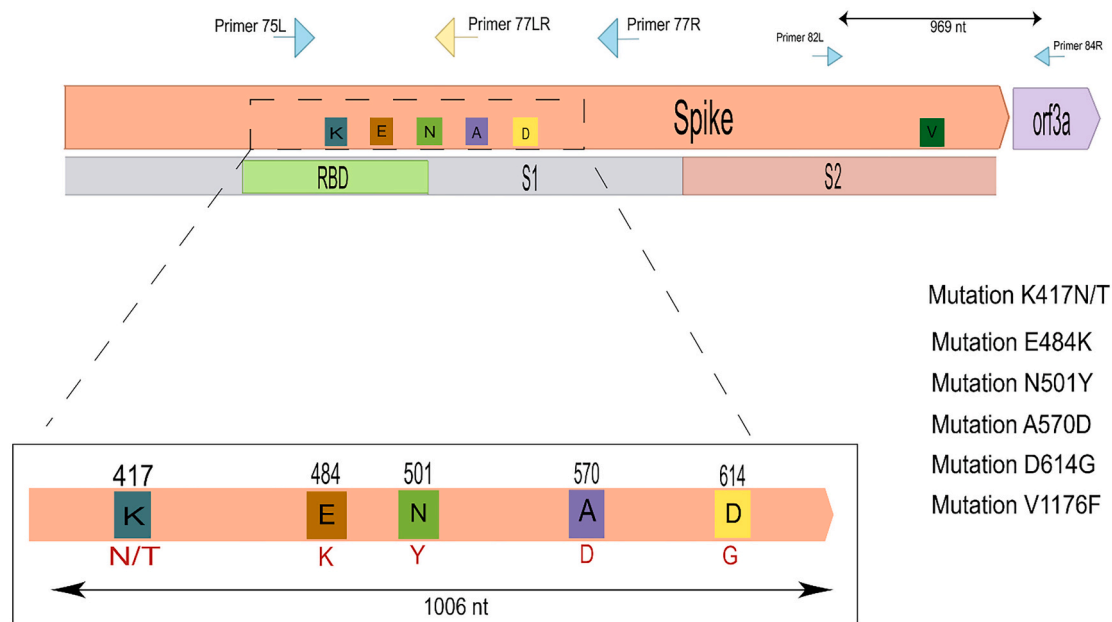


Fig. 1. Schematic representations of the Spike protein and the 1006 bp amplicon. PCR amplifications were performed with primers that were used both for PCR and Sanger sequencing (blue arrows). Yellow arrows indicate primers used only for sequencing. Colored rectangles represent changes in amino acids found in different VOI and VOC's and their respective positions in the SARS-CoV-2 genome, Lysine (K) blue, Glutamic acid (E) orange, Asparagine (N) green, Alanine (A) purple, Aspartic acid (D) yellow, Valine (V) dark green. It is also represented the 969 bp PCR amplicon generated with primers annealing within the spike and orf3a protein genes (small blue arrows), which allowed us to verify for the presence or absence of a mutation at position 1176 that differs in the VOIs N.9 and P.2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

Amino acid changes found in circulating variants that can be identified by sequencing the amplified fragments.

Variants	Mutations in the amplified fragment
P.1 (VOC)	K417T, E484K, N501Y
P.2 (VOI)	E484K
B.1.1.7 (VOC)	N501Y, A570D
B.1.351 (VOC)	K417N, E484K, N501Y
B.1.617.2 (VOC)	L452R, T478K

discriminate between these two variants (Fig. 1). Five samples classified as P.2/N.9 were sequenced and all of them had the V1176F mutation and was then classified as P.2 (data not shown). Based on the sequences of the amplified 1006 bp fragment generated from all 228 samples collected from individuals in São Paulo and Belo Horizonte, we were able to detect, besides the lineage that is closely related to the original strain isolated in Wuhan (B.1) the B.1.1.7 variant, which have been previously identified in England as well as the P.1 and P.2/N.9 variants which have been previously identified in Brazil. Fig. 2 showed that, while in São Paulo city, the P.2/N.9 variant appeared in September 2020 (Fig. 2A and C), in Belo Horizonte, the same variant was first detected in November (Fig. 2B and D). The emergence of the P.2 variant was announced in December, after being identified in the state of Rio de Janeiro [25] and rapidly spread throughout the country. Our data showed that this lineage was actually circulating in the neighbor state of São Paulo at least 3 months before this announcement and was already present in Belo Horizonte, also a large city close to Rio de Janeiro, one month before. Fig. 2 also showed that, in the two cities, the P.2 variant was rapidly replaced by the P.1 variant, which was initially described in the city of Manaus, more than 3000 Km distant from São Paulo and Belo Horizonte. In both cities, the P.1 variant was first detected in February. It is interesting to note that in both cities, one month after its appearance, the P.1 variant became largely predominant among all other circulating lineages, being identified as 75% and 86% of all samples analyzed in Belo Horizonte and São Paulo, respectively, in March 2021.

Whereas in Belo Horizonte, the P.1 variant corresponded to all samples collected in the month of April, one sample corresponding to the P.2 variant was still detected in April in São Paulo city. It is also noteworthy that, in contrast to the P.2 variant, which dominated over the B.1 lineage, and the P.1 variant, which rapidly dominated over the P.2 variant, the appearance of the B.1.1.77 variant in both cities during the month of March did not result in a similar, fast dissemination. Because the B.1.1.77 variant was introduced in both cities after the P.1 variant was already in place, it can be assumed that its dissemination capacity is significantly lower compared to P.1.

Phylogenetic and molecular evolution analyses were performed using nucleotide sequences derived from 219 samples. Nine of the 228 samples were not included in this analysis because the contig size became too short after trimming the sequence extremities. Although these sequences also contain mutations of interest, they were excluded from the phylogenetic analyses because the differences in size cause a considerable impact on the multiple alignment. As shown in Fig. 3A, all generated sequences were grouped in four main clusters corresponding to the B.1 lineage and the B.1.1.7, P.1 and P.2/N.9 variants. This analysis also confirms that the B.1.1.7 was derived from the B.1 lineage, as previously described, and also indicated that the P.1 and P.2/N.9 variants have emerged independently in Brazil.

To confirm that the sequence obtained from 1006 bp fragment of the Spike gene was sufficient to unambiguously ascertain the identity of the variants, we selected 16 samples to determine their complete genome sequences using the Ion Torrent Sequencing Platform. The viral samples that have their complete genomes sequenced were indicated by an arrow on Fig. 3A. The phylogenetic tree shown in Fig. 3B indicates that, for all genomes that were fully sequenced, the classification based on partial Sanger sequencing as belonging to B.1 lineage and the B.1.1.7, P.1 and P.2/N.9 variants could be fully validated by the complete NGS genome analyses.

After comparing the cost and the turnaround time required to perform the analyses using Sanger sequencing and NGS, we showed that the Sanger method is highly advantageous. As shown in Supplementary Table 2, the time required to analyze 16 samples is less than 15 h when

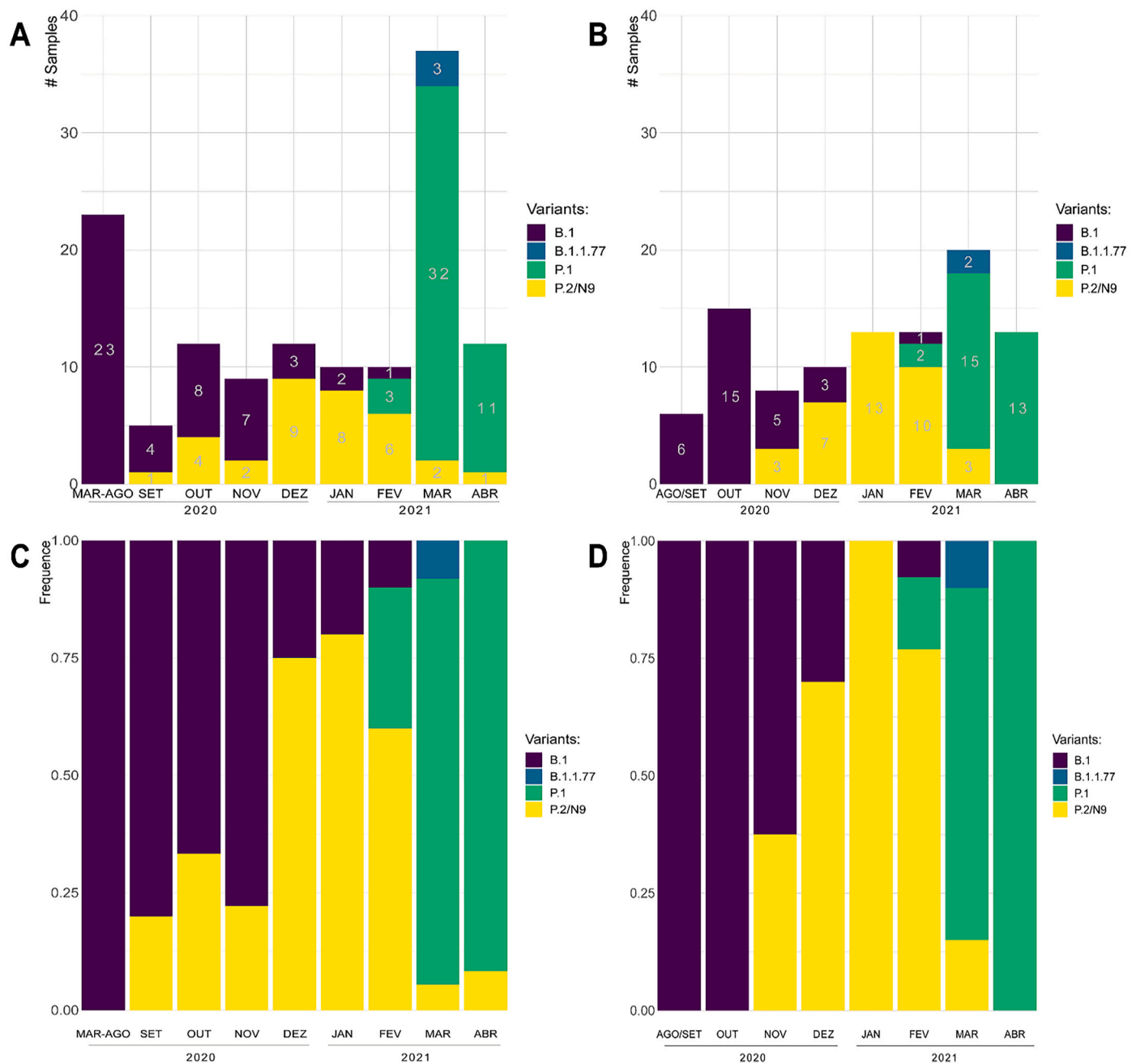


Fig. 2. Stacked bar plots of the SARS-CoV-2 variants genotyped by Sanger sequencing in the cities of São Paulo (A, C) and in the metropolitan region of Belo Horizonte (B, D). Colors of the bars represent the number of samples (A, B) or the frequency (C, D) of the variants by period: purple, B.1; blue, B.1.1.77; green, P.1; yellow P.2 / N.9. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

using Sanger sequencing compared to 51 h for NGS. Compared to NGS, the cost per sample is approximately 20 times lower (\$14.59) for Sanger sequencing. Similar strategy has been described by Jørgensen et al., (2021) with an estimated cost of 4€ per sample and a turnaround time of less than 30 h [26].

Using Sanger sequencing, we have also analyzed two samples that were collected during the third week of May 2021, at the international airport of Belo Horizonte. These samples were from RT-PCR-positive individuals who had contact with a flight crew with Indian origin. In less than 5 days, we were able to amplify the 1006 bp fragment, sequence and verify that in both cases, they correspond to the P.1 variant, and not to the new emerging variant B.1.617.2, which has the two distinct mutations in this region. Taken together, these results showed the power of the Sanger sequencing strategy to rapidly generate

data that can be used to perform genomic surveillance during the COVID-19 pandemic. It also confirmed that it can be employed for genomic epidemiology studies of other diseases. With a relatively small number of samples, we showed that the introduction of VOI or VOCs in two large urban areas can rapidly reshape the genomic profile, which may have a significant impact in the course of the COVID-19 pandemic.

4. Discussion

In the first months of the SARS-CoV2 pandemics it became clear that ample qPCR testing and contact tracing would be essential to mitigate local and nationwide virus spread, limiting overall impacts of the disease [27,28]. Indeed, countries that exceeded in testing their population had relative success in diminishing the burden of the epidemics, both in

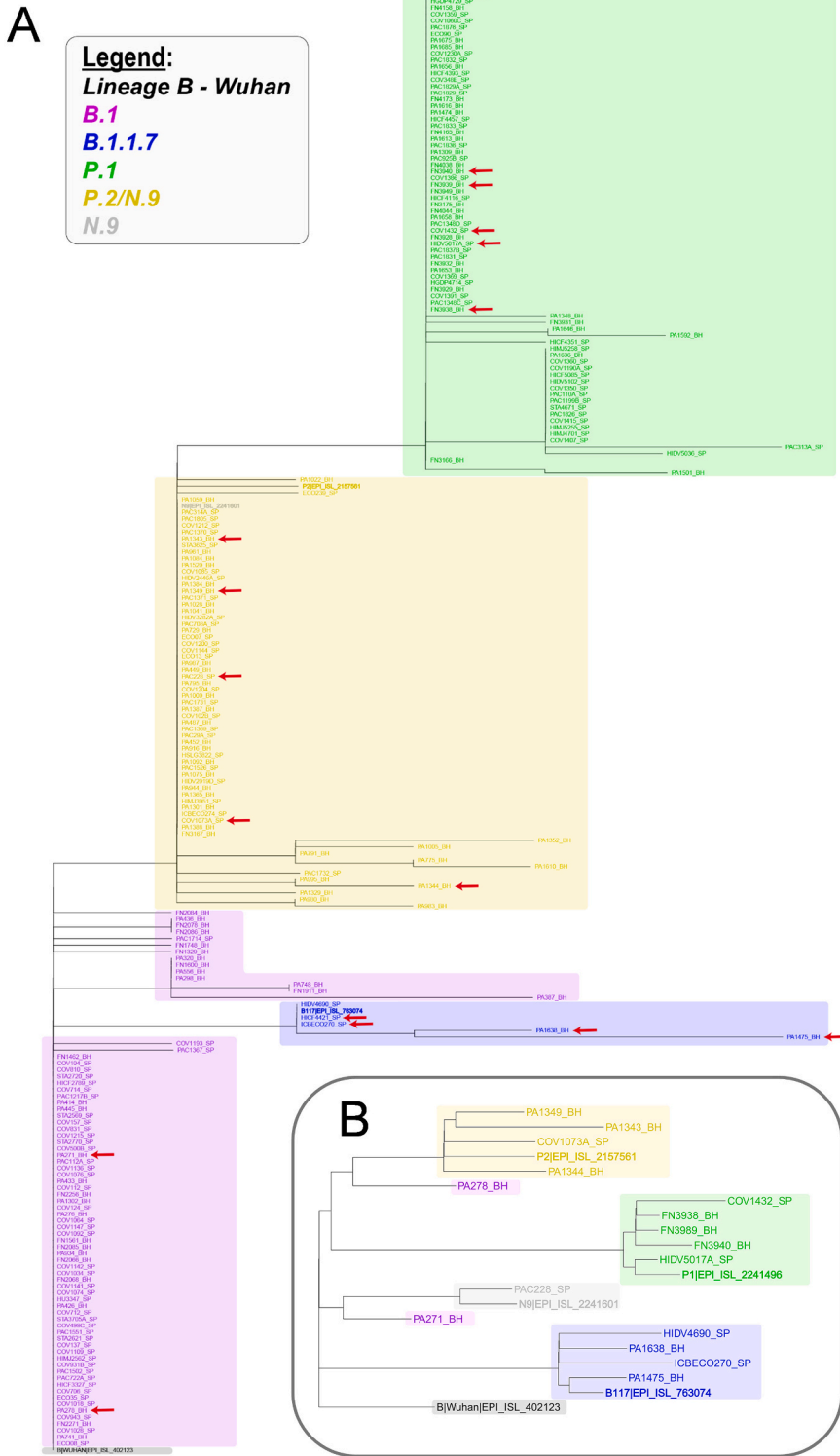


Fig. 3. Dendrogram showing the relationship between the sequences obtained by Sanger (A) and by NGS sequencing of the complete genome of some of these samples (B). Samples whose complete genome has been sequenced are indicated in A by the red arrow. The genotyped sequences of each of the variants of interest are represented in colors: purple, B.1; blue, B.1.1.7; green, P.1; yellow, P.2 / N.9. The dendrograms were rooted using the corresponding 770 bp region of Lineage B (Wuhan, EPI_ISL_402123,2019). Reference sequences for each variant of interest are highlighted in bold. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

terms of number of affected people as well as in social and economic impacts [29,30]. Nonetheless, over and over the new coronavirus has proven to be an elusive enemy to defeat. One of the most difficult challenges related to the current moment of the pandemic is certainly to identify and contain the spread of virus variants presenting mutations that confer dangerous selective advantages to them. Again, widespread testing is essential to this purpose. However, different from the simpler RT-qPCR-based diagnostic assay to detect SARS-CoV2 RNA in suspected patients, detecting the occurrence and circulation of genetic variants is a

much more complex task. To date, the monitoring of viral variants' arising and circulation is done almost exclusively through NGS [31–33]. That can be a huge constraint for the global monitoring of SARS-CoV2 mutants, as equipment and trained personal to perform the technique and analyze the results are not broadly available, especially in developing and under-developed countries where the epidemics soar out of control. Although it is not a suitable method to identify new variants, since complete genome sequences are necessary for the characterization of a new SARS-CoV-2 variants, partial genome sequence of the Spike

gene using the Sanger protocol is a fast and cost-effective way to unambiguously assign a viral sequence to one of the previously characterized variants.

Two recent and somewhat opposed examples illustrate the importance of early detection of SARS-CoV2 variants to avoid their uncontrolled spread. The B.1.617 variant, usually called Indian variant, was first reported in late 2020 at the Maharashtra Indian state [34], and today has spread to more than 40 countries. After the first identification of B.1.617 variant, three subtypes with slightly different set of mutations have been described through a massive effort using NGS to track the spread and evolution of this new variant [35]. The early detection of this VOC harboring a worrisome set of mutations previously related to increased virus infectivity and partial escape from antibody neutralization has helped to create an important state of preparedness abroad. In UK, coordinated efforts to monitor the spread of the Indian variants through massive acquisition of NGS data has helped scientists and sanitary authorities to model the new mutants' spread within the country. This network has been able to determine, for example, that the B.1.617.2 subtype has outcompeted the other two Indian B.1.617 subtypes and now seems to be replacing the former prevalent B.1.1.7 variant [35].

The second example comes for the P.1 VOC arising in Brazil in November 2020. The P.1 harbors a plethora of lineage-defining mutations, including some particularly worrisome modification in the S protein that have been related to increased transmissibility (K417T, E484K, and N501Y) and neutralization antibodies' partial escape (E484K) [9,36]. The rise of this new VOC has been pointed out as one of the pivotal reasons for the recent and overwhelming wave of SARS-CoV2 cases in the city of Manaus, the capital of the Amazon State, leading to a full collapse of the city's health system [10]. This shocking wave of infections and deaths took place despite the fact that most of the city's population was thought to be immune after a massive first wave of COVID-19 hit the city in March/April 2020 [36]. Still, the new P.1 variant circulated undetected for months until it was described in four Japanese travelers returning from Manaus to Japan [37]. Only then, tipped by that finding, scientists were able to detect the virus circulating in Manaus through networking NGS efforts [10,36]. By then, the P.1 variant had already spread from Manaus and quickly became the prevalent variant circulating in many other Brazilian regions [38].

If the P.1 variant had been detected earlier, it is possible to speculate that the second large COVID-19 wave hitting Brazil could have been at least partially attenuated. And it is also possible to speculate that constraints in doing massive NGS throughout the country can be partially responsible for such failure. Sanger nucleotide sequencing is much easier to execute and analyze and largely available in different research laboratories than NGS. When tailored to focus in genomic regions bearing signature mutations, the sequence length limitation of the Sanger sequencing strategy ceases to be a problem and become an advantage, as a much larger number of samples can be sequenced in a shorter period of time. The data present here showed that the method was able to accurately detect the appearance of the P.1 VOC in two of the largest Brazilian cities and was also able to unveil the strong capacity of the P.1 variant to disseminate and quickly replace the P.2 variant, which was prevalent lineage in both cities. Although beyond the scope of this study, a better understanding of the molecular mechanism behind the increased dissemination capacity as well as the implications of the spreading of distinct variants regarding vaccination and other methods for the control of the pandemic are certainly necessary. Further efforts towards mitigation of new COVID-19 waves may include studies that rely on a broad implementation of Sanger sequencing genotyping strategy such as the one described here.

Data availability

Sequences generated using the Sanger method are available in GenBank [39] with identifiers **MZ410314** to **MZ410541**, and sequences generated using the Ion Torrent Platform are available in GISAID data

bank [40] with identifiers **EPI_ISL_2551523** to **EPI_ISL_2551536**, and **EPI_ISL_2627737**, **EPI_ISL_2627738**.

Financial support

This work was supported by the Brazilian Ministry of Science, Technology and Innovation (MCTI) through the "Rede Virus" initiative and its many individual projects: Sub-rede Corona-ômica, sub-rede Diagnóstico, sub-rede Laboratórios de Campanha and sub-rede Vacinas. The research leading to these results also received funding from the Secretaria de Ensino Superior do Ministério da Educação [grant number SEI 23072.211119/2020-10], from FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), Finep (Financiadora de Estudos e Projetos), from Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES), from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), from FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) [Project 2020/06409-1] as well as from the INCTV (Instituto Nacional de Ciência e Tecnologia de Vacinas), Prefeitura de Belo Horizonte, the Sociedade Beneficente Israelita Brasileira Hospital Albert Einstein and Associação Brasileira de gestão em projetos. GPS, EGD, BLT, FGF, SMRT, APF are recipients of CNPq fellowships.

Acknowledgments

We thank Dr. Renato S. Aguiar, Dr. Betânia P. Drumond and Dr. Jônatas S. Abrahão for technical discussions, Dr. Unaí Tupinambás, Dr. Claudia R. L. Alves, Dr. Elaine L. Machado and MSc Murilo S. Costa for setting up the PBH-UPA cohort. We also thank Dr. Pedro Augusto Alves and Carolina P. Rocha for sequencing reactions performed at the Plataforma de Sequenciamento Capilar do Centro de Pesquisa Rene Rachou and Igor Pereira Godinho for administrative support. We also thanks Dr. Milena De Flavia from University Hospital at São Paulo University (HU-USP); Dr. Flávia Jaqueline Almeida from Santa Casa da Misericórdia Hospital (SCMH); Dr. Andressa Simões Aguiar from Hospital São Luiz Gonzaga (HSLG), Infant Hospital Candido Fontoura (IHCF) and Geriatric Hospital Dom Pedro; Dr. Luciana Becker Mau from Pediatric Hospital Menino Jesus (PHMJ) and Dr. Carolina Sucupira from Infant Hospital Darcy Vargas.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ygeno.2021.10.015>.

References

- [1] F. Konings, M.D. Perkins, J.H. Kuhn, M.J. Pallen, E.J. Alm, B.N. Archer, A. Barakat, T. Bedford, J.N. Bhiman, L. Caly, L.L. Carter, A. Cullinane, T. de Oliveira, J. Druce, I. El Masry, R. Evans, G.F. Gao, A.E. Gorbalenya, E. Hamblion, B.L. Herring, E. Hodcroft, E.C. Holmes, M. Kakkar, S. Khare, M.P.G. Koopmans, B. Korber, J. Leite, D. MacCannell, M. Marklewitz, K.C. Maurer-Stroh, J.A.M. Rico, V.J. Munster, R. Neher, B.O. Munnink, B.I. Pavlin, M. Peiris, L. Poon, O. Pybus, A. Rambaut, P. Resende, L. Subissi, V. Thiel, S. Tong, S. van der Werf, A. von Gottberg, J. Ziebuhr, M.D. Van Kerkhove, SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse, *Nat. Microbiol.* (2021), <https://doi.org/10.1038/s41564-021-00932-w>.
- [2] N.G. Davies, C.I. Jarvis, W.J. Edmunds, N.P. Jewell, K. Diaz-Ordaz, R.H. Keogh, Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7, *Nature* 593 (2021) 270–274, <https://doi.org/10.1038/s41586-021-03426-1>.
- [3] H.H. Gan, A. Twaddle, B. Marchand, K.C. Gunsalus, Structural modeling of the SARS-CoV-2 Spike/human ACE2 complex interface can identify high-affinity variants associated with increased transmissibility, *J. Mol. Biol.* 433 (2021) 167051, <https://doi.org/10.1016/j.jmb.2021.167051>.
- [4] A. Khan, T. Zia, M. Suleman, T. Khan, S.S. Ali, A.A. Abbasi, A. Mohammad, D. Wei, Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: an insight from structural data, *J. Cell. Physiol.* (2021), <https://doi.org/10.1002/jcp.30367>.
- [5] S. Cele, I. Gazy, L. Jackson, S.-H. Hwa, H. Tegally, G. Lustig, J. Giandhari, S. Pillay, E. Wilkinson, Y. Naidoo, F. Karim, Y. Ganga, K. Khan, M. Bernstein, A.B. Balazs, B. I. Gosnell, W. Hanekom, M.-Y.S. Moosa, R.J. Lessells, T. de Oliveira, A. Sigal,

- Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma, *Nature* 593 (2021) 142–146, <https://doi.org/10.1038/s41586-021-03471-w>.
- [6] S. Jangra, C. Ye, R. Rathnasinghe, D. Stadlbauer, F. Krammer, V. Simon, L. Martinez-Sobrido, A. Garcia-Sastre, M. Schotsaert, H. Alshammary, A. A. Amoako, M.H. Awawda, K.F. Beach, M.C. Bermúdez-González, R.L. Chernet, L. Q. Eaker, E.D. Ferreri, D.L. Floda, C.R. Gleason, G. Kleiner, D. Jurczynszak, J. C. Matthews, W.A. Mendez, L.C.F. Mulder, K.T. Russo, A.-B.T. Salimbangon, M. Saksena, A.S. Shin, L.A. Sominsky, K. Srivastava, SARS-CoV-2 spike E484K mutation reduces antibody neutralisation, *Lancet Microbe* (2021), [https://doi.org/10.1016/S2666-5247\(21\)00068-9](https://doi.org/10.1016/S2666-5247(21)00068-9).
- [7] P. Supasa, D. Zhou, W. Dejnirattisai, C. Liu, A.J. Mentzer, H.M. Ginn, Y. Zhao, H.M. E. Duyvesteyn, R. Nutalai, A. Tuekprakhon, B. Wang, G.C. Paesen, J. Slon-Campos, C. López-Camacho, B. Hallis, N. Coombes, K.R. Bewley, S. Charlton, T.S. Walter, E. Barnes, S.J. Dunachie, D. Skelly, S.F. Lumley, N. Baker, I. Shaik, H.E. Humphries, K. Godwin, N. Gent, A. Sienkiewicz, C. Dold, R. Levin, T. Dong, A.J. Pollard, J. C. Knight, P. Klenerman, D. Crook, T. Lambe, E. Clutterbuck, S. Bibi, A. Flaxman, M. Bittaye, S. Belij-Rammerstorfer, S. Gilbert, D.R. Hall, M.A. Williams, N. G. Paterson, W. James, M.W. Carroll, E.E. Fry, J. Mongkolsapaya, J. Ren, D. I. Stuart, G.R. Screaton, Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera, *Cell* 184 (2021), <https://doi.org/10.1016/j.cell.2021.02.033>, 2201–2211.e7.
- [8] X. Xie, Y. Liu, J. Liu, X. Zhang, J. Zou, C.R. Fontes-Garfias, H. Xia, K.A. Swanson, M. Cutler, D. Cooper, V.D. Menachery, S.C. Weaver, P.R. Dormitzer, P.-Y. Shi, Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera, *Nat. Med.* 27 (2021) 620–621, <https://doi.org/10.1038/s41591-021-01270-4>.
- [9] N.R. Faria, T.A. Mellan, C. Whittaker, I.M. Claro, D.S. da Candido, S. Mishra, M.A. E. Crispim, F.A.S. Sales, I. Hawrylyuk, J.T. McCrone, R.J.G. Hulsvit, L.A.M. Franco, M.S. Ramundo, J.G. de Jesus, P.S. Andrade, T.M. Coletti, G.M. Ferreira, C.A. M. Silva, E.R. Manuli, R.H.M. Pereira, P.S. Peixoto, M.U.G. Kraemer, N. Gaburo, C. C. da Camilo, H. Hoeltgebaum, W.M. Souza, E.C. Rocha, L.M. de Souza, M.C. de Pinho, L.J.T. Araujo, F.S.V. Malta, A.B. de Lima, J.P. do Silva, D.A.G. Zauli, A.C. S. de Ferreira, R.P. Schnekenberg, D.J. Laydon, P.G.T. Walker, H.M. Schlüter, A.L. P. dos Santos, M.S. Vidal, V.S. Del Caro, R.M.F. Filho, H.M. dos Santos, R.S. Aguiar, J.L. Proença-Modena, B. Nelson, J.A. Hay, M. Monod, X. Miscouridou, H. Coupland, R. Sonabend, M. Vollmer, A. Gandy, C.A. Prete, V.H. Nascimento, M. A. Suchard, T.A. Bowden, S.L.K. Pond, C.-H. Wu, O. Ratmann, N.M. Ferguson, C. Dye, N.J. Loman, P. Lemey, A. Rambaut, N.A. Fraiji, M.P.S.S. do Carvalho, O. G. Pybus, S. Flaxman, S. Bhatt, E.C. Sabino, Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil, *Science* 372 (80) (2021) 815–821, <https://doi.org/10.1126/science.abb2644>.
- [10] F.G. Naveca, V. Nascimento, V.C. de Souza, A.L. de Corado, F. Nascimento, G. Silva, A. Costa, D. Duarte, C. Pessoa, M. Mejía, M.J. Brandão, M. Jesus, L. Gonçalves, C.F. da Costa, V. Sampaio, D. Barros, M. Silva, T. Mattos, G. Pontes, L. Abdalla, J.H. Santos, I. Arantes, F.Z. Dezordi, M.M. Siqueira, G.L. Wallau, P. C. Resende, E. Delatorre, T. Gräf, G. Bello, COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence, *Nat. Med.* (2021), <https://doi.org/10.1038/s41591-021-01378-7>.
- [11] IBGE - Instituto de Geografia e Estatística, Estimate of the Population of City of São Paulo in 2020. <https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama>, 2021.
- [12] IBGE - Instituto de Geografia e Estatística, Estimate of the Population of City of Belo Horizonte in 2020. <https://cidades.ibge.gov.br/brasil/mg/belo-horizonte/panorama>, 2021.
- [13] Informative COVID-19 Bulletin of the São Paulo City (2021/04/30), São Paulo. https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/saude/20210430_boletim_covid19_diario.pdf, 2021.
- [14] Informative COVID-19 Bulletin of the Belo Horizonte City (2021/04/30), Belo Horizonte. https://prefeitura.pbh.gov.br/sites/default/files/estrutura-de-governo/saude/2021/boletim_epidemiologico_assistencial_259_covid-19_30-04-2021.pdf, 2021.
- [15] V.M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D.K. Chu, T. Bleicker, S. Brünink, J. Schneider, M.L. Schmidt, D.G. Mulders, B.L. Haagmans, B. van der Veer, S. van den Brink, L. Wijsman, G. Godinger, J.-L. Romette, J. Ellis, M. Zambon, M. Peiris, H. Goossens, C. Reusken, M.P. Koopmans, C. Drosten, Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR, *Eurosurveillance* 25 (2020), <https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045>.
- [16] J.R. Tyson, P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J.H. Choi, H. Lapointe, K. Kamelian, A.D. Smith, N. Prystajek, I. Goodfellow, S.J. Wilson, R. Harrigan, T.P. Snutch, N.J. Loman, J. Quick, Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore, *BioRxiv Prepr. Serv. Biol.* (2020), <https://doi.org/10.1101/2020.09.04.283077>.
- [17] GeneStudio Inc, version 2.2.0.0. <http://genestudio.com/>, 2011.
- [18] K. Katoh, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res.* 30 (2002) 3059–3066, <https://doi.org/10.1093/nar/gkf436>.
- [19] J. Rozewicki, S. Li, K.M. Amada, D.M. Standley, K. Katoh, MAFFT-DASH: integrated protein sequence and structural alignment, *Nucleic Acids Res.* (2019), <https://doi.org/10.1093/nar/gkz342>.
- [20] Benchling [Biology Software]. <https://benchling.com>, 2021.
- [21] A. Larsson, AliView: a fast and lightweight alignment viewer and editor for large datasets, *Bioinformatics* 30 (2014) 3276–3278, <https://doi.org/10.1093/bioinformatics/btu531>.
- [22] pangolin - Phylogenetic Assignment of Name Global Outbreak Lineages. <https://pangolin.cog-uk.io/>, 2021.
- [23] B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era, *Mol. Biol. Evol.* 37 (2020) 1530–1534, <https://doi.org/10.1093/molbev/msaa015>.
- [24] G. Yu, Using ggtree to visualize data on tree-like structures, *Curr. Protoc. Bioinformatics* 69 (2020), <https://doi.org/10.1002/cpbi.96>.
- [25] C.M. Voloch, R. da Silva Francisco, L.G.P. de Almeida, C.C. Cardoso, O. J. Brustolini, A.L. Gerber, A.P.C. de Guimarães, D. Mariani, R.M. da Costa, O. C. Ferreira, A.C. Cavalcanti, T.S. Frauches, C.M.B. de Mello, I.C. de Leitão, R. M. Galliez, D.S. Faffe, T.M.P.P. Castiñeiras, A. Tanuri, A.T.R. de Vasconcelos, Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil, *J. Virol.* 95 (2021), <https://doi.org/10.1128/JVI.00119-21>.
- [26] T.S. Jørgensen, K. Blin, F. Kuntke, H.K. Salling, T.Y. Michaelsen, M. Albertsen, H. Larsen, A rapid, cost efficient and simple method to identify current SARS-CoV-2 variants of concern by Sanger sequencing part of the spike protein gene, *MedRxiv (Preprint)* (2021), <https://doi.org/10.1101/2021.03.27.21252266>.
- [27] J. Peto, Covid-19 mass testing facilities could end the epidemic rapidly, *BMJ* (2020), <https://doi.org/10.1136/bmj.m1163>.
- [28] C.A. Hogan, S. Gombar, H. Wang, K. Röltgen, R.-Z. Shi, M. Holubar, S. Chang, G. M. Lee, S.D. Boyd, J. Zehnder, B.A. Pinsky, Large-scale testing of asymptomatic healthcare personnel for severe acute respiratory syndrome coronavirus 2, *Emerg. Infect. Dis.* 27 (2021) 250–254, <https://doi.org/10.3201/eid2701.203892>.
- [29] C.J. Wang, C.Y. Ng, R.H. Brook, Response to COVID-19 in Taiwan, *JAMA* 323 (2020) 1341, <https://doi.org/10.1001/jama.2020.3151>.
- [30] H. Sung, C.-K. Yoo, M.-G. Han, S.-W. Lee, H. Lee, S. Chun, W.G. Lee, W.-K. Min, Preparedness and rapid implementation of external quality assessment helped quickly increase COVID-19 testing capacity in the Republic of Korea, *Clin. Chem.* 66 (2020) 979–981, <https://doi.org/10.1093/clinchem/hvaa097>.
- [31] M.-M. Aynaud, J.J. Hernandez, S. Barutcu, U. Braunschweig, K. Chan, J.D. Pearson, D. Trcka, S.L. Prosser, J. Kim, M. Barrios-Rodiles, M. Jen, S. Song, J. Shen, C. Bruce, B. Hazlett, S. Poutanen, L. Attisano, R. Bremner, B.J. Blencowe, T. Mazzulli, H. Han, L. Pelletier, J.L. Wrana, A multiplexed, next generation sequencing platform for high-throughput detection of SARS-CoV-2, *Nat. Commun.* 12 (2021) 1405, <https://doi.org/10.1038/s41467-021-21653-y>.
- [32] R.C. Bhojyar, A. Jain, P. Sehgal, M.K. Divakar, D. Sharma, M. Imran, B. Jolly, G. Ranjan, M. Rophina, S. Sharma, S. Siwach, K. Pandhare, S. Sahoo, M. Sahoo, A. Nayak, J.N. Mohanty, J. Das, S. Bhandari, S.K. Mathur, A. Kumar, R. Sahlot, P. Rojariani, J.V. Lakshmi, A. Surekha, P.C. Sekhar, S. Mahajan, S. Masih, P. Singh, V. Kumar, B. Jose, V. Mahajan, V. Gupta, R. Gupta, A. Arumugam, A. Singh, A. Nandy, P.V. Ravagendran, R.M. Jha, A. Kumari, S. Gandotra, V. Rao, M. Faruq, S. Kumar, B. Reshma, N. Varma, S.S. Roy, A. Sengupta, S. Chattopadhyay, K. Singhal, S. Pradhan, D. Jha, S. Naushin, S. Wadhwa, N. Tyagi, M. Poojary, V. Scaria, S. Sivasubbu, High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next-generation sequencing, *PLoS One* 16 (2021), <https://doi.org/10.1371/journal.pone.0247115>.
- [33] M. Chiara, A.M. D'Erchia, C. Gissi, C. Manzari, A. Parisi, N. Resta, F. Zambelli, E. Picardi, G. Pavesi, D.S. Horner, G. Pesole, Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities, *Brief. Bioinform.* 22 (2021) 616–630, <https://doi.org/10.1093/bib/bbaa297>.
- [34] S. Cherian, V. Potdar, S. Jadhav, P. Yadav, N. Gupta, M. Das, S. Das, A. Agarwal, S. Singh, P. Abraham, S. Panda, S. Mande, R. Swarup, B. Bhargava, R. Bhushan, N. team, I. Consortium, Convergent evolution of SARS-CoV-2 spike mutations, L452R, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India, *BioRxiv* (2021), <https://doi.org/10.1101/2021.04.22.440932>, 2021.04.22.440932.
- [35] D. Adam, What scientists know about new, fast-spreading coronavirus variants, *Nature* 594 (2021) 19–20, <https://doi.org/10.1038/d41586-021-01390-4>.
- [36] E.C. Sabino, L.F. Buss, M.P.S. Carvalho, C.A. Prete, M.A.E. Crispim, N.A. Fraiji, R.H. M. Pereira, K.V. Parag, P. da Silva Peixoto, M.U.G. Kraemer, M.K. Oikawa, T. Salomon, Z.M. Cucunuba, M.C. Castro, A.A. de Souza Santos, V.H. Nascimento, H.S. Pereira, N.M. Ferguson, O.G. Pybus, A. Kucharski, M.P. Busch, C. Dye, N. R. Faria, Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence, *Lancet* 397 (2021) 452–455, [https://doi.org/10.1016/S0140-6736\(21\)00183-5](https://doi.org/10.1016/S0140-6736(21)00183-5).
- [37] T. Fujino, H. Nomoto, S. Kutsuna, M. Ujiie, T. Suzuki, R. Sato, T. Fujimoto, M. Kuroda, T. Wakita, N. Ohmagari, Novel SARS-CoV-2 variant in travelers from Brazil to Japan, *Emerg. Infect. Dis.* 27 (2021), <https://doi.org/10.3201/eid2704.210138>.
- [38] G.R. Barbosa, L.V.L. Moreira, A.F.O. Justo, A.H. Perosa, L.K. de Souza Luna, A.P. C. Chaves, M.S. Bueno, D.D. Conte, J.M.A. Carvalho, J. Prates, P.S. Dantas, K. S. Faico-Filho, C. Camargo, P.C. Resende, M.M. Siqueira, N. Bellei, Rapid spread and high impact of the variant of concern P.1 in the largest city of Brazil, *J. Infect.* (2021), <https://doi.org/10.1016/j.jinf.2021.04.008>.
- [39] K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, GenBank, *Nucleic Acids Res.* 44 (2016) D67–D72, <https://doi.org/10.1093/nar/gkv1276>.
- [40] S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health, *Glob. Challenges* 1 (2017) 33–46, <https://doi.org/10.1002/gch2.1018>.