Seroprevalence, prevalence, and genomic surveillance:

monitoring the initial phases of the SARS-CoV-2 pandemic in Betim, Brazil

Ana Valesca Fernandes Gilson Silva [1,#,*], Diego Menezes [2,3,#], Filipe Romero Rebello Moreira [4,#], Octavio Alcântara Torres [1], Paula Luize Camargos Fonseca [2,3], Rennan Garcias Moreira [5], Hugo José Alves [2,3], Vivian Ribeiro Alves [1], Tania Maria de Resende Amaral [1], Adriano Neves Coelho [1], Júlia Maria Saraiva Duarte [3], Augusto Viana da Rocha [1], Luiz Gonzaga Paula de Almeida [6], João Locke Ferreira de Araújo [2,3], Hilton Soares de Oliveira [1], Nova Jersey Claudio de Oliveira [1], Camila Zolini de Sá [4], Jôsy Hubner de Sousa [7], Elizângela Gonçalves de Souza [1], Rafael Marques de Souza [2,3], Luciana de Lima Ferreira [2,3], Alexandra Lehmkuhl Gerber [6], Ana Paula de Campos Guimarães [6], Paulo Henrique Silva Maia [1], Fernanda Martins Marim [2,3], Lucyene Miguita [8], Cristiane Campos Monteiro [1], Tuffi Saliba Neto [1], Fabricia Soares Freire Pugêdo [1], Daniel Costa Queiroz [2,3], Damares Nigia Alborguetti Cuzzuol Queiroz [1], Luciana Cunha Resende-Moreira [9], Franciele Martins Santos [7], Erika Fernanda Carlos Souza [1], Carolina Moreira Voloch [4], Ana Tereza Vasconcelos [6], Renato Santana de Aguiar [2,3,10,*], Renan Pedra de Souza [2,3,*]

[1] Escola de Saúde Pública de Betim, Betim, MG, Brazil
[2] Programa de Pós Graduação em Genética; Departamento de Genética, Ecologia e Evolução; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[3] Laboratório de Biologia Integrativa; Departamento de Genética, Ecologia e Evolução; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[4] Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro, RJ, Brazil
[5] Centro de Laboratórios Multiusuários, Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[6] Laboratório Nacional de Computação Científica, Petropólis, RJ, Brazil
[7] Programa de Pós-graduação em Biologia Celular, Departamento de Morfologia; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[8] Departamento de Patologia; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[9] Departamento de Botânica; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[10] Instituto D'Or de Pesquisa e Ensino (IDOR), Rio de Janeiro, RJ, Brazil
[#] Authors contributed equally
[*] Corresponding author

Renan P. Souza (renanpedra@gmail.com) or Renato S. Aguiar (santanarnt@gmail.com) Universidade Federal de Minas Gerais. Av. Antônio Carlos, 6627 ICB – Pampulha, 31270901 – Belo Horizonte – Minas Gerais – Brazil. Phone: +553134092895.

Ana Valesca F. G. Silva (anavalescafernandes@hotmail.com) Escola de Saúde Pública de Betim. R. Para de Minas, 640 – Brasileia, 32600412 – Betim – Minas Gerais – Brazil. Phone: +553135123000

Abstract:

The Covid-19 pandemic has created an unprecedented need for epidemiological monitoring using diverse strategies. We conducted a project combining prevalence, seroprevalence, and genomic surveillance approaches to describe the initial pandemic stages in Betim City, Brazil. We collected 3239 subjects in a population-based age-, sex- and neighbourhood-stratified, household, prospective; cross-sectional study divided into three surveys 21 days apart sampling the same geographical area. In the first survey, overall prevalence (participants positive in serological or molecular tests) reached 0.46% (90% CI 0.12% – 0.80%), followed by 2.69% (90% CI 1.88% – 3.49%) in the second survey and 6.67% (90% CI 5.42% - 7.92%) in the third. The underreporting reached 11, 19.6, and 20.4 times in each survey, respectively. We observed increased odds to test positive in females compared to males (OR 1.88 95% CI 1.25 – 2.82), while the single best predictor for positivity was ageusia/ anosmia (OR 8.12, 95% CI 4.72 – 13.98). Thirty-five SARS-CoV-2 genomes were sequenced, of which 18 were classified as lineage B.1.1.28, while 17 were B.1.1.33. Multiple independent viral introductions were observed. Integration of multiple epidemiological strategies was able to describe Covid-19 dispersion in the city adequately. Presented results have helped local government authorities to guide pandemic management.

1. Introduction

Since its emergence in December 2019, the new human coronavirus has had a tremendous impact on humanity due to the pandemic nature of its infection, called Covid-19 [1]. The SARS-CoV-2 pathogen was described on January 24, 2020. In Brazil, the first case of Covid-19 was reported on February 26, 2020, in the city of São Paulo [2]. The virus spread rapidly, and the country had the highest number of cases and deaths in Latin America, experiencing its first peak wave in late July 2020. Although most cases were identified in the most prominent Brazilian cities, São Paulo and Rio de Janeiro, dispersion to other municipalities were quickly reported. Betim, a town located in the Minas Gerais State in Brazil with an estimated population of 439,340 in 2019, had its first reported SARS-CoV-2 case on March 23, 2020, in two patients returning from Europe. Two months later, on May 23, 2020, only 73 confirmed cases had been reported, although 4380 suspected cases were identified in public databases indicating limited testing availability.

Brazilian public healthcare system has prioritized testing subjects with symptoms due to scarce diagnostic tests, particularly in the early days of the pandemic. Since data suggest that symptomatic cases represent a fraction of persons infected with SARS-CoV-2, official statistics were expected to be underestimated [3]. Epidemiological surveillance using prevalence studies is needed to evaluate the true extent of SARS-CoV-2 dispersion, significantly extending testing to asymptomatic subjects. Combining serological and molecular tests may be a more robust strategy to uncover viral diffusion in a territory, avoiding each test's kinetic detection limitations. Valid prevalence and seroprevalence estimates for a population rely on two major factors: (i) a representative population sample and (ii) accurate diagnostic testing [4].

While the epidemiological investigation is essential for controlling Covid-19, genomic surveillance is equally crucial. Robust SARS-CoV-2 variant monitoring can track viral evolution, detect new variants, describe patterns and clusters of transmission, outbreak tracking, among others. Therefore, it can provide actionable information on implementing a more targeted public health strategy that addresses local priorities through stakeholder engagement and mitigation efforts [5]. Here, we conducted a study combining seroprevalence, prevalence, and genomic surveillance approaches to understand the SARS-CoV-2 epidemic spread in Betim city.

2 - Materials and Methods:

2.1 – Seroprevalence and prevalence

The Research Ethics Committee approved the present experiment under protocol CAAE 31459220.2.0000.5651. We conducted a population-based age-, sex- and neighbourhood-stratified, household, prospective; cross-sectional study repeated every 21 days in the same geographic area to determine the extent of SARS-CoV-2 transmission in Betim, Minas Gerais, Brazil. Three surveys were held: June 3-5, June 23-25, and July 13-15, 2020. The sample size (n = 1,080 each survey) was estimated considering dichotomous outcome (positive or negative), the population of 439,340 inhabitants, the confidence level of 90%, the maximum margin of error of 2.5%, and lack of a priori information on the prevalence of SARS-COV-2 in the municipality's population (the latter represented by p = q = 0.5) and using the equation below:

$$ n = \frac{z_{\frac{\alpha}{2}}^2 . \hat{p} . \hat{q} . N}{E^2 . (N - 1) + z_{\frac{\alpha}{2}}^2 . \hat{p} . \hat{q}} $$

Random sampling was employed to ensure representativeness of the population, stratified by sex, age (0 to 5; 6 to 19; 20 to 39; 40 to 59 and 60 years or older) and city neighbourhoods (Centro, Alterosas, Imbiruçu, Norte, Teresópolis, PTB, Citrolândia, Vianópolis, Icaivera, and Petrovale). Every census tract (population stratum created by Governmental agencies) was sampled with at least one address. In case of refusal or closed households, the closest home was selected. Thirty-six teams (one driver, one nurse, and one community health worker) worked on active sampling subjects in 1080 addresses during three days. Clinical and epidemiological data were obtained using a questionnaire during interviews with participants or their legal guardians who signed the Informed Consent. Biological samples were collected using a nasal swab to conduct RT-PCR and capillary blood obtained by fingerstick for the serological test.

RT-PCR to detect SARS-CoV-2 RNA was initially conducted in pools of ten samples [6]. Whenever pools were positive, individual samples were examined. Molecular diagnosis was established according to the CDC 2019-Novel Coronavirus Real-Time RT-PCR Diagnostic Panel (N1, N2 and RNP primers). Serological tests were conducted using the SARS-CoV-2 Antibody Test (Guangzhou Wondfo Biotech Co., Ltd.) that detects IgM/IgG antibodies. The same test was used in a previous study in Brazil [7].

Reported sensitivity is 86.43% (95% CI: 82.41% ~ 89.58%) and specificity 99.57% (95% CI: 97.63% ~ 99.92%). We have validated antibody tests using serum samples from subjects who were SARS-CoV-2 positive confirmed with RT-PCR.

Associations of each variable of interest with surveys (Table 1) and positive status (Table 2) were assessed using chi-square tests. Odds ratios were estimated using logistic regression with the *glm* function. Spatial geostatistical modelling and prediction were carried out using the *gstat* and *predict* functions from the *gstat* package. All analyses were carried out in R software (version 4.1.1).

2.2 – Genomic surveillance

Whole viral genome amplification and DNA library preparation was carried out as described elsewhere [8]. Briefly, QIAseq SARS-CoV-2 Primer Panel - QIAGEN kit was used to amplify positive samples, following manufacturer instructions. In total, 39 of the 84 detectable samples were eligible for library preparation based on their CTs ≤ 30. Library concentration was measured using the QIAseq Library Quant Assay - QIAGEN kit, and the fragment integrity and size were evaluated using Bioanalyzer (Agilent Technologies, Waldbronn, DE). Sequencing was carried out on a MiSeq (Illumina, San Diego, CA, USA).

The raw data generated were filtered by Trimmomatic v0.39 [9], which trimmed low-quality bases (Phred score < 30) and removed short reads (<50 nucleotides) as well as adapters and primer sequences. Reads were then mapped against the SARS-CoV-2 reference genome (accession number: NC_045512.2) with Bowtie2 [10]. The resulting BAM files were manipulated with SAMtools, BCFtools [11], and BEDtools [12] to generate consensus genome sequences. Bases with less than 10x sequencing depth were masked. In total, 35 of the 39 genome sequences presented coverage greater than 79% and average sequencing depth greater than 200x. Sequencing metadata is available in **Table S1**. The 35 consensus genome sequences were submitted to the PANGOLIN 2.0 lineage classification tool (database version February 2, 2021) [13].

To confirm the PANGOLIN identification and further contextualize the diversity of lineages circulating in Betim, we performed a set of phylogenetic analyses. First, a global dataset was assembled from a subset of high-quality data available on GISAID and the newly generated genomes (*n* = 3,814). This dataset contained all Brazilian sequences and one per week for each country, as available on GISAID until January 12, 2021. These sequences were aligned with MAFFT v7.475[14], and a maximum likelihood tree was

inferred on IQ-Tree 2 [15], under the GTR+F+I+G4 model [16], [17]. Shimoidara-Hasegawa approximate likelihood ratio test (SH-aLRT) was used to assess branches' statistical support [18].

Two subsets of the previous dataset were assembled to explore the temporal dynamics of introduction and circulation of SARS-CoV-2 in Betim, comprehending sequences belonging to lineages B.1.1.28 ($n = 258$) and B.1.1.33 ($n = 284$). The parameterization of the phylogeographic model was set to be primarily informative concerning introductions of SARS-CoV-2 in Betim. Therefore, we set the model with six discrete categories: Betim City, Minas Gerais State, Rio de Janeiro State, São Paulo State, other Brazilian States, and foreign sequences. These locations were represented by 18, 2, 22, 71, 79, and 66 sequences in dataset B.1.1.28 while B.1.1.33 dataset composition was 17, 20, 53, 52, 73, and 69 sequences from each region, respectively.

Maximum likelihood trees were inferred from these datasets, and their temporal signal was evaluated with tempest v1.5.3 [19]. Time scaled phylogenies were then inferred from these datasets with BEAST v1.10.4 [20], using: (***i***) the HKY+I+G4 nucleotide substitution model [17], (***ii***) the strict molecular clock model, (***iii***) the non-parametric coalescent skygrid tree prior [21] and (***iv***) a symmetric discrete phylogeographic model [22]. A normal prior distribution (mean = 1.13 x 10⁻³; std= 5.1 x 10⁻⁴) on clock rate was assumed, based on a previous estimate [23]. The cutoff values of the skygrid tree prior were set based on the previously estimated dates for the emergence of each lineage [23]. The number of grids of the tree priors was set to match the approximate number of weeks comprehended between the estimated dates for lineages' emergence and the dates of the most recently sampled sequences (41 weeks, both datasets). Two and three independent chains of 200 million generations sampling every 10,000 states were performed for datasets B.1.1.33 and B.1.1.28, respectively. Tracer v1.7.1 [24] was used to verify mixing and convergence of chains (effective sample size > 200 for all parameters), which were then combined with logcombiner v1.10.4 after 10% burning removal. Maximum clade credibility trees were generated with treeannotator v1.10.4. All logs and trees are available in https://github.com/LBI-lab/SARS-CoV-2_phylogenies.git.

3 - Results

3.1 – Seroprevalence and prevalence

Table 1 presents clinical and epidemiological data obtained from participants. No significant difference was observed for the presence of any prior health condition across surveys (pneumopathy, chronic neurological disease, pregnant, postpartum, chronic cardiovascular disease, chronic kidney disease, obesity, asthma, immunodepression, chronic liver disease, diabetes, hypertension, transplanted, cancer or any comorbidity) indicating proper sampling was conducted since there was no reason to find significant differences in the period. Four symptoms (cough, sore throat, myalgia, and rhinorrhea) and contact with a symptomatic person increased while international travel decreased. Prevalence and seroprevalence increased across surveys.

Pandemic progression in Betim city is presented in **Figure 1**. Confirmed cases underestimation was found in all three surveys. In the first survey, overall prevalence (participants positive in serological or molecular tests) reached 0.46% (90% CI 0.12% – 0.80%), followed by 2.69% (90% CI 1.88% – 3.49%) in the second survey and 6.67% (90% CI 5.42% - 7.92%) in the third. The underreporting was obtained by the difference between survey prevalence and official data, and its magnitude reached 11, 19.6, and 20.4 times (distance between black dots and red curve in Figure 1B). Active transmission areas (RT-PCR positive participants) were observed increasing across time (**Figure 1C-E**). By the third survey, almost all populated city areas were likely to have viral circulation (**Figure 1E**). The same pattern of increase was observed in overall prevalence for most administrative regions (**Figure 1F-G**).

We have also evaluated whether clinical and epidemiological variables were associated with molecular or serological test positivity (**Table 2**). Several significant results were observed, mostly with reported symptoms (fever, cough, sore throat, dyspnoea, myalgia, rhinorrhea, respiratory discomfort, nausea/ vomit, headache, prostration, ageusia/ anosmia). We also observed increased odds to test positive in females compared to males (OR 1.88 95% CI 1.25 – 2.82) and clear enrichment of positive cases in certain city regions (e.g., Imbiruçu and Terezópolis). Surprisingly, people with obesity were more likely to be positive (OR 3.33, 95% CI 1.68 – 6.59). The single best predictor for positivity was ageusia/ anosmia (OR 8.12, 95% CI 4.72 – 13.98). Non-significant results can be found in **Table S2**.

3.2 – Genomic viral surveillance

In total, 35 novel SARS-CoV-2 genome sequences were obtained (GISAID EPI_ISL_5416087-5416121). The sequences were classified by PANGOLIN 2.0 to assess the genetic diversity of SARS-CoV-2 circulating in Betim. 18 of the 35 genomes were classified as lineage B.1.1.28, while 17 were B.1.1.33 (*Probability* = 1.0). Further, a maximum likelihood tree was inferred from the global dataset GISAID [25].

The analysis supported these results, revealing sequences from the Betim cluster within several clades of these lineages confirming the circulation of (B.1.1.28 and B.1.1.33 during the first wave of COVID-19 pandemics in the city (**Figure 2**). The spread of Betim sequences across the tree suggests multiple independent introductions occurred in the town. Further, eight clades majorly composed by Betim sequences were inferred with variable degrees of statistical support (median SH-aLRT = 82.75, range: 0 - 100), suggesting the occurrence of local transmission in the city after initial introduction events. In addition to these clusters, nine introductions supported by single sequences have also been detected. Most Betim sequences or clusters are closely related to sequences from Rio de Janeiro and São Paulo, two neighbouring States connected by highways to Minas Gerais. To formally assess the dynamics of introduction and spread of SARS-CoV-2 in Betim, separated datasets for lineages B.1.1.28 and B.1.1.33 were evaluated. Regression between sampling times and genetic distances revealed both datasets had moderate temporal signal (B.1.1.28: R^2 = 0.49; B.1.1.33: R^2 = 0.58), justifying molecular clock analysis.

The time-scaled phylogeographic analysis performed with dataset B.1.1.28 suggests this lineage emerged on February 22, 2020, in São Paulo (95% highest posterior density, HPD: February 11, 2020 - March 05, 2020; geographic model posterior probability, PP = 0.91), later spreading to other Brazilian states (**Figure 3A**). The phylogeny reveals that two introduction events, dated between April 19, 2020 (95% HPD: April 17, 2020 - May 11, 2020) and April 22, 2020 (95% HPD: April 20, 2020 – May 27, 2020), led to the emergence of Betim clusters (harbouring between two and six sequences). Additionally, four introductions related to single sequences have been detected. The phylogeographic model suggests that three introductions occurred from another Brazilian region to Betim, in addition to other single events from RJ, another one from SP, and another from foreign sequences. All events presented high statistical support (PP > 92% for all events).

The phylogeographic reconstruction performed for dataset B.1.1.33 infers the origin of this lineage on February 06, 2020, in Rio de Janeiro (95% HPD: January 14, 2020 – February 25, 2020, PP = 0.78). The model supports the occurrence of many Betim clusters. One cluster comprises four sequences, dating to May 27, 2020 (95% HPD: May 01, 2020 - June 03, 2020) grouped with other sequences from other Brazilian regions and foreign. The model has also estimated eight introductions supported by single sequences. According to our phylogeny, the B.1.1.33 introductions came from different locations, such as the states of Rio de Janeiro, São Paulo, Minas Gerais, other Brazilian regions, and foreign sequences (PP > 0.81 for all events) (**Figure 3B**). The patterns reconstructed by both phylogeographic inferences are consistent, indicating the emergence of lineages B.1.1.28 and B.1.1.33 was followed by multiple importation events to diverse regions within the country, likely driven by human mobility. Additionally, evolutionary rate estimates also differed between datasets (B.1.1.28: 8.6372×10^{-4}, 95% HPD: 7.8379×10^{-4} - 9.4559×10^{-4}; B.1.1.33: 6.8743×10^{-4}, 95% HPD 6.1784×10^{-4} - 7.5446×10^{-4}).

4 - Discussion

Betim is a medium-sized Brazilian city (439,340 inhabitants, 343 thousand square kilometres) crossed by national roads connecting major Brazilian cities and serving as a local hub for the Brazilian Public Health System. Understanding its pandemic dynamic may provide relevant information for municipalities with similar features. Here, we estimated the overall prevalence of active infections, seroprevalence and conducted genomic surveillance before the first pandemic wave in Betim. Brazilian molecular diagnostic capacity was insufficient in the first months of the pandemic [26]. Therefore, Covid-19 cases may have been included in the official statistics as severe acute respiratory infection cases with unknown aetiology. Data until May 2020 indicated a positive association between higher per-capita income and molecular Covid-19 diagnosis, while the severe acute respiratory infection cases with unknown aetiology were associated with lower per-capita income, suggesting a possible diagnosis bias related to economic status [27]. Inadequate diagnosis availability may lead to underreporting [28]. Our data estimated underreporting rates up to 20 times.

No studies have been conducted in Brazil evaluating active infection prevalence using adequate sampling. Our study design was inspired by previous research conducted in Santa Clara, USA, using pooled samples [29]. Pooled PCR tests were initially suggested to be used in asymptomatic people [6] and later were recommended for surveillance studies in populations with low infection prevalence [30]. Seroprevalence studies were conducted during the first wave in Brazil that peaked in July 2020. Two of the highest city seroprevalences reported during the period were Boa Vista (25.4% in June 2020) [7] and São Luiz (40.4% between the end of July and August 2020) [31], both in the northern area of the country. A nationwide survey carried out in May and June 2020 presented seroprevalence lower than two per cent during both surveys in all sampled cities neighbouring Betim (less than 200km), corroborating our findings [7]. Furthermore, seroprevalences higher than ten per cent were solely found in towns in the North Region [7]. In December 2020, Manaus, the largest city in the North Region, experienced a resurgence of Covid-19 [32] despite high seroprevalence [33], likely due to the gamma variant [34].

Previous seroprevalence studies have indicated ethnic and socioeconomic bias for SARS-CoV-2 infection in Brazil since the pandemic's beginning [35], [36]. Results from Rio de Janeiro in April 2020 indicated that younger blood donors with lower education

levels were more likely to test positive for SARS-CoV-2 antibodies [35]. A nationwide study revealed that the poorest quintile was 2.16 times more likely to test positive with the lowest risks among white, educated, and wealthy individuals [36]. Likewise, we found one of the highest prevalences in the poorest neighbourhood, Terezópolis, that include the largest slum of the city where more than 23 thousand people live.

Further modelling results showed higher infection rates among young adults, lower socioeconomic status, and people without healthcare access in the less developed North and Northeast areas until August 2020 [37]. Betim also presents most of its inhabitants with less than 59 years (90.7%), but no age effect was observed in the infectivity rates. Increased female infection odds were observed, although previous reports indicated a gender predisposition towards death in some Brazilian regions with higher male risk [38]. One possible explanation could be that 70% of the global health workforce are women [39] and a gender bias of pandemic perception and attitude [40].

Covid-19 diffusion presents strong socio-spatial determinants. Relocation diffusion from more- to less-developed regions and hierarchical diffusion from countries with higher population and density were relevant since early 2020 [41]. Data indicated a similar pattern in the São Paulo State with contiguous diffusion from the capital metropolitan area and hierarchical with long-distance spread through major highways that connects São Paulo city with cities of regional relevance [42]. Modelling results revealed that São Paulo city may have accounted for more than 85% of the initial case spread in the entire country [43]. Betim is directly connected to São Paulo city by a main national highway which may have contributed to Covid-19 diffusion.

Genomic surveillance is a powerful tool to elucidate viral dispersion patterns. The first sequencing work conducted in Brazil evaluated the first six positive individuals and reported the same predominant lineages found in Italy [44]. Later, a study with samples collected until late April 2020 from different country areas showed the dominance of clade B-derived lineages. At the national level, the respective frequency of these clades was seen in a 98.98%/1.02% ratio [23]. In Minas Gerais State, A lineages represented 2.5% of the infections, B.1 appeared in 92.5% of the samples, and B was responsible for 5% of the cases [45]. The exclusivity of lineages B.1.1.28 and B.1.1.33 circulating in Betim-MG from June to July 2020, given that multiple introductions from different country regions were demonstrated, is representative of the extent of these lineages' dominance in the Brazilian scenario at the moment. Independent introductions also

emphasize the importance of inter-state mobility barriers as a measure to control the epidemic.

Our study presents some limitations. First, the household survey is less likely to sample severe cases, thus underestimating symptomatic Covid-19. Second, all clinical data were self-reported, which may lead to reporting bias [46]. Third, we could not sequence all PCR positive samples due to the low viral load and sequencing technology employed. Nevertheless, our study shows the potential to integrate different epidemiological inquiries (prevalence, seroprevalence, and genomic surveillance) to describe pandemic dispersion adequately. Moreover, our findings present original and relevant evidence that has helped local government authorities to guide pandemic management.

Conflict of interest

None

Acknowledgement

References:

[1]   P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, 2020, doi: 10.1038/s41586-020-2012-7.

[2]   D. B. Araujo *et al.*, "SARS-CoV-2 isolation from the first reported patients in brazil and establishment of a coordinated task network," *Mem. Inst. Oswaldo Cruz*, vol. 115, 2020, doi: 10.1590/0074-02760200342.

[3]   S. L. Wu *et al.*, "Substantial underestimation of SARS-CoV-2 infection in the United States," *Nat. Commun.*, vol. 11, no. 1, 2020, doi: 10.1038/s41467-020-18272-4.

[4]   O. Byambasuren *et al.*, "Comparison of seroprevalence of SARS-CoV-2 infections with cumulative and imputed COVID-19 cases: Systematic review," *PLoS ONE*, vol. 16, no. 4 April. 2021, doi: 10.1371/journal.pone.0248946.

[5]   J. D. Robishaw *et al.*, "Genomic surveillance to combat COVID-19: challenges and opportunities," *The Lancet Microbe*, vol. 2, no. 9, 2021, doi: 10.1016/s2666-5247(21)00121-x.

[6]   S. Lohse *et al.*, "Pooling of samples for testing for SARS-CoV-2 in asymptomatic people," *The Lancet Infectious Diseases*, vol. 20, no. 11. 2020, doi: 10.1016/S1473-3099(20)30362-5.

[7]   P. C. Hallal *et al.*, "SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys," *Lancet Glob. Heal.*, vol. 8, no. 11, 2020, doi: 10.1016/S2214-109X(20)30387-9.

[8]   F. R. R. Moreira *et al.*, "Epidemic spread of sars-cov-2 lineage b.1.1.7 in Brazil," *Viruses*, vol. 13, no. 6. 2021, doi: 10.3390/v13060984.

[9]   A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, 2014, doi: 10.1093/bioinformatics/btu170.

[10]   B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-

efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, 2009, doi: 10.1186/gb-2009-10-3-r25.

[11]   H. Li *et al.*, "The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup," *Bioinformatics*, vol. 25, no. 16, 2009.

[12]   A. R. Quinlan and I. M. Hall, "BEDTools: A flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, 2010, doi: 10.1093/bioinformatics/btq033.

[13]   A. Rambaut *et al.*, "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nat. Microbiol.*, vol. 5, no. 11, 2020, doi: 10.1038/s41564-020-0770-5.

[14]   K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability," *Mol. Biol. Evol.*, vol. 30, no. 4, 2013, doi: 10.1093/molbev/mst010.

[15]   B. Q. Minh *et al.*, "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Mol. Biol. Evol.*, vol. 37, no. 5, 2020, doi: 10.1093/molbev/msaa015.

[16]   S. Tavaré, "Some probabilistic and statistical problems in the analysis of DNA sequences," *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, vol. 17. 1986.

[17]   Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods," *J. Mol. Evol.*, vol. 39, no. 3, 1994, doi: 10.1007/BF00160154.

[18]   S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0," *Syst. Biol.*, vol. 59, no. 3, 2010, doi: 10.1093/sysbio/syq010.

[19]   A. Rambaut, T. T. Lam, L. M. Carvalho, and O. G. Pybus, "Exploring the

temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)," *Virus Evol.*, vol. 2, no. 1, 2016, doi: 10.1093/ve/vew007.

[20] M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut, "Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10," *Virus Evol.*, vol. 4, no. 1, 2018, doi: 10.1093/ve/vey016.

[21] M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, and M. A. Suchard, "Improving bayesian population dynamics inference: A coalescent-based model for multiple loci," *Mol. Biol. Evol.*, vol. 30, no. 3, 2013, doi: 10.1093/molbev/mss265.

[22] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard, "Bayesian phylogeography finds its roots," *PLoS Comput. Biol.*, vol. 5, no. 9, 2009, doi: 10.1371/journal.pcbi.1000520.

[23] D. S. Candido *et al.*, "Evolution and epidemic spread of SARS-CoV-2 in Brazil," *Science (80-.).*, vol. 369, no. 6508, pp. 1255–1260, 2020, doi: 10.1126/SCIENCE.ABD2161.

[24] A. Rambaut, A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard, "Posterior summarization in Bayesian phylogenetics using Tracer 1.7," *Syst. Biol.*, vol. 67, no. 5, 2018, doi: 10.1093/sysbio/syy032.

[25] Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data – from vision to reality," *Eurosurveillance*, vol. 22, no. 13. 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[26] R. M. T. Grotto *et al.*, "Increasing molecular diagnostic capacity and COVID-19 incidence in Brazil," *Epidemiol. Infect.*, 2020, doi: 10.1017/S0950268820001818.

[27] W. M. de Souza *et al.*, "Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil," *Nat. Hum. Behav.*, vol. 4, no. 8, 2020, doi: 10.1038/s41562-020-0928-4.

[28] E. Kupek, "How many more? Under-reporting of the COVID-19 deaths in Brazil

in 2020," *Trop. Med. Int. Heal.*, vol. 26, no. 9, 2021, doi: 10.1111/tmi.13628.

[29]   C. A. Hogan, M. K. Sahoo, and B. A. Pinsky, "Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2," *JAMA - Journal of the American Medical Association*, vol. 323, no. 19. 2020, doi: 10.1001/jama.2020.5445.

[30]   L. Mutesa *et al.*, "A pooled testing strategy for identifying SARS-CoV-2 at low prevalence," *Nature*, vol. 589, no. 7841, 2021, doi: 10.1038/s41586-020-2885-5.

[31]   A. A. M. da Silva *et al.*, "Population-based seroprevalence of SARS-CoV-2 and the herd immunity threshold in Maranhão," *Rev. Saude Publica*, vol. 54, 2020, doi: 10.11606/s1518-8787.2020054003278.

[32]   E. C. Sabino *et al.*, "Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence," *The Lancet*, vol. 397, no. 10273. 2021, doi: 10.1016/S0140-6736(21)00183-5.

[33]   L. F. Buss *et al.*, "Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic," *Science (80-.).*, vol. 371, no. 6526, 2021, doi: 10.1126/science.abe9728.

[34]   N. R. Faria *et al.*, "Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil," *Science (80-.).*, vol. 372, no. 6544, 2021, doi: 10.1126/science.abh2644.

[35]   L. A. Filho *et al.*, "Seroprevalence of anti-SARS-CoV-2 among blood donors in Rio de Janeiro, Brazil," *Rev. Saude Publica*, vol. 54, 2020, doi: 10.11606/s1518-8787.2020054002643.

[36]   B. L. Horta *et al.*, "Prevalence of antibodies against SARS-CoV-2 according to socioeconomic and ethnic status in a nationwide Brazilian survey," *Rev. Panam. Salud Publica/Pan Am. J. Public Heal.*, vol. 40, 2020, doi: 10.26633/RPSP.2020.135.

[37]   E. E. Campos de Lima, E. Gayawan, E. A. Baptista, and B. L. Queiroz, "Spatial pattern of COVID-19 deaths and infections in small areas of Brazil," *PLoS One*,

vol. 16, no. 2 Febuary, 2021, doi: 10.1371/journal.pone.0246808.

[38]   P. Baqui, I. Bica, V. Marra, A. Ercole, and M. van der Schaar, "Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study," *Lancet Glob. Heal.*, vol. 8, no. 8, 2020, doi: 10.1016/S2214-109X(20)30285-0.

[39]   G. Lotta, M. Fernandez, D. Pimenta, and C. Wenham, "Gender, race, and health workers in the COVID-19 pandemic," *The Lancet*, vol. 397, no. 10281. 2021, doi: 10.1016/S0140-6736(21)00530-4.

[40]   V. Galasso, V. Pons, P. Profeta, M. Becher, S. Brouard, and M. Foucault, "Gender differences in COVID-19 attitudes and behavior: Panel evidence from eight countries," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 44, 2020, doi: 10.1073/pnas.2012520117.

[41]   T. Sigler *et al.*, "The socio-spatial determinants of COVID-19 diffusion: the impact of globalization, settlement characteristics and population," *Global. Health*, vol. 17, no. 1, 2021, doi: 10.1186/s12992-021-00707-2.

[42]   C. M. C. Branco Fortaleza *et al.*, "The use of health geography modeling to understand early dispersion of COVID-19 in São Paulo, Brazil," *PLoS One*, vol. 16, no. 1 January, 2021, doi: 10.1371/journal.pone.0245051.

[43]   M. A. L. Nicolelis, R. L. G. Raimundo, P. S. Peixoto, and C. S. Andreazzi, "The impact of super-spreader cities, highways, and intensive care availability in the early stages of the COVID-19 epidemic in Brazil," *Sci. Rep.*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-92263-3.

[44]   J. G. de Jesus *et al.*, "Importation and early local transmission of covid-19 in brazil, 2020," *Rev. Inst. Med. Trop. Sao Paulo*, vol. 62, 2020, doi: 10.1590/S1678-9946202062030.

[45]   J. Xavier *et al.*, "The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing," *Emerg. Microbes Infect.*, vol. 9, no. 1, 2020, doi: 10.1080/22221751.2020.1803146.

[46] M. Baker, M. Stabile, and C. Deri, "What do self-reported, objective, measures of health measure?," *J. Hum. Resour.*, vol. 39, no. 4, 2004, doi: 10.2307/3559039.

Table 1: Clinical and epidemiological data obtained from participants. Bolded p values indicate p < 0.05.

Variable	Level	Overall n (%)	First survey n (%)	Second survey n (%)	Third survey n (%)	p-value
Administrative Regions	Alterosas	634 (19.6%)	198 (18.4%)	218 (20.2%)	218 (20.2%)	0.9584
	Citrolândia	219 (6.8%)	83 (7.7%)	68 (6.3%)	68 (6.3%)	
	Icaivera	62 (1.9%)	20 (1.9%)	21 (1.9%)	21 (1.9%)	
	Imbiruçu	565 (17.4%)	183 (17.0%)	191 (17.7%)	191 (17.7%)	
	Norte	333 (10.3%)	111 (10.3%)	111 (10.3%)	111 (10.3%)	
	Petrovale	41 (1.3%)	13 (1.2%)	14 (1.3%)	14 (1.3%)	
	PTB	290 (9.0%)	108 (10.0%)	91 (8.4%)	91 (8.4%)	
	Sede	583 (18.0%)	201 (18.6%)	191 (17.7%)	191 (17.7%)	
	Terezópolis	319 (9.8%)	109 (10.1%)	105 (9.7%)	105 (9.7%)	
	Vianópolis	193 (6.0%)	53 (4.9%)	70 (6.5%)	70 (6.5%)	
Sex	Female	1628 (50.3%)	548 (50.8%)	536 (49.6%)	544 (50.4%)	0.8619
Age range	0 - 5	217 (6.7%)	71 (6.6%)	73 (6.8%)	73 (6.8%)	1.0000
	6 - 19	650 (20.1%)	218 (20.2%)	217 (20.1%)	215 (19.9%)	
	20-39	1067 (32.9%)	354 (32.8%)	355 (32.9%)	358 (33.1%)	
	40-59	871 (26.9%)	291 (27.0%)	289 (26.8%)	291 (26.9%)	
	Above 60	434 (13.4%)	145 (13.4%)	146 (13.5%)	143 (13.2%)	
Pneumopathy	Yes	30 (0.9%)	7 (0.6%)	13 (1.2%)	10 (0.9%)	0.4042
Chronic neurological disease	Yes	39 (1.2%)	16 (1.5%)	10 (0.9%)	13 (1.2%)	0.4948
Pregnant	Yes	28 (0.9%)	10 (0.9%)	11 (1.0%)	7 (0.6%)	0.6257
Postpartum	Yes	9 (0.3%)	2 (0.2%)	3 (0.3%)	4 (0.4%)	0.7165
Chronic cardiovascular disease	Yes	96 (3.0%)	34 (3.2%)	39 (3.6%)	23 (2.1%)	0.1154
Chronic kidney disease	Yes	50 (1.5%)	24 (2.2%)	12 (1.1%)	14 (1.3%)	0.0799
Obesity	Yes	105 (3.2%)	33 (3.1%)	37 (3.4%)	35 (3.2%)	0.8903
Asthma	Yes	173 (5.3%)	65 (6.0%)	58 (5.4%)	50 (4.6%)	0.3537
Immunodepression	Yes	22 (0.7%)	9 (0.8%)	5 (0.5%)	8 (0.7%)	0.5507
Chronic liver disease	Yes	15 (0.5%)	4 (0.4%)	7 (0.6%)	4 (0.4%)	0.5478
Diabetes	Yes	228 (7.0%)	78 (7.2%)	74 (6.9%)	76 (7.0%)	0.9430
Hypertension	Yes	563 (17.4%)	190 (17.6%)	186 (17.2%)	187 (17.3%)	0.9698
Transplanted	Yes	4 (0.1%)	2 (0.2%)	1 (0.1%)	1 (0.1%)	0.7780
Cancer	Yes	23 (0.7%)	10 (0.9%)	8 (0.7%)	5 (0.5%)	0.4342
Any comorbidity	Yes	955 (29.5%)	327 (30.3%)	320 (29.6%)	308 (28.5%)	0.6552
Fever	Yes	224 (6.9%)	66 (6.1%)	70 (6.5%)	88 (8.1%)	0.1398
Cough	Yes	648 (20.0%)	185 (17.1%)	213 (19.7%)	250 (23.1%)	**0.0022**
Sore throat	Yes	397 (12.3%)	112 (10.4%)	125 (11.6%)	160 (14.8%)	**0.0051**
Dyspnoea	Yes	141 (4.4%)	49 (4.5%)	46 (4.3%)	46 (4.3%)	0.9336
Myalgia	Yes	284 (8.8%)	74 (6.9%)	99 (9.2%)	111 (10.3%)	**0.0165**
Rhinorrhea	Yes	717 (22.1%)	205 (19.0%)	240 (22.2%)	272 (25.2%)	**0.0025**
Respiratory discomfort	Yes	188 (5.8%)	63 (5.8%)	58 (5.4%)	67 (6.2%)	0.7084
Nausea/ vomit	Yes	120 (3.7%)	37 (3.4%)	39 (3.6%)	44 (4.1%)	0.7156
Headache	Yes	790 (24.4%)	244 (22.6%)	259 (24.0%)	287 (26.6%)	0.0936
Prostration	Yes	188 (5.8%)	60 (5.6%)	51 (4.7%)	77 (7.1%)	0.0523
Diarrhea	Yes	211 (6.5%)	59 (5.5%)	76 (7.0%)	76 (7.0%)	0.2336
Conjunctivitis	Yes	32 (1.0%)	13 (1.2%)	11 (1.0%)	8 (0.7%)	0.5478
Ageusia/ anosmia	Yes	101 (3.1%)	30 (2.8%)	30 (2.8%)	41 (3.8%)	0.2914
Loss of voice	Yes	56 (1.7%)	18 (1.7%)	13 (1.2%)	25 (2.3%)	0.1381
Sought health assistance	Hospital	138 (4.3%)	41 (3.8%)	41 (3.8%)	56 (5.2%)	0.1492
	Basic Health Unit	129 (4.0%)	42 (3.9%)	41 (3.8%)	46 (4.3%)	
	Emergency Care Unit	127 (3.9%)	38 (3.5%)	35 (3.2%)	54 (5.0%)	
	None	2845 (87.8%)	958 (88.8%)	963 (89.2%)	924 (85.6%)	
Admitted to a health institution	Yes	38 (1.2%)	11 (1.0%)	12 (1.1%)	15 (1.4%)	0.7085
International travel	Yes	14 (0.4%)	10 (0.9%)	4 (0.4%)	0 (0.0%)	**0.0043**
Household contact with symptomatic person	Yes	640 (19.8%)	157 (14.6%)	193 (17.9%)	290 (26.9%)	**< 0.0001**
Sorological test	Reactive	39 (1.2%)	3 (0.3%)	8 (0.7%)	28 (2.6%)	**< 0.0001**
	Non-reactive	3200 (98.8%)	1076 (99.7%)	1072 (99.3%)	1052 (97.4%)	
PCR test	Detected	84 (2.6%)	2 (0.2%)	22 (2.0%)	60 (5.6%)	**< 0.0001**
	Undetected	3112 (96.1%)	1035 (95.9%)	1057 (98.0%)	1020 (94.4%)	
	Indeterminate	42 (1.3%)	42 (3.9%)	0 (0.0%)	0 (0.0%)	
Prevalence	Sorological reacive and/ or PCR detected	106 (3.3%)	5 (0.5%)	29 (2.7%)	72 (6.7%)	**< 0.0001**

Table 2: Significant associations of clinical and epidemiological data with positive test (serological or molecular). Non-significant associations are presented in Table S1. Bolded p values indicate p < 0.05.

Variable	Level	Positive	Negative	p-value
Survey	First	5 (4.7%)	1074 (34.3%)	
	Second	29 (27.4%)	1051 (33.5%)	
	Third	72 (67.9%)	1008 (32.2%)	**< 0.0001**
Administrative Regions	Alterosas	18 (17.0%)	616 (19.7%)	
	Citrolândia	4 (3.8%)	215 (6.9%)	
	Icaivera	0 (0.0%)	62 (2.0%)	
	Imbiruçu	32 (30.2%)	533 (17.0%)	
	Norte	11 (10.4%)	322 (10.3%)	
	Petrovale	0 (0.0%)	41 (1.3%)	
	PTB	8 (7.5%)	282 (9.0%)	
	Sede	15 (14.2%)	568 (18.1%)	
	Terezópolis	17 (16.0%)	302 (9.6%)	
	Vianópolis	1 (0.9%)	192 (6.1%)	0.0024
Sex	Female	69 (65.1%)	1559 (49.8%)	**0.0026**
Fever	No	88 (83.0%)	2927 (93.4%)	
	Yes	18 (17.0%)	206 (6.6%)	**< 0.0001**
Cough	No	73 (68.9%)	2518 (80.4%)	
	Yes	33 (31.1%)	615 (19.6%)	**0.0053**
Sore throat	No	77 (72.6%)	2765 (88.3%)	
	Yes	29 (27.4%)	368 (11.7%)	**< 0.0001**
Dyspnoea	No	96 (90.6%)	3002 (95.8%)	
	Yes	10 (9.4%)	131 (4.2%)	**0.0180**
Myalgia	No	72 (67.9%)	2883 (92.0%)	
	Yes	34 (32.1%)	250 (8.0%)	**< 0.0001**
Rhinorrhea	No	70 (66.0%)	2452 (78.3%)	
	Yes	36 (34.0%)	681 (21.7%)	**0.0041**
Respiratory discomfort	No	90 (84.9%)	2961 (94.5%)	
	Yes	16 (15.1%)	172 (5.5%)	**< 0.0001**
Nausea/ vomit	No	94 (88.7%)	3025 (96.6%)	
	Yes	12 (11.3%)	108 (3.4%)	**< 0.0001**
Headache	No	50 (47.2%)	2399 (76.6%)	
	Yes	56 (52.8%)	734 (23.4%)	**< 0.0001**
Prostration	No	83 (78.3%)	2968 (94.7%)	
	Yes	23 (21.7%)	165 (5.3%)	**< 0.0001**
Ageusia/ anosmia	No	87 (82.1%)	3051 (97.4%)	
	Yes	19 (17.9%)	82 (2.6%)	**< 0.0001**
Obesity	No	96 (90.6%)	3038 (97.0%)	
	Yes	10 (9.4%)	95 (3.0%)	**< 0.0001**
Sought health assistance	Hospital	8 (7.5%)	130 (4.1%)	
	None	81 (76.4%)	2764 (88.2%)	
	Basic Health Unit	8 (7.5%)	121 (3.9%)	
	Emergency Care Unit	9 (8.5%)	118 (3.8%)	**0.0032**
Household contact with symptomatic person	No	71 (67.0%)	2528 (80.7%)	
	Yes	35 (33.0%)	605 (19.3%)	**0.0007**

Figure 1: Covid-19 pandemic progression in Betim. (A) Absolute number of new cases according to official city statistics. (B) Cumulative number of cases according to official city statistics. Black dots indicate estimated overall prevalence (immunological and molecular tests) in the current study with its 95% confidence interval. Distance from black dots and red curve represent underreporting. (C-E) Dispersion of positive molecular tests across each survey. In the third survey (panel E), most populated areas of the city already had a non-null probability of presenting residents with a positive molecular test. (F-G) Overall prevalence (immunological and molecular tests) comparison in each of the ten administrative regions of the city across successive surveys. An increase was observed in most areas from the first to the second survey and, more substantially, from the second to the third survey.
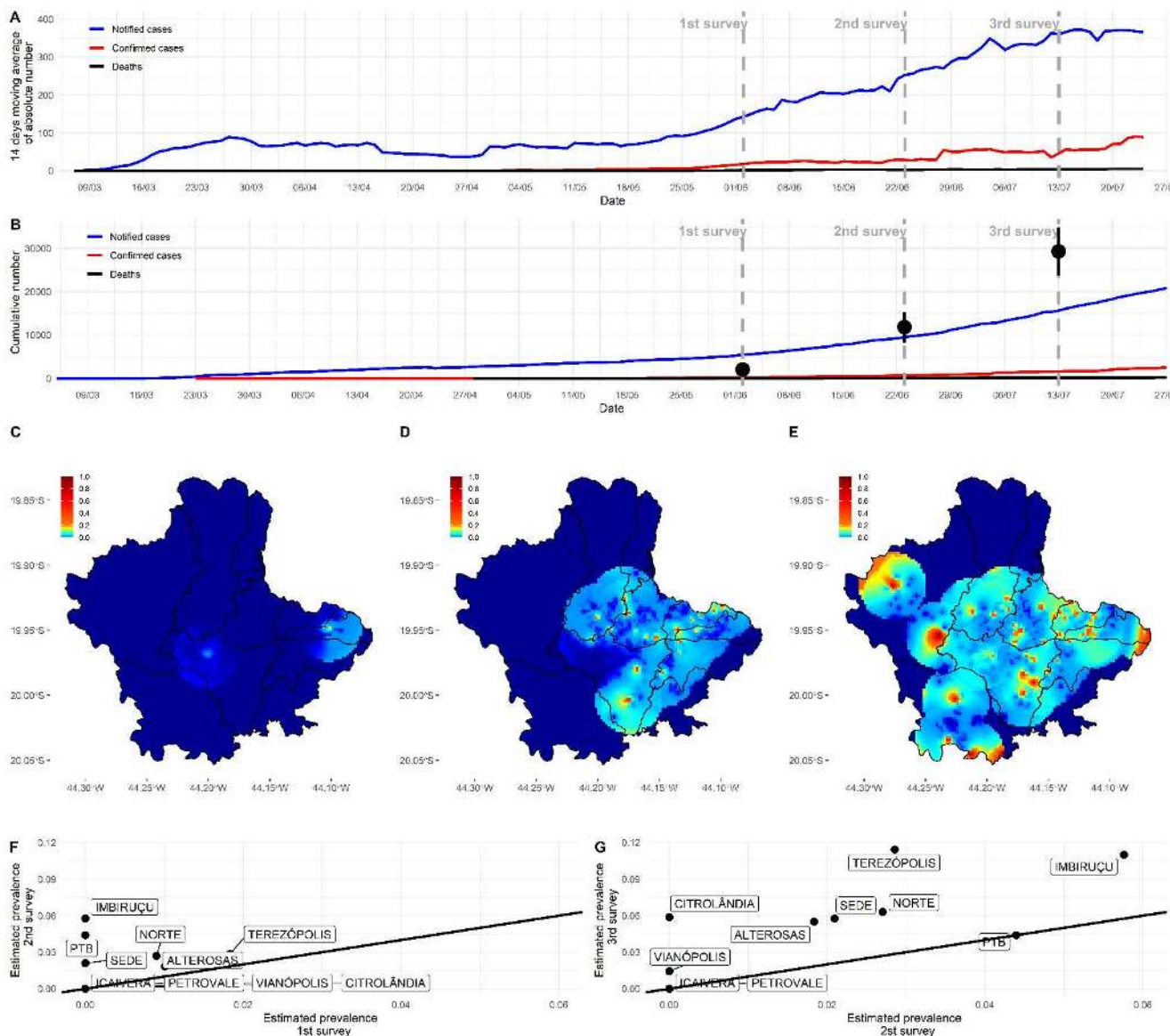
Figure 2: Phylogenetic characterization of SARS-CoV-2 genomes characterized in Betim. A maximum-likelihood tree was inferred on IQ-Tree under the GTR+F+I+G4 model with a comprehensive reference dataset, encompassing all Brazilian sequences plus one international sequence per country per week, from late 2019 to January 12 2021 ($n = 3,814$). The phylogeny depicted exhibits a subtree of 2,023 tips that harbours all relevant diversity considered for this study, mainly lineages B.1.1.28 (light salmon) and B.1.1.33 (light blue) where the novel genome sequences sparsely clustered. Tip shapes mark sequences characterized in this study. The scale bar indicates average nucleotide substitutions per site.
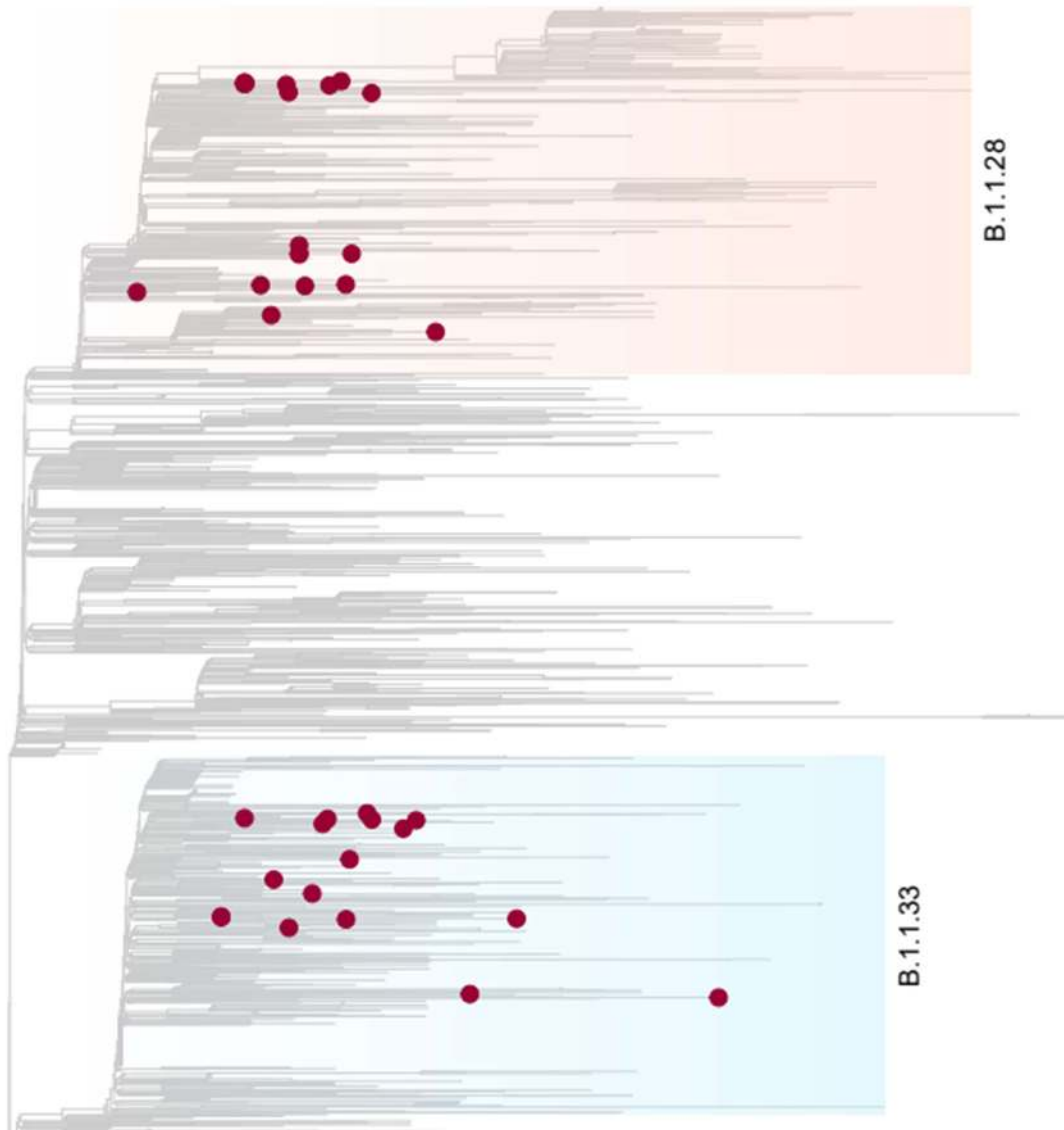
Figure 3: Spread of B.1.1.28 and B.1.1.33 lineages in Betim city. (A) Time-resolved maximum clade credibility phylogeny from a dataset comprehending 240 publicly available B.1.1.28 sequences and the 18 genomes generated in this study. **(B)** Time-resolved maximum clade credibility phylogeny from a dataset including 267 publicly available B.1.1.33 sequences and the 17 genomes generated in this study. For both analyses, the HKY+I+G4 nucleotide substitution model was used. The diamond indicates sequences from Betim city obtained in this study. The trees inferred are available on https://github.com/LBI-lab/SARS-CoV-2_phylogenies.git.

Table S1: Sequencing statistics.

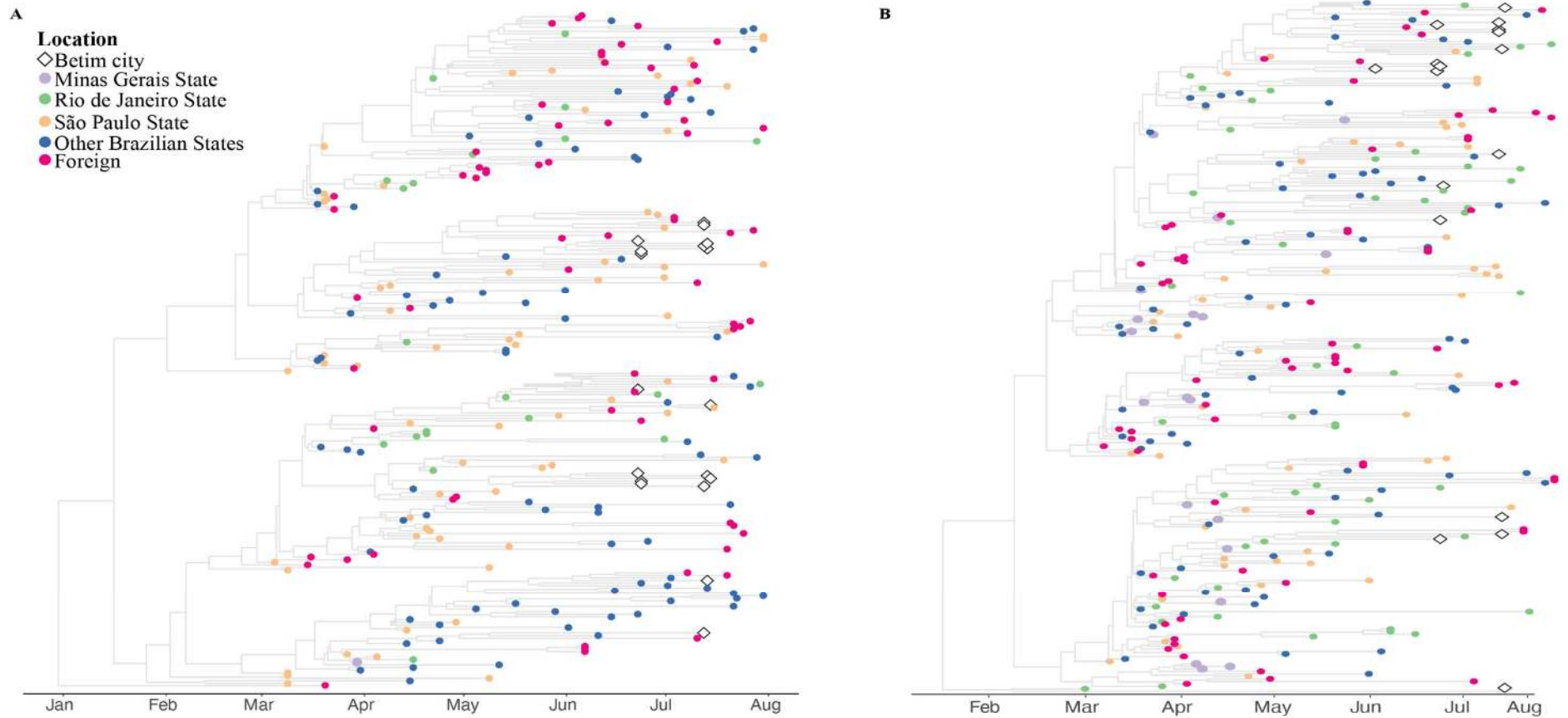| ID	Raw	Paired_filtered	Unpaired_filtered	Mapped	Average_depth	Coverage
LBIb1520	790922	614298	86891	680628	2670.83	0.998562065
LBIb1756	895726	698392	96497	756729	2972.99	0.998495185
LBIb1853	595660	110010	221855	312410	996.48	0.949438202
LBIb1521	970750	684820	140045	787387	3056	0.998361423
LBIb1367	898558	520784	172820	321874	1262.4	0.998327983
LBIb1730	604232	162584	211009	118340	449543	0.992409042
LBIb1834	713926	506904	100255	519339	2027.59	0.998361423
LBIb1338	443152	218098	99202	144518	562251	0.996956929
LBIb1155	587778	276476	135914	60791	239	0.973481808
LBIb1128	589594	280462	142500	64184	250	0.976056715
LBIb1957	1096024	975312	32986	854776	4308.04	0.877407705
LBIb0013	288132	240058	3544	233727	1080.49	0.991873997
LBIb2769	555052	486824	11649	250565	1210.8	0.806046014
LBIb2626	312768	260432	8929	210031	1129.37	0.943151418
LBIb2405	931682	830622	12449	812147	4070.16	0.993077849
LBIb2427	245794	200684	9298	102113	532031	0.790730337
LBIb2421	970744	866408	19461	797757	3964.35	0.935995185
LBIb2791	410522	373574	5169	370680	2036.86	0.990001338
LBIb2224	2768498	2618126	30325	2600398	13991.7	0.993010968
LBIb2933	680422	622700	7978	613536	3254.58	0.973147405
LBIb2624	556428	511752	7062	507512	2796	0.9179374
LBIb2621	802890	747922	10676	717957	4050.13	0.977962814
LBIb2905	1291410	1067698	50017	1000586	5093.82	0.955557785
LBIb3231	3115324	2888612	39358	2854534	15183.4	0.99705725
LBIb2256	949972	892682	10741	883323	4972.17	0.97950107
LBIb2808	834356	739114	14535	723412	3741.48	0.943820225
LBIb2964	822522	722510	12942	314131	1762.05	0.934691011
LBIb3167	1060064	972588	12679	965926	5015.61	0.993846977
LBIb2674	1398626	1279472	28774	1236914	6484.77	0.9941145
LBIb1905	2295810	1661904	174732	1370724	6910.15	0.97752809
LBIb1806	2111520	1906066	61989	1885222	10040.7	0.989332531
LBIb1706	2368546	1554024	290025	1170912	6275.13	0.976290797
LBIb2296	2508976	1884492	202645	1600481	7961.24	0.991372392
LBIb2892	2369302	2159526	63471	2149739	11672.1	0.991673355
LBIb2494	2285012	2088736	52678	1919668	10072	0.986824505

Table S2: Association of clinical and epidemiological data with a positive test (serological or molecular). Bolded p values indicate p < 0.05.

Variable	Level	Positive	Negative	p-value
	First	5 (4.7%)	1074 (34.3%)	
	Second	29 (27.4%)	1051 (33.5%)	
Survey	Third	72 (67.9%)	1008 (32.2%)	**< 0.0001**
	Alterosas	18 (17.0%)	616 (19.7%)	
	Citrolândia	4 (3.8%)	215 (6.9%)	
	Icaivera	0 (0.0%)	62 (2.0%)	
	Imbiruçu	32 (30.2%)	533 (17.0%)	
	Norte	11 (10.4%)	322 (10.3%)	
	Petrovale	0 (0.0%)	41 (1.3%)	
	PTB	8 (7.5%)	282 (9.0%)	
	Sede	15 (14.2%)	568 (18.1%)	
	Terezópolis	17 (16.0%)	302 (9.6%)	
Administrative Regions	Vianópolis	1 (0.9%)	192 (6.1%)	**0.002424**
Sex	Female	69 (65.1%)	1559 (49.8%)	**0.002642**
	Male	37 (34.9%)	1574 (50.2%)	
	0-5	3 (2.8%)	214 (6.8%)	
	06-19	15 (14.2%)	635 (20.3%)	
	20-39	42 (39.6%)	1025 (32.7%)	
	40-59	31 (29.2%)	840 (26.8%)	
Age range	Above60	15 (14.2%)	419 (13.4%)	0.190538
	No	106 (100.0%)	3119 (99.6%)	
International travel	Yes	0 (0.0%)	14 (0.4%)	1
	No	88 (83.0%)	2927 (93.4%)	
Fever	Yes	18 (17.0%)	206 (6.6%)	**0.000075**
	No	73 (68.9%)	2518 (80.4%)	
Cough	Yes	33 (31.1%)	615 (19.6%)	**0.005304**
	No	77 (72.6%)	2765 (88.3%)	
Sore throat	Yes	29 (27.4%)	368 (11.7%)	**0.000003**
	No	96 (90.6%)	3002 (95.8%)	
Dyspnoea	Yes	10 (9.4%)	131 (4.2%)	**0.018051**
	No	72 (67.9%)	2883 (92.0%)	
Myalgia	Yes	34 (32.1%)	250 (8.0%)	**< 0.0001**
	No	70 (66.0%)	2452 (78.3%)	
Rhinorrhea	Yes	36 (34.0%)	681 (21.7%)	**0.004198**
	No	90 (84.9%)	2961 (94.5%)	
Respiratory discomfort	Yes	16 (15.1%)	172 (5.5%)	**0.000079**
	No	94 (88.7%)	3025 (96.6%)	
Nausea/ vomit	Yes	12 (11.3%)	108 (3.4%)	**0.000075**
	No	50 (47.2%)	2399 (76.6%)	
Headache	Yes	56 (52.8%)	734 (23.4%)	**< 0.0001**
	No	83 (78.3%)	2968 (94.7%)	
Prostration	Yes	23 (21.7%)	165 (5.3%)	**< 0.0001**
	No	95 (89.6%)	2933 (93.6%)	
Diarrhea	Yes	11 (10.4%)	200 (6.4%)	0.150267
	No	105 (99.1%)	3102 (99.0%)	
Conjunctivitis	Yes	1 (0.9%)	31 (1.0%)	1
	No	87 (82.1%)	3051 (97.4%)	
Ageusia/ anosmia	Yes	19 (17.9%)	82 (2.6%)	**< 0.0001**
	No	102 (96.2%)	3081 (98.3%)	
Loss of voice	Yes	4 (3.8%)	52 (1.7%)	0.206495
	No	104 (98.1%)	3105 (99.1%)	
Pneumopathy	Yes	2 (1.9%)	28 (0.9%)	0.593166
	No	106 (100.0%)	3094 (98.8%)	
Chronic neurological disease	Yes	0 (0.0%)	39 (1.2%)	0.482095
	No	104 (98.1%)	3107 (99.2%)	
Pregnant	Yes	2 (1.9%)	26 (0.8%)	0.533509
	No	105 (99.1%)	3125 (99.7%)	
Postpartum	Yes	1 (0.9%)	8 (0.3%)	0.699884
	No	106 (100.0%)	3037 (96.9%)	
Chronic cardiovascular disease	Yes	0 (0.0%)	96 (3.1%)	0.123958
	No	103 (97.2%)	3086 (98.5%)	
Chronic kidney disease	Yes	3 (2.8%)	47 (1.5%)	0.489013
	No	96 (90.6%)	3038 (97.0%)	
Obesity	Yes	10 (9.4%)	95 (3.0%)	**0.000721**
	No	102 (96.2%)	2964 (94.6%)	
Asthma	Yes	4 (3.8%)	169 (5.4%)	0.609912
	No	104 (98.1%)	3113 (99.4%)	
Immunodepression	Yes	2 (1.9%)	20 (0.6%)	0.348298
	No	106 (100.0%)	3118 (99.5%)	
Chronic liver disease	Yes	0 (0.0%)	15 (0.5%)	1
	No	100 (94.3%)	2911 (92.9%)	
Diabetes	Yes	6 (5.7%)	222 (7.1%)	0.71047
Hypertension	No	91 (85.8%)	2585 (82.5%)	0.445923

	Yes	15 (14.2%)	548 (17.5%)	
	No	106 (100.0%)	3129 (99.9%)	
Transplanted	Yes	0 (0.0%)	4 (0.1%)	1
	No	106 (100.0%)	3110 (99.3%)	
Cancer	Yes	0 (0.0%)	23 (0.7%)	0.766303
	No	73 (68.9%)	2211 (70.6%)	
Any comorbidity	Yes	33 (31.1%)	922 (29.4%)	0.787174
	Hospital	8 (7.5%)	130 (4.1%)	
	None	81 (76.4%)	2764 (88.2%)	
	Basic Health Unit	8 (7.5%)	121 (3.9%)	
Sought health assistance	Emergency Care Unit	9 (8.5%)	118 (3.8%)	**0.003285**
	No	103 (97.2%)	3098 (98.9%)	
Admitted in health institution	Yes	3 (2.8%)	35 (1.1%)	0.249183
	No	71 (67.0%)	2528 (80.7%)	
Household contact with symptomatic person	Yes	35 (33.0%)	605 (19.3%)	**0.000774**

Table S3: Acknowledgement to sequences obtained from GSAID.