Genomics and epidemiology of SARS-CoV-2 in Brazil

Darlan da Silva Candido

Merton College, University of Oxford

A thesis submitted for the degree of *Doctor of Philosophy*

Hilary 2022

Abstract

Genomics and epidemiology of SARS-CoV-2 in Brazil

Darlan da Silva Candido, Merton College, University of Oxford

A thesis submitted for the degree of Doctor of Philosophy, Hilary, 2022

As of the 24th January 2021, it is estimated that the coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to over 350 million reported cases and over 5.6 million deaths worldwide. Brazil has the third highest case count, over 24 million, and the second highest death count, over 623,000. In this thesis, I apply genomic and epidemiological approaches to describe and understand SARS-CoV-2 importation, transmission, spread, evolution and response during the first year of the COVID-19 pandemic in Brazil.

Chapter 2 provides and overview of the early importation, spread and response. I start by identifying the probable air routes for SARS-CoV-2 importation into Brazil. I also provide a description of the first SARS-CoV-2 cases reported in Latin America, followed by epidemiological estimates of the basic reproduction number for the most affected Brazilian states. This chapter ends with a description of the implementation and easing of non-pharmaceutical interventions (NPIs) in 72.3% of the Brazilian municipalities.

In Chapter 3, I couple genomic insights obtained from a novel representative dataset of 427 SARS-CoV-2 genomes from Brazil with human mobility data to describe SARS-CoV-2 importation and genomic diversity, reconstruct SARS-CoV-2 nationwide spatial spread and investigate the impact of NPIs implemented in Brazil.

Chapter 4 covers the application of genomic epidemiology approaches to the identification and description of new SARS-CoV-2 variants of concern (VOCs). I describe the first two cases of the Alpha VOC in Brazil and provide a genomic characterization of the first cases of the Gamma VOC in Manaus, north Brazil.

Finally, I apply epidemiological and genomic approaches to uncover the dynamics of hospital-associated transmission in the largest hospital complex in Latin America. Chapter 5 shows evidence for SARS-CoV-2 within-hospital transmission to be higher in non-COVID-19 hospitals.

Preface

As with any piece of epidemiological work, the next pages you are about to read are full of complicated mathematics and difficult biological terms. However, they attempt to describe and understand one of the worst moments in humankind history. As of 24th January 2021, the COVID-19 pandemic has taken at least 5.6 million lives, not to mention the numerous people who will suffer with the consequences of long-COVID, economic depression and mental health issues. Please, also look at this piece of work as a snapshot of the history of the pandemic in Brazil, of those directly affected, and of those who devoted their time and skills to study and respond to it, while also going through the same struggles. There is more in these pages than just science.

That being said, there is also hope. In less than one year of pandemic, global scientific efforts have been able to develop several vaccines effective against SARS-CoV-2, which is unheard of in vaccine history. It is estimated by the WHO Regional Office for Europe that at least 470,000 lives have already been saved in the region by such vaccines in those aged 60 years and over only. Science has also been able to save lives through the implementation of masks, discovery of effective drugs for mitigating symptoms and the use of monoclonal antibodies. Scientific advances have also led to the highest genomic output of any outbreak in history with over 7 million genomes made publicly available on GISAID to date. So please, do also see these pages as a testament of what science can do in our lives and how important it is. But also see beyond that.

Acknowledgments

First and foremost, I would like to thank my main supervisor, Professor Nuno Faria, for believing in me and guiding me through such a beautiful, but also bumpy, road. I started my time in Oxford as a DPhil in Medical Sciences studying immunology of inherited cardiomyopathies. After some struggling, I decided to change back to something related to public health of infectious diseases, something more meaningful to the Brazilian population. Very randomly, I came across Nuno's Zoology page and, in a very cheeky move, I sent him an email. Next thing I know, we had an amazing meeting in which Nuno trusted my academic potential, even though I made it clear to him I knew nothing about phylogenetics, arboviruses (my pre-pandemic thesis) nor sequencing. The next couple of years were very intense for me, but Nuno was always there guiding me through skills I did not have. Then, the pandemic hit and those skills I had just learned were suddenly essential for a rapid response to the pandemic in Brazil, and resulted in this thesis. However, apart from all of that, the main thing Nuno has taught me is that anything is possible. I know, it sounds cliché, but it is true. And, for that, I will always be grateful.

A massive thank you to my co-supervisor, professor Oliver Pybus, for trusting my potential, for opening his research group to me and for all the guidance throughout this almost 4 years.

I would also like to thank professor Ester Sabino, my unofficial co-supervisor in Brazil, for all the opportunities, amazing research collaborations and all the guidance she has been giving me since we first met during my Master's in Brazil. It all means a lot to me.

A big thank you to the Evolve.Zoo team (Bernardo, Lucy, Sarah, Marina, Tanya, Alex, Sabrina, Louis, Moritz, Jessie and Jayna) for the everyday support and laughs. An especial thank you to Bernardo for being the best PhD-mate I could have ever asked for. Thank you for teaching me so much, for all the productive talks and walking along this

journey with me (watch out for this one, he is going to go far!). And also, a big thank you to Lucy Matkin for being the best project manager ever and saving my life so many times, but also for being such a dear friend (you rock girl!).

To my Brazilian collaborators at the Tropical Medicine Institute of the University of São Paulo thank you for welcoming me with open arms and making this journey so enjoyable. Thank you also for sharing some of the most challenging experiences as well in the heat of the early COVID-19 response in Brazil. Those venting therapy sessions really helped me keep my sanity. To Ingra, an especial thank you for teaching me everything I know about genome sequencing. Some of you have become dear friends and almost family (Ingra, Flávia, Mariana and Jaque), and I hope to keep these tight relationships for years to come.

To my friends from Oxford (Anjali, Luciana, Annika, Andreza, Daniel, Isabella, Licio, Joana, Nici, Carla, David, Raphael, Mariana, José Inácio, Bárbara, Natália, Vinícius, Ariane, Giacomo, Teodora, Beatrice and so many others) thank you so much for all the support but especially for all the good memories I will cherish from this time of our lives. To Anjali, my partner in academic life, thank you so much for all the walks, the calls and the daily shared life. There is more of you in this thesis than you can tell.

To my Brazilian friends (Jéssica, Cláudia, Marryer, Andressa, Herta, Amanda, Letícia, Luan, Junior, Alysson, Luiza, Pedro and so many others) thank you so much for always making me feel home and like nothing has changed whenever I go back. Thank you also for the mental health support through the difficult times in 2021.

To Ben, my boyfriend, thank you for being the kindest human being I have ever met. Thank you for supporting me throughout this journey and for the patience you had in the busy times. Thank you also for all the happiness and joy you bring into my life and for being more than a boyfriend, but a partner. I love you.

Last but not least, thank you to my family, especially to my Mom, my heroine. Thank you for having always trusted me and loved me unconditionally. I know it has not always been easy for you. This achievement is as much yours as it is mine. As for my dad, I wish he was here to see this. I love you.

"None of us, including me, ever do great things. But we can all do small things, with great love, and together we can do something wonderful."

– Mother Teresa

Statement of contributions and associated publications

The work presented in this thesis has mostly been developed over the last two years. My thesis was not initially meant to be about SARS-CoV-2, but rather about the genomic epidemiology of arboviruses such as Zika, dengue and yellow fever in Brazil. However, in early 2020, my research focus was shifted to reflect the urgent need to better understand how the pandemic that changed our modern world spread across different settings. In this thesis, you will find a compilation of my scientific contributions to the COVID-19 pandemic scientific response in Brazil. This work was highly collaborative, especially with the CADDE project team. My contributions to each manuscript in this thesis are stated below.

Chapter 2.1: Routes for COVID-19 importation in Brazil

I was responsible for all aspects of this work, from conception, to analysis, interpretation and writing. Flight data was provided by AW and KK. NRF supervised the entire work from conception. LA assisted with figures. All authors reviewed it.

***Candido DDS**, Watts A, Abade L, Kraemer MUG, Pybus OG, Croda J, de Oliveira W, Khan K, Sabino EC, Faria NR. Routes for COVID-19 importation in Brazil. J Travel Med. 2020 May 18;27(3).*

Chapter 2.2: Importation and early local transmission of COVID-19 in Brazil, 2020

This manuscript was a joint effort with collaborators in Brazil. JGJ and CS were responsible for genome sequencing of the first SARS-CoV-2 cases in Latin America, together with other collaborators from Brazil and with the support of AR, NJL and NRF. I was involved in data curation, analysis, preparation of figures, interpretation, writing and reviewing. ECS and NRF supervised and were involved in all aspects of this work.

Jesus JG de, Sacchi C*, **Candido D da S***, Claro IM, Sales FCS, Manuli ER, et al. Importation and early local transmission of COVID-19 in Brazil, 2020. Rev Inst Med Trop Sao Paulo. 2020 May 11;62:e30.*

Chapter 2.3: "Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil"

This manuscript was a joint effort with the CADDE project team. My main contribution to this work was providing an assessment of the early SARS-CoV-2 epidemic spread in Brazil by estimating basic reproduction number estimates for four Brazilian States and five countries, including Brazil. This analysis was developed in collaboration with AZ, especially for the model development. NRF supervised and was involved in all aspects of this work.

de Souza WM, Buss LF*, **Candido D da S***, Carrera J-P*, Li S*, Zarebski AE, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat Hum Behav. 2020 Aug;4(8):856–65.*

Chapter 2.4: **Dataset on SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities**

This work was led by AASS, myself and WMS. We equally contributed to data analysis, interpretation, study design, figures, writing and reviewing of the manuscript.

de Souza Santos AA, **Candido D da S***, de Souza WM*, Buss L, Li SL, Pereira RHM, et al. Dataset on SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities. Sci Data. 2021 Mar 4;8(1):73.*

Chapter 3: Evolution and epidemic spread of SARS-CoV-2 in Brazil

I led this work together with NRF and ECS and was involved in all aspects of it from conception. This work was highly collaborative CADDE project initiative involving the generation of a large representative genomic dataset from Brazil. I oversaw the generation of genome sequences including sampling strategy and metadata collection, together with genome sequencing teams in Brazil. Performed genomic and epidemiological data curation. Performed all phylogenetic analysis (with NRF, SD and PL), interpretation, writing and revisions with NRF, while overseeing all other aspects of this work. Assessment of the impact of NPIs through estimation of the reproduction number over time was performed by colleagues at the Imperial College London.

Candido DS*, *Claro IM*, de Jesus JG*, Souza WM*, Moreira FRR*, Dellicour S*, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020 Sep 4;369(6508):1255–60.*

Chapter 4.1: Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020

This work was mainly led by IMC and NRF. I was responsible for the phylogenetic analysis and interpretation with NRF and IMC. I was also involved in writing and reviewing the manuscript.

Claro IM, da Silva Sales FC, Ramundo MS, ***Candido DS***, Silva CAM, de Jesus JG, et al. Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020. Emerging Infect Dis. 2021 Mar;27(3):970–2.

Chapter 4.2: Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil.

This work was also highly collaborative, especially between the CADDE project team and the Imperial College team. Together with IMC, I oversaw the generation of all genome sequences and performed metadata collection, since the beginning of the Manaus investigation which was initially published as a *virological.org* post and later extended into this manuscript. NRF and I were also responsible for the phylogenetic and phylogeographic analysis in this manuscript with the support of PL. Other authors were involved in data generation and in the modelling of P.1 epidemiological characteristics.

Faria NR*, Mellan TA*, Whittaker C*, Claro IM*, ***Candido D da S****, Mishra S*, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 2021 May 21;372(6544):815–21.

Chapter 5: SARS-CoV-2 hospital-associated transmission dynamics in São Paulo, Brazil: a retrospective genomic surveillance study

I led and I am responsible for all aspects of this work. I oversaw the generation of genome sequences with IMC and performed metadata collection with IMC and MSR. I am responsible for all epidemiological and phylogenetic analysis in this manuscript, interpretation, figures, writing and revising. NRF, ECS and SFC supervised the development of this work. It has recently been submitted to Clinical Infectious Diseases.

indicates shared first authorship

- Other manuscripts related to the work presented in this thesis

Prete CA, Buss L, Dighe A, Porto VB, ***Candido DS***, Ghilardi F, et al. Serial interval distribution of SARS-CoV-2 infection in Brazil. J Travel Med. 2021;28(2).

Claro IM, Ramundo MS, Coletti TM, da Silva CAM, Valenca IN, ***Candido DS,*** et al. Rapid viral metagenomics using SMART-9N amplification and nanopore sequencing. Wellcome Open Res. 2021 Sep 20;6:241.

Souza WM, Amorim MR, Sesti-Costa R, Coimbra LD, Brunetti NS, Toledo-Teixeira DA, et al. Neutralisation of SARS-CoV-2 lineage P.1 by antibodies elicited through natural SARS-CoV-2 infection or vaccination with an inactivated SARS-CoV-2 vaccine: an immunological study. Lancet Microbe. 2021 Oct;2(10):e527–35.

Romano CM, Felix AC, Paula AV, Jesus JG, Andrade PS, ***Candido DS***, et al. SARS-CoV-2 reinfection caused by the P.1 lineage in Araraquara city, Sao Paulo State, Brazil. Rev Inst Med Trop Sao Paulo. 2021;63:e36.

Brizzi A, Whittaker C, Servo LMS, Hawryluk I, Prete CA, de Souza WM, et al. Report 46: Factors driving extensive spatial and temporal fluctuations in COVID-19 fatality rates in Brazilian hospitals. medRxiv. 2021 Nov 2; Currently submitted to Nature Medicine;

Gutierrez B, Castelan HG, ***da Silva Candido D***, Jackson B, Fleishon S, Ruis C, et al. Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America. medRxiv. 2021 Nov 21;

Gutierrez B, Márquez S, Prado-Vivar B, Becerra-Wong M, Guadalupe JJ, ***Candido DDS***, et al. Genomic epidemiology of SARS-CoV-2 transmission lineages in Ecuador. Virus Evol. 2021 Jun 4;7(2):veab051.

- Other manuscripts published during my DPhil but not related to the work developed in this thesis

Hill SC, (…), ***Candido DS***, et al. Emergence of the Asian lineage of Zika virus in Angola: an outbreak investigation. Lancet Infect Dis. 2019 Oct;19(10):1138-1147. doi: 10.1016/S1473-3099(19)30293-2.

de Lima STS*, de Souza WM*, Cavalcante JW*, ***Candido DS****, Fumagalli MJ*, Carrera JP, et al. Fatal Outcome of Chikungunya Virus Infection in Brazil. Clin Infect Dis. 2021;73(7):e2436-e43.

De Jesus JG, Dutra KR, Salles FCS, Claro IM, Terzian AC, ***Candido DS***, (...), et al. Early identification of dengue virus lineage replacement in Brazil using portable genomic surveillance. Memórias do Instituto Oswaldo Cruz.

Cunha MS, Faria NR, Caleiro GS, ***Candido DS***, et al. Genomic evidence of yellow fever virus in Aedes scapularis, southeastern Brazil, 2016. Acta Trop. 2020 Feb 7;205:105390. doi: 10.1016/j.actatropica.2020.105390.

Naveca FG, Claro I. Giovanetti M, de Jesus JG, Xavier J, Iani FCM, ***Candido DS.***, et al. Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon. PLoS Negl Trop Dis. 2019;13(3):e0007065.

Xavier J, Giovanetti M, Fonseca V, Thézé J, Gräf T, Fabri A, et al. Circulation of chikungunya virus East/Central/South African lineage in Rio de Janeiro, Brazil. PLoS ONE. 2019 Jun 11;14(6):e0217871.

As Darlan's supervisor, I confirm the information presented here is representative of Darlan's contributions to the work listed in this thesis.

Prof. Nuno R. Faria

Reader in Virus Evolution
Department of Infectious Disease Epidemiology
School of Public Health
Imperial College London

Associate Professor
Department of Zoology
University of Oxford

United Kingdom

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid increase in human mobility, population size and density in urban areas, land use, migration accompanied by climate change, natural disasters and civil conflicts have been some of the major drivers of the emergence and re-emergence of infectious diseases (1-4). A review published in 2008 concluded that most of the 1,399 known human pathogens are bacteria (38.67%), fungi (23.2%), helminths (20.4%) and viruses (13.5%). However, viruses disproportionally account for most of the recently described human pathogens (from 1980 onwards), 66.7% (58/87), which mostly emerged in human populations from spill over events from animal reservoirs (1). Most of these new pathogenic viruses, 76.3% (45/58), are single-stranded RNA viruses, e.g. the Human Immunodeficiency Viruses type 1 and type 2 (HIV-1 and HIV-2), Cote d'Ivoire and Reston Ebola viruses, coronaviruses such as HKU1, NL63 and the severe acute respiratory syndrome virus (SARS) (1).

Several zoonotic viruses have achieved pandemic spread and affected thousands/millions of people worldwide. The HIV-1 pandemic has taken approximately 32 million lives; SARS in 2002 caused 8096 cases and 774 deaths, and the pandemic A(H1N1) 2009 influenza virus caused between 151,700-575,400 deaths (5). We have also seen the re-emergence and international spread of previously described viruses, such as the chikungunya and Zika epidemics in the Americas from 2013 (6) and 2015 (7), which occurred in tandem with endemic circulation of dengue virus (8). More recently, the world has been facing the biggest public health threat of the century – the emergence and pandemic

spread of a new human coronavirus, the severe acute respiratory syndrome virus coronavirus-2 (SARS-CoV-2), causative agent of the coronavirus disease 2019 (COVID-19) (9). Since its emergence in late 2019 (9), and as of 24th January 2022, SARS-CoV-2 spread globally and has caused at least 350 million cases and 5.6 million deaths (10). Understanding virus transmission, spread, evolution and its application to public health has never been timelier in an increasingly connected world.

Epidemiology can be defined as "the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the prevention and control of health problems" (11). In its broader definition, epidemiology does not only concern infectious/communicable diseases, however several epidemiological contributions to public health are infectious disease-related, especially viral epidemiology; for instance, the eradication of smallpox in the 1980's after a massive global campaign led by the WHO. Another example is the identification of risk factors, transmission routes, and effective interventions capable of reducing the spread and increasing the life expectancy and quality of life of people living with HIV (PLWH) (11). In this sense, viral epidemiology is not only interested in the distribution and determinants of diseases/health conditions caused by viruses but also in the dynamics of these viral infections in the population and the interaction between viruses and their hosts (12). Traditional virus epidemiology approaches rely on the description of where, when, who and how, mainly by investigating and documenting cases and their characteristics, while also using mathematical approaches to understand disease distribution, identify risk factors and model disease dynamics (12-14). However, although extremely important and informative, epidemiological methods of disease surveillance are inherently prone to several types of biases and other limitations, which may affect results and conclusions driven from this type of data and analysis (15). Such limitations include, but are not limited to, lack of information on pathogen biological

characteristics, heterogeneity of collected metadata quality within- and between-regions and contact tracing inherent biases (selection, recall and response) (15).

Modern epidemiology relies on the integration of several scientific and technological advances to its traditional approaches to circumvent some of these ingrown obstacles and expand its ability to understand infectious disease dynamics. These advances span from the increased capacity of generation and analysis of laboratory surveillance and outbreak data, to computer sciences and digital health-related approaches (14). For instance, social media have been used to track and predict the spread and transmission of infectious diseases (16); anonymised human mobility data has been integrated into epidemiological models by using satellite and mobile phone data; and the advances in virus genome sequencing and genetic analysis now allow us to track the transmission and evolution of viral pandemics in near- to real-time (14). This thesis relies on the extensive use of genomic epidemiology coupled with traditional epidemiological methods and, when possible, the use of human mobility data to uncover and understand virus transmission, spread and evolution in Brazil.

1.1 Traditional Epidemiology Approaches to Infectious Diseases

Traditional epidemiology applies standard frequency measures and methods to understand disease distribution and association with population and environmental characteristics. To better understand disease distribution, the most basic measures are incidence, the number of new cases of an infection/disease in a population over a period of time and prevalence, the total number of cases of an infection/disease in a population at a certain point in time. Both measures can be highly informative of the dynamics of the measured outcome, especially when tracked over time (12).

However, the epidemiology of infectious diseases differs from that of non-communicable diseases as events result from the transmission between individuals and, as such, are not independent from each other. Understanding transmission and how to control it are two of the main objectives of viral epidemiology. Several factors can influence virus transmission, including incidence, prevalence, mode of pathogen transmission, pathogen susceptibility, population density and mobility patterns (13). For this reason, mathematical models designed to understand and predict virus transmission within a population usually account for different key parameters for virus transmission and also individual infection status (compartments). The most common infection compartments are susceptible, exposed, infectious, recovered or immune. The SIR model is the simplest of all compartmental models as it only assumes 3 compartments: susceptible, infectious and recovered. In SIR models, individuals are assumed to become infected and infectious after contact with another infected individual and subsequently to become non-infectious (recovery) (17). However, the type of infection compartments will vary according to the pathogen and type of transmission model (18). The transmission rate of a virus in a given population is also affected by biological factors derived from the contact with its host, such as (1) the latent infection period, defined as the time between infection and being infectious, (2) average infectiousness, defined by the average transmissibility and contact rate, (3) incubation period, the time between infection and symptom onset, and (4) the serial interval, defined as the time between symptom onset in a pair infector and infected (19).

The extent of transmission of a pathogen can be estimated through epidemiological parameters such as the basic reproduction number, R_0. This represents the average number of secondary cases derived from one single infectious individual in a completely naïve and susceptible population. As such, an $R_0 > 1$ means a growing epidemic, while an $R_0 < 1$ means the epidemic is declining. It is often used to understand how fast a pathogen can spread in

specific populations, the magnitude an epidemic can reach and the extent of the mitigation strategies needed to control its spread (20). For instance, R_0 can be used to estimate the proportion of a population who would need to acquire immunity to a pathogen so that the threshold in which transmission is interrupted, the so-called herd immunity, can be reached (21). This is a concept particularly useful in vaccinology to estimate the number of people who should be targeted by vaccination programs to prevent new outbreaks. This proportion can be estimated by the following equation: $P= (1-1/R_0)$ (20, 22). Thus, an R_0 of 2 would require 50% of the population to be immune to a pathogen to stop its spread, while for R_0 of 3, 4, 5 and 10 this proportion would be 66.7%, 75%, 80% and 90%, respectively (22). Examples of currently known R_0 are 14.5 for the 1960-68 measles epidemic in Ghana, 3.5 for the 2002-2003 severe acute respiratory syndrome (SARS) epidemic, 1.51 for 2014 Ebola virus epidemic in Guinea, 1.51 for the pandemic A(H1N1) 2009 influenza virus in South Africa, and 6 for the 1955-1960 poliomyelitis epidemic in Europe (18).

R_0 is usually calculated as a function of the probability of transmission between infected and susceptible individuals (β), the type and duration of contact (c), and the duration of the infectious period (D) ($R0=\beta cD$) (22). Although often considered a fixed biological measure for each pathogen, R_0 is a measure of transmissibility that is influenced by the biological characteristics of a pathogen, the sociodemographic characteristics of the population in which it is spreading and the environment (23). It is usually calculated using individual-level data or population-level data. In theory, the simplest way of calculating R_0 is through epidemiological surveillance and contact tracing of all primary cases of an infection and the secondary cases linked to them. However, this often becomes impractical in large outbreaks (24). On the other hand, population-based models estimate R by applying ordinary differential equations to cumulative incidence data, while making assumptions on individual-level parameters, such as contact rate, duration of infectious period and

probability of infection upon contact (often named as infectivity and susceptibility) (25). However, R_0 values achieved by individual and population-level data approaches cannot be compared, as the R_0 obtained from population-level data represents the value of a threshold parameter, which governs whether an epidemic would occur, rather than the actual value of R_0 (26).

As R_0 is reported as an average and its 95% confidence interval, it will intrinsically fail to represent important local variation at different scales of analysis (27). For instance, R_0 might be higher in regions with higher population density, human mobility or even given cultural factors, such as frequency and intensity of human contact and other practices (28, 29). This is particularly important when R_0 is used to inform public health interventions for pathogens for which superspreading events account for a large share of the transmission events, as these events will skew R_0 towards higher levels (30), even though transmission might be low for most infected individuals. In fact, previously estimated R_0, such as for measles, when R_0 of 12-18 were estimated with data collected between 1912-1928 in the United States and 1944-1979 in the United Kingdom, have probably become obsolete as characteristics of our society and its organization have dramatically changed since they were first estimated (23).

When estimated during ongoing epidemics rather than at their beginning, the reproduction number is then referred to as R effective (R_e), as the pool of susceptible individuals decreases and the proportion of immune individuals increases through natural exposure or vaccine-mediated immunity, thus not representing a completely naïve population anymore. As an epidemic progresses, R_e generally becomes lower than R_0 as a result both of the implementation of public health interventions and of the depletion of susceptible individuals in the population (31). R_e can also be estimated for specific moments in time, and as such, it is called instantaneous reproduction number or R_t. R_t has been used

to evaluate the effectiveness of public health interventions (e.g., vaccination, social distancing, isolation) in reducing pathogen epidemic spread in given populations (31-39). For instance, timeseries of basic epidemiological measures, incidence and prevalence, and R_t estimates were used to assess the impact of public health interventions during the SARS epidemic in Hong Kong in 2003 (31).

For relying on data usually collected through syndromic or disease-specific surveillance of infectious diseases, instantaneous measures of transmissibility such as *Re* or *Rt* are subject to the inherent limitations and biases faced by other traditional epidemiology estimates, especially in low- and middle-income countries (LMICs). Such limitations include lack of standardization, limited laboratory capacity for diagnostic resulting in underestimation of cases and limitations regarding data processing and sharing (40). When relying on data from contact tracing strategies for the reconstruction of transmission chains, traditional epidemiology approaches are also prone to response, selection and recall bias. Even when such biases can be circumvented, information regarding lineage specific dynamics, number of lineage introductions into specific populations, or adequately identifying pairs of infectors and infected individuals cannot be retrieved from traditional epidemiology sources (15). Fortunately, the development of new genomic sequencing technologies and their integration with traditional epidemiology have provided insightful venues to complement and maximize epidemiological insights.

1.2 Genomic Epidemiology Approaches to Infectious Diseases

The 21st century, especially in the last decade, has seen an enormous growth in our capacity to perform whole genome sequencing (WGS) at increasingly faster pace and lower costs. Together with an increasing computational power to process and analyse such large amounts of genetic information, genome sequencing has been essential for the birth of a new

field of epidemiology. Genomic epidemiology can be defined as "the study of the transmission of infectious diseases through the use of pathogen genomes" (41). More so than only investigating pathogen genomic data, genomic epidemiology integrates genomic data generated during outbreaks with the associated epidemiological metadata to understand the dynamics of pathogen emergence, transmission and spread at different geographical and time scales. Some of the key objectives of genomic epidemiology studies can be: pathogen identification during an outbreak of infectious disease, identification of the origins of an outbreak or epidemic, investigate potential cross species transmission, understanding pathogen transmission and the key factors driving its spread, investigate transmission chains and tracking pathogen evolution at the within-host and population-level, and track the emergence of novel pathogen strains potentially associated with increased transmissibility, immune escape or disease severity (3). Genomic epidemiology relies on the analysis of pathogen genomic data using molecular phylogenetic tools to investigate the patterns of gene flow in pathogen epidemic histories.

Molecular phylogenetics is the field of biology that studies the evolutionary relationships between genetic or protein sequences and thereby their shared ancestry (42). The reconstruction of evolutionary relationships between biological entities (species, individuals, or genes) is usually depicted as a phylogenetic tree. Phylodynamics can be defined as "the study of how evolutionary, epidemiological and immunodynamic processes act and potentially interact to shape phylogenies" (43-45). Since its conception, the field has focused on understanding pathogen transmission dynamics and impact on diversity (43, 45). In turn, phylogeography focuses on the study of the principles and processes governing virus geographical spread through the analyses of spatial patterns across phylogenies (46).

Viruses are the ideal model organisms for understanding pathogen evolution, especially RNA viruses, given the fast pace at which their genomes undergo nucleotide

substitution and acquire new mutations, their short genomes and large population sizes (44, 46). These characteristics allow for the study of microevolutionary changes, since mutation fixation in such populations occur at the same timescale as the epidemiological and ecological processes shaping them while leaving a footprint in their genomes (46). The epidemiological and evolutionary insights obtained from reconstructing evolutionary, spatial and temporal dynamics from virus genomes often obtained during outbreaks now constitutes a key part of outbreak response and public health interventions/response(47).

1.2.1 Evolutionary Models

The reconstruction of viral evolutionary history through the inference of phylogenetic trees starts by the selection of an appropriate nucleotide substitution model. The simplest way of estimating the genetic distance between a pair of sequences is counting the number of nucleotide positions differing between them. However, such approach does not consider three important factors: (i) multiple nucleotide substitution events might have happened at the same site, (ii) rates of evolution might be different across different sites, and (iii) types of substitution differ in their rate of occurrence (48). For this reason, several mathematical models have been developed to reconstruct the nucleotide substitution process within a genetic sequence and thus adequately estimate genetic diversity and infer phylogenetic trees. Jukes Cantor 1969 (JC69) is the first and the simplest of these models and it assumes that rates of evolution across sites are equal, and that all types of nucleotide substitutions happen at the same rate and same frequency for all bases. Later, different rates of substitution for transitions and transversions were included in the Kimura 1980 (K80) model (49). Currently, the most complex and flexible nucleotide substitution model is the general time reversible model (GTR) which allows for different rates for all possible types of substitutions and different nucleotide frequencies (12 parameters) (50). Substitution

models can be further optimised by considering a proportion of invariable sites (I) and rate variation across sites using a gamma distribution (G). Although more than 1600 nucleotide substitution model combinations have been developed, with different parameter combinations, more complex models, such as GTR + G + I tend to provide a better fit to nucleotide datasets. However, several statistical tools such as JModelTest (51) are available and can be used to identify the most appropriate nucleotide substitution model for a given dataset, as selection of inadequate models can affect phylogenetic inference by under parametrisation or can become computationally costly by over parametrisation(52).

1.2.2 Population Dynamic Models

One of the main advances in the study of genomic epidemiology is the integration of coalescent models to study pathogen population dynamics. In coalescent theory, the history of sampled individuals can be represented as a genealogy, in which the most recent common ancestors of sampled individuals are represented by coalescent events (internal nodes) based on the divergence of the linages descending from them. Thus, the coalescent travels backwards from the moment individuals were sampled (present), and coalescent events continue to occur until the common ancestor of all sampled individuals (root) is reached and only one lineage remains (53, 54). Population genetics models developed under the coalescent theory can be used to link ecological processes and phylogenetic structure, by assuming that the pattern of coalescent events across the tree can inform on the occurrence of transmission events and population-level processes through time (45, 55), such as population size, migration, recombination, and selection. In fact, such models have been commonly used to understand viral transmission dynamics (demographic history, e.g., population growth and decline) within a population (55, 56) and infer important epidemiological parameters from genetic data, including the effective population size (N_e), growth rate (r), the doubling-time (λ) and the basic reproduction number (R_0) (46). RNA

viruses are ideal candidates for the application of coalescent-based models, given that their fast evolutionary rate often leads to well-resolved phylogenetic trees (46).

Several phylodynamic coalescent-based methods have been developed since the coalescent theory was first proposed in 1982 by Kingman (57). Over time, models have become increasingly complex in terms of the population dynamics they consider. Initial parametric models focused on simple constant, exponential growth (56) and logistic growth population sizes, while providing maximum likelihood (58) or Bayesian frameworks for estimation of key parameters of interest (59, 60). However, such models are most useful when the populations under study are known to follow such population dynamics *a priori* (59, 61). More recent coalescent models use non-parametric approaches for estimation of N_e, such as Skyline (59), Skyride (62) and Skygrid (63). These models approximate N_e as a linear function of parametric functions and allow inference of flexible time-changing population dynamics from heterochronous sequences. More recent developments include the ability to test the association between N_e and temporal covariates under a generalised linear model (GLM) framework while considering the shared ancestry and phylogenetic uncertainty (64).

Recently, birth-death (BD) models have been used as an alternative to coalescent methods for the inference of epidemiological parameters from genetic data. The rationale for using birth-death models comes from two limitations of coalescent-based methods, (i) not being able to differentiate between recovery and death, and (ii) assuming that only a small random proportion of cases have been sequenced (65). On the other hand, BD models estimate separate birth (transmission) and death rates, and explicitly consider sampling proportion as an additional parameter (65). For a completely susceptible population, these two rates can be used to estimate R_0. BD models have also been extended to nonparametric models allowing for time-changing infection rates (66). Birth-death models have

successfully been applied to investigate population dynamics of different viruses such as HIV (67) and hepatitis C virus (HCV) (66).

1.2.3 Molecular Clock Models

To study the temporal dynamics of virus epidemics, molecular clock models are typically used to convert genetic divergence into calendar time units (45, 68). For this reason, the sampling (or collection) dates of biological samples used for genome sequencing have become an essential piece of metadata to estimate time-calibrated phylogenetic trees and to assist with adequate phylogenetic rooting (69, 70) . Molecular dating has become an important tool for the understanding of viral epidemics. It has been used to date the introduction of viruses or new lineages into a specific region (71), identify surveillance gaps and cryptic transmission (71-73), date spill over events (74, 75), reconstruct historical spread (71, 73, 75-77) and even estimate the date of infection in an individual patient (78). Finally, dated phylogenies often form the basis to estimate important epidemiological parameters over time, such as the reproduction number, the impact of interventions, and factors associated with spread over time (43, 67, 73, 76, 77, 79, 80).

The earliest molecular clock models assumed that evolution occurred at a constant evolutionary rate across the phylogeny and were named as strict molecular clock models (81). This is a simplistic way of describing the relationship between divergence and time, and subsequent studies showed that strict molecular clock models did not often provide realistic descriptions of virus evolutionary histories (82-84). However, more recent models can account for variation in the evolutionary rate across different lineages in a phylogeny, and are named as relaxed molecular clock models (85). Examples of relaxed clock models include the local molecular clock models (86-88), autocorrelated molecular clock models (89, 90), uncorrelated-relaxed clock model (85) and the random local clock models (91).

The uncorrelated-relaxed clock model, in particular, has been an important improvement to previous approaches (local and autocorrelated) as it allows for co-estimation of phylogeny and divergence dates rather than requiring a fixed tree topology (85). The random local clock model, in turn, allows for a given number of evolutionary rates to be randomly assigned for specific subset of lineages from the tree, varying from 0 lineages (strict clock) to every single lineage in the phylogeny (uncorrelated-relaxed clock) (91).

1.2.4 Phylogeographic Models

As host population characteristics play a major role in shaping virus population structure, especially RNA viruses, initial phylogeographic analyses focused on understanding the information that phylogenetic tree topologies could provide on the geographical pattern of spread of different viruses (92). Accordingly, RNA virus phylogeographic dynamics can be divided into five patterns: (i) no clear structure, (ii) wave-like transmission, (iii) source-sink or core-satellite, (iv) gravity-like, and (v) strong spatial subdivision. It is possible for the same virus to present with different dynamics depending on the spatial scale. Such dynamics reflect relative rates of virus gene flow, how long such viruses have been associated to human populations and their mode of transmission (92).

Alongside population dynamics and dating of evolutionary events, advances in coalescent theory and molecular clock models have been fundamental for the development of tools for assessing the relationship between genetic evolution and spatial spread (92). Population structure can be investigated according to arbitrary discrete or continuous traits, such as geographic locations, body compartments, viral loads, host species, morphological characters, habitat preferences, antigenicity or cellular compartments (68). However, geographic locations are the most commonly used traits in phylodynamics and several statistical approaches have been developed for the analysis of geo-referenced datasets (68).

Most recent approaches rely on stochastic phylogenetic frameworks for inferring discrete and continuous diffusion under a suite of coalescent and molecular clock models (68, 93). Probabilistic models, especially Bayesian frameworks, allow for the integration of different evolutionary models and data sources, and offer increased flexibility in hypothesis testing compared to parsimonious approaches (64, 93, 94).

Discrete trait phylogeographic approaches model spatial diffusion using continuous-time Markov Chain (CTMC), initially introduced within a maximum likelihood framework by Pagel et al in 1999 (95), and later extended to Bayesian frameworks (96, 97). CTMCs are similar to common nucleotide, codon and amino acid substitutions models in its use of infinitesimal matrices of exchange rate between trait locations, which can be symmetrical or asymmetrical. Discrete trait phylogeographic analysis (DTA) is typically applied to infer spatial origins and spread patterns from virus genetic data with information on sampling location. The statistical performance of geographical DTA can be improved by applying a Bayesian stochastic search variable selection procedure (BSSVS), which allows location-exchange rates to be zero with some probability and identifies those that are more likely to explain virus spatial diffusion patterns (47, 93). Markov jumps and rewards can also be used to estimate the expected number of transitions between specific levels of the traits and the time spent at each trait level along the tree (waiting times) (97-99). For instance, Markov jumps can be used to estimate the number of imports and exports between specific countries (100, 101).

However, when fine-scale continuous geographical data is available, e.g., latitude and longitude of sample collection, continuous phylogeographic approaches may be preferred to discrete phylogeographic approaches. Continuous phylogeography assumes that traits evolve according to a Brownian motion process (random motion of particles suspended in a medium) and uses random walk models (succession of random steps) to infer

phylogeographic diffusion while also reconstructing evolutionary history (102). Such approaches have been recently used to reconstruct the continuous geographical diffusion of yellow fever virus in southeast Brazil (75) and rabies virus epidemic in North American racoons (102). Tools for the visualisation of discrete and continuous phylogeographic analysis have been recently developed, such as the program SpreaD3 ("Spatial Phylogenetic Reconstruction of Evolutionary Dynamics using Data-Driven Documents") (103) and the R package "seraphim" (104).

1.2.5 The Rise of Genomic Epidemiology

Although genomic epidemiology approaches have been applied to the study of several pathogens in the past decades, most of these investigations have been retrospectively performed, and near-real time or real-time applications have only recently been achieved. For instance, the 2002-2003 SARS-CoV epidemic was the first time genomic sequencing technologies were applied during an outbreak of international concern. However, within the first month after pathogen identification, only 3 sequences had been made available and 31 sequences within the first 3 months (105). Although genomic data was useful to uncover information about SARS-CoV origins, genomic structure, evolution, host interaction, and guided the design of diagnostic molecular assays, limited sequencing capacity and delays in data generation and sharing limited its public health impact (105-110).

During the influenza A(H1N1) 2009 pandemic (H1N1pdm), which was first detected in April 2009, genetic data was used for the first time, together with traditional epidemiological analysis to assess virus transmission and to aid in public policy decisions (105). 23 hemagglutinin A sequences and 11 whole-genomes sequences were used to assess the pandemic's potential by estimating its most recent common ancestor (TMRCA) around 2 January 2009 [95% credible interval (CrI): 3 November 2008 to 2 March 2009], a doubling time of 10 days (95% CrI: 4.5 to 37.5 days) and an R_0 of 1.22; 95% CrI: 1.05 to 1.60, similar

to the values found with epidemiological data and comparable to those of previous influenza pandemics (79). Such results were essential to understand the extent of mitigation measures needed for public health control. Further studies used larger genomic data to investigate H1N1pdm origins (111), revealing at least 3 months of cryptic transmission between the first detection and estimated TMRCA (72). Unfortunately, given limitations in sampling influenza in swine populations, the origins of H1N1pdm remained unknown until very recently, when a study published in 2016 suggested an origin in Central Mexico (112). The rapid sharing of H1N1pdm genetic data on the then recently created Global Initiative on Sharing Avian Influenza Data (GISAID) online platform was a key factor for early analyses and subsequent responses to the curb pandemic spread (113, 114).

Although useful for the study of SARS-CoV, H1N1pdm, and the Middle Eastern Respiratory Syndrome (MERS), it was during the 2013-2016 Ebola Virus (EBOV) Makona epidemic in Guinea, Liberia, Sierra Leone that genomic epidemiology was first conducted in real-time (115). Most early genomic epidemiological studies had been performed on tens to few hundreds of sequences, while much larger EBOV virus genome datasets were generated during the epidemic, representing 5% of all reported cases (115).

Phylogenetic analysis of EBOV Makona revealed important aspects about its zoonotic origins and transmission dynamics. Firstly, analyses of genomic data revealed that the EBOV epidemic was the result of a novel and independent cross-species transmission event into the human population (69, 116) and new lineages emerging during the outbreak descended from earlier ones rather than from new spill-over events from a zoonotic reservoir species. Phylogenetic analysis also estimated the EBOV epidemic to have started around December 2013, that EBOV Makona only diverged from other EBOV lineages about a decade before the outbreak ignited, and that all EBOV lineages share a common ancestor that dates back to around 1975, consistent with the first reported EBOV cases in 1976 (70,

73, 115). Continued virus genome sequencing and analysis during the epidemic was also crucial to clarify the evolutionary rate of EBOV Makona, which was debated across the scientific community and the media (115). Early estimates suggested evolutionary rates that were twice as fast compared to those estimated for previous EBOV outbreaks. However, later analyses showed that such results were related to the time-dependency of evolutionary rates, given that mildly deleterious variants would not have been eliminated yet by genetic drift (115, 117). Secondly, although no formal assessment of the phenotypic impact of amino acid mutations was performed at the time, genomic sequencing identified the first mutation on the EBOV glycoprotein receptor-binding domain, A82V (115, 118), later shown to carry increased in-vitro infectivity potentially associated with increased membrane fusion (119). Thirdly, genomic epidemiology studies also revealed a previously unknown role of sexual transmission in the spread of EBOV (115, 120-124).

Most importantly, this was the first instance genomic epidemiology was used in real-time to identify transmission chains and inform public health strategies for epidemic control (115). Phylogenetic analyses were used to uncover virus spread at different geographical and temporal scales, revealing frequent transmission events within and between countries (125, 126), as well as providing important information about transmission networks and virus persistence at localized scales (127). Genomic epidemiology approaches were also used to estimate EBOV Makona's R_0, ranging between 1.65-2.18, and its potential for epidemic spread (128). Combined, these contributions were essential to reveal gaps in epidemiological surveillance, track virus spread at regional and local scales, identify transmission hotspots and design more effective infectious disease control measures (115).

The EBOV Makona epidemic also revolutionized the use of portable genome sequencing for real-time generation of pathogen genome sequences in the field (126). In a seminal paper by Quick and colleagues at the time describing the use of the recently

developed Oxford Nanopore MinION portable sequencer, together with necessary reagents, were transported in a suitcase from the United Kingdom to European Mobile Laboratory located at the Donka Hospital in Conakry, Guinea. A total of 142 EBOV genomes were generated on-site between March and October 2015, aiding the detection of local transmission chains in the country (126). This approach would be later on deployed during the following public health emergency of international concern (PHEIC) caused by the Zika virus (71, 129), and subsequent epidemics caused by Chikungunya virus (16, 130, 131), dengue virus (DENV) (132) and yellow fever virus (YFV) (75) in Brazil. The ZiBRA project (Zika in Brazil Real Time Analysis) used portable genome sequencing in Brazil for the first time during the Zika virus PHEIC in 2015-2016, when a minivan was converted into a mobile laboratory and researchers travelled between 6 public health laboratories in Northeast Brazil tested over 1200 samples and sequencing the first large dataset of ZIKV genomes (133). The study revealed that the ZIKV Asian genotype was introduced from French Polynesia to Brazil 18 months before its first detection in the country (71). More recently, portable genome sequencing was used to confirm that human YFV cases in Southeast Brazil were being sustained by frequent spill over events from non-human primates to humans mediated by sylvatic mosquito vectors in forested areas, rather than via human-to-human transmission (75). And finally, portable sequencing was also used to identify the replacement of the CHIKV Asian by the CHIKV East-Central-South-African genotype in North Brazil (16) and the introduction of a novel DENV serotype 2 genotype associated with a large epidemic in 2019 in Southeast Brazil (132). Building such human and technological capacity on genome sequencing, especially portable genome sequencing, would prove crucial for the early and rapid response to SARS-CoV-2 in Brazil.

1.3 SARS-CoV-2 Evolutionary Origins and Global Spread

1.3.1 Coronavirus Family

Coronaviruses are a large family of viruses which circulate mostly in mammals (e.g. bats, camels, humans, etc) and birds. They are causative agents of gastrointestinal infections in animals and respiratory illness in humans. They are part of the realm *Riboviria*, order *Nidovirales*, suborder *Cornidovirineae* and family *Coronaviridae* which includes 2 subfamilies (*Coronavirinae* and *Torovirinae)*, 5 genera, 27 subgenera and 39 species (134). Viruses from the *Coronaviridae* family are the largest RNA viruses described to date, with polyadenylated and capped genomes ranging between 25-32 kb. They are enveloped, single-stranded, positive-sense RNA viruses and their virions are spherical with a large glycoprotein, Spike (S), which extends from the envelope, resulting in its crown-like shape (135). Nucleocapsids can be flexible (*Coronavirinae*) or helical and doughnut shaped (*Torovirinae*). Viruses from the *Coronavirinae* subfamily are categorised into 4 genera: *alphacoronaviruses*, *betacoronaviruses*, *gamacoronaviruses,* and *deltacoronaviruses (134, 135)*. While *alpha* and *betacoronaviruses* infect mammals, *gamma* and *delta* coronaviruses mostly infect birds, but also some mammals. Humans are primarily infected by *alpha* and *betacoronaviruses* (136, 137).

To date, seven human coronaviruses (HCoVs) have been described (138). Commonly acquired human coronaviruses (HCoV-229E, HCoV-NL63, HCoV-OC43, HCoV-HKU1) are endemic and mostly cause mild respiratory disease (137-139). However, 3 *betacoronaviruses* are highly pathogenic, causing severe respiratory infections in humans and having been linked to severe global outbreaks: the severe acute respiratory syndrome virus (SARS-CoV), the Middle Eastern respiratory syndrome virus (MERS-CoV) and the severe acute respiratory syndrome virus 2 (SARS-CoV-2), the causative agent of the current global pandemic (9, 136, 140).

1.3.2 Highly-pathogenic human coronaviruses

In November and December 2002, the first cases of a new upper respiratory tract infection rapidly evolving to severe pneumonia and potentially death were reported in province of Guangdong, China (141, 142). The disease was named as severe acute respiratory syndrome (SARS) and its causative agent was later described as the first highly pathogenic human coronavirus (SARS-CoV). To that date, coronaviruses were only known to cause mild respiratory diseases in humans (136). Approximately two thirds of the initial SARS-CoV cases were linked to workers handling live animals/food, suggesting that zoonotic spillover events in wildlife markets could explain the origins of the outbreak in humans (142). However, it is now believed that bats may be the natural hosts of SARS-CoV, while civets (143, 144) would be intermediate hosts between bats and humans (136). In February 2003, a traveller from Guangdong infected multiple people while staying in a hotel in Hong Kong. Some of these individuals were also travellers, which in turn infected multiple people when returning to other countries, including Canada, Vietnam, and Singapore (145). It is estimated that SARS-CoV caused over 8000 SARS cases and 700 deaths in at least 26 countries worldwide, figuring as the first pandemic of the 21st century and the first in history to be caused by a coronavirus (141, 146). SARS-CoV is transmitted by mucosal contact with respiratory droplets or fomites from infected individuals. Given the close contact with SARS patients and the higher risk imposed by aerosol-generating procedures, cases were mostly reported in hospital settings (141, 147), with outbreaks sometimes involving more than 100 individuals (147-149). Cases have also been linked to superspreading events (142, 150).

The first Middle Eastern respiratory syndrome (MERS) case was described in June 2012 in Saudi Arabia in a patient presenting an unusual pneumonia, multiple organ failure and subsequent death (151). As of October 2021, a total of 2578 MERS cases and 888 deaths

have been reported, equating to a case fatality ratio of 34.4%. Cases have been reported in 27 countries worldwide, with 2178 (84.5%) being reported in Saudi Arabia (152, 153). Individuals aged 50-59 are at the highest risk for primary infection (presumably camel-to-human), while individuals aged 30-39 are at highest risk for secondary infection (human-to-human transmission) (152, 153). Mortality is higher with increased age (153). Information from some case reports, serological surveys and genome sequencing studies suggest frequent spillover through direct contact between humans and camels (154-158). Recent phylodynamic analysis of MERS-CoV genomes from humans and camels has revealed camels to be MERS-CoV main reservoirs and responsible for the MERS-CoV long-term evolution, while humans are only transient and final hosts, with large clusters of transmission happening in specific environments, such as household contact and healthcare setting (74). These results confirm previous hypotheses that outbreaks in humans are dependent from contact with and spillover events from camels, and that MERS-CoV is unlikely to become endemic in the human population, as human-to-human transmission is not sustainable and clusters are more likely to die out (74). Most affected countries have implemented important and strong control measures which have resulted both in a massive decrease in the number of reported cases since the largest 2014-2015 MERS-CoV outbreak and in a reduction of its global spread (159).

These SARS-CoV and the MERS-CoV international epidemics have encouraged the scientific international community to investigate highly pathogenic coronaviruses and for countries to learn from their previous experiences and prepare for a possible future coronavirus pandemic (160, 161). Since then, several new coronaviruses have been described in bats and in other animals, suggesting that spillover of new coronaviruses to human populations are very likely to occur in the future. Fortunately, advances were also made regarding our understanding of the biology and control of highly pathogenic

coronaviruses, including novel vaccines and treatments (162). These advances would be essential for the response to first coronavirus pandemic in history.

1.3.3 Timeline of SARS-CoV-2 Early Cases

SARS-CoV-2 is a novel betacoronavirus and it is the aetiological agent of the coronavirus disease 2019 (COVID-19) (163). The first cases of COVID-19 were reported on the 31st of December 2019 by the WHO China Country Office, initially as a cluster of cases of a severe pneumonia of unknown aetiology in Wuhan, Hubei province (164). Cases were initially linked to the Huanan Seafood Market, where most of the patients worked or had recently visited (165, 166). By the 4th of January 2020, a total of 44 patients had been reported, mostly presenting fever. However, 11 patients presented severe symptoms such as difficulty in breathing and invasive lung lesions (164). The next few days were followed by several measures and initial communications by the WHO on the novel outbreak, including guidelines on how other countries could detect and handle new cases (167). On the 7th of January 2020, Chinese scientists reported having identified a new beta coronavirus from bronchoalveolar lavage fluid collected on the 26 of December 2019? from a patient residing in Wuhan. This new virus, suspected to be the causative agent of the Wuhan new pneumonia cases (9, 165, 166), was then named as 2019-nCoV. On the 11th of January 2020, China reported the first death associated to the new pneumonia from a 61-year old man who was a regular customer at the Huanan Market in Wuhan. Chinese scientists shared the first genomic sequence of the new coronavirus on the 10th of January 2020 (168). Early sharing of the virus complete genome was considered a turning point in the management of the outbreak as it allowed the rapid development of diagnostic tests, vaccines and treatments. Soon after, reports of clusters amongst family members and within-hospital transmission

suggested that human-to-human transmission was the main transmission mode of the novel 2019-nCoV (169-171).

While the cases increased in Wuhan and 2019-nCoV spread across other Chinese provinces, on the 13th of January 2020, Thailand officials reported a returning traveller from Wuhan to be the first COVID-19 case outside of China (172). This report was followed by other imported cases in Japan and in North Korea on the 15th and 20th of January 2020, respectively. On the 21st of January 2020, the first COVID-19 case outside of the Asia was confirmed by the Center for Disease Control and Prevention, CDC in a Washington State resident who had returned from Wuhan on the 15th of January 2020 (173).

To prevent further national and international spread of COVID-19 cases, on the 23rd January 2020, the Chinese government isolated the city of Wuhan by prohibiting any movement in or out of the city, in addition to raising the national public health emergency response to level 4, its highest. A day later, Wuhan restrictions were expanded to the Province of Hubei (174). However, by the 30th of January, a total of 7818 cases had been reported globally, 7,736 of them within China and 82 cases in 18 countries across 4 different continents. The exponential growth of the 2019-nCoV Chinese epidemic and the rapid global spread of 2019-nCoV outside of China led the WHO Emergency Committee to declare the COVID-19 epidemic a Public Health Emergency of International Concern (PHEIC) on the 30th January 2020 (167, 175).

The months of February and March 2020 were marked by the first deaths outside China, worldwide spread and by a change of the novel PHEIC epicenter. On the 2nd of February 2020 (176), the first COVID-19 death outside China was reported in the Philippines, followed by a death in Japan on the 14th of February 2020 (177). The first death outside of Asia was reported in France on the 16th of February 2020 (178). By the 11th of March 2020, 118,319 cases and 4,292 deaths had been reported across 113

countries/areas/territories worldwide and COVID-19 was officially declared a pandemic by the WHO (179). Two days later, Europe was declared the new epicenter of the COVID-19 epidemic, with Italy and Spain being the worst hit countries (180). Given the upsurge of cases and deaths, the Italian government enforced strict containment measures, including social distancing and severe movement restrictions, to a national level, which would only be removed 2 months later, on the 4th of May 2020 (181). By the lifting of restrictions, Italy had recorded 210,000 cases and almost 29,000 deaths, much higher numbers than China, the initial epicenter of the pandemic, that reported 84,400 cases and 4,643 deaths during the same period (182). Spain, the second most affected European country, at the time, reached its peak reporting of daily new cases on 1st April 2020, with a total of 8,195 cases (10).

By then, Europe had become the new global epicentre of the pandemic, with the United States (US) reporting an increasing number of COVID-19 cases in New York city (183). By the end of March 2020, the US recorded some of the highest daily number of cases worldwide and accounted for 1 in each 5 cases globally (184). This would be followed by a largely fragmented response to the pandemic with politically motivated public health decisions being made at the state level (185).

The first COVID-19 case in Latin America was reported on 26th of February 2020 in São Paulo, Brazil (186) followed by reports from Mexico on the 28th of February 2020 (187) and Ecuador on the following day (188). By the 10th of March 2020, other 10 Latin American countries/territories had reported at least one COVID-19 case (189). The first COVID-19 death in Latin America was reported on the 7th of March in Argentina (190). The global shortage of testing and personal protective equipment (PPE), lack of preparedness and uncoordinated responses, in addition to the decades of failing socio-political systems and struggles with high indices of inequality, low access to health services and poverty led Latin America into a turmoil (191-194). Only 4% of the Latin American demand for medical

supplies required to reduce exposure and mortality, including PPE, ventilators and test kits, were met by local production, making the region highly vulnerable to shortages and dependent on the production from foreign nations, which were already suffering with their own internal demands and implementing export restrictions (195, 196). Moreover, Latin America had recently gone through major epidemics caused by DENV, CHIKV, ZIKV, and YFV which strained the local public health systems and population resilience (197). Amidst reports of people dying on the streets in Ecuador (198) and mass grave yards in Manaus, Brazil, Latin America became the new epicenter of pandemic on the 22nd of May 2020 (199). By then, Brazil was the second worst-hit country by the SARS-CoV-2 pandemic in the world and the epicentre of the epidemic in Latin America with 291,579 cases and 18,859 deaths (200). However, due to the scarcity of available testing supplies, underreporting of SARS-CoV-2 cases is high in Latin America and the true number of cases are likely much larger than the reported numbers (201, 202).

1.3.5 SARS-CoV-2 Variants

Several processes impact and shape the evolution of viral genomes such as polymerase error during replication, host enzymes, spontaneous chemical reactions and mutagen agents found in the environment (203). Although exceptions to the rule exist, RNA viruses tend to have higher mutation rates than DNA viruses. This faster evolutionary rate can be partially explained by a less effective proof-reading replication system which uses low fidelity RNA polymerases. While most RNA viruses have mutation rates at the scale of 10^{-3} substitution/per site/year, variations within the group also exist and evolutionary rates can range between 10^{-2} and 10^{-5} substitution/per site/year (203). For instance, reverse transcriptases (RT) used for replication by retroviruses such as the human immunodeficiency virus (HIV) have much higher fidelity than that of RNA-dependent RNA

polymerases (RdRp) from other RNA viruses, and thus lead to lower mutation rates (203, 204). However, DNA viruses lacking exonuclease proofreading have similar fidelity rates to RdRp RNA viruses, suggesting that the lack of proofreading mechanisms is likely the main driver for the more error-prone replication process of RNA viruses, rather than lower polymerase fidelity by itself (205). The evolution of RNA viruses is likely constrained at a range to allow for the emergency of enough diversity and fast adaption, but also to ensure genome stability as most random mutations are deleterious and could result in lower fitness (205-208). Controlling of replication fidelity has been used as a mechanism to produce new RNA virus vaccines (209).

In fact, large RNA viruses belonging to the *Nidovirales* order and *Coronaviridae* and *Roniviridae* families with genomes >20,000 base pairs (bp) such as coronaviruses have been shown to harbour proofreading systems to improve polymerase fidelity and reduce mutation rates (210). These mechanisms would be essential to maintain the integrity of such large genomes, which are much larger than the average 10,000 bp of most RNA viruses (211, 212). Examples of such mechanisms would be the presence of 3′→5′ exoribonuclease activity associated to the Non-structural protein 14 (nsp14-ExoN) and endoribonuclease activity of nsp15 of coronaviruses (213). nsp14-ExoN can proofread single 3' mismatched nucleotides (214, 215) and abolishment of its activity reduces coronavirus replication fidelity by up to 20-fold (216). Such proofreading mechanisms have led coronaviruses to present evolutionary rates which can be 10-fold lower compared to seasonal influenza A virus (217). However, nsp14 has also been shown to be required for coronavirus recombination (218).

SARS-CoV-2 has an estimated evolutionary rate ranging from 0.8 to 2 x 10^{-3} substitutions per site per year (s/s/y) (219-221). This equates to approximately 1 mutation every two weeks or 2-3 mutations a month. Given this slow evolutionary rate, classification

of new lineages and their association to specific geographic locations have been challenging tasks as lineages often differ by a single nucleotide change (217). Early studies analysing the population structure of SARS-CoV-2 phylogenies identified two large SARS-CoV-2 lineages, A and B, which co-circulated since the beginning of the outbreak in Wuhan, suggesting that at least two multiple independent spillover events may have happened. Lineage B sequences differ at two nucleotide positions (8,782 in ORF1ab and 28,144 in ORF8) from the closest sequences of known bat viruses (RaTG13 and RmYN02), while lineage A viruses have these positions conserved, suggesting lineage A to be the closest common ancestor to the human circulating SARS-CoV-2 strains (217, 222, 223). The first SARS-CoV-2 publicly available genomes belonged to SARS-CoV-2 lineage B and were linked to exposure to the Huanan seafood market while the earliest SARS-CoV-2 lineage A sequences have been linked to other wet markets, and to unrelated cases in Wuhan and other parts of China (9, 166, 224, 225).

Although both SARS-CoV-2 lineages A and B spread globally, lineage B soon became dominant worldwide. This early dominance of B lineages has been associated with the acquisition of one amino acid substitution in the spike protein, D614G, which was passed down to subsequent B lineages. The fact that 614G strains were associated to potentially higher virus loads and infectious titers raised the question of whether D614G had increased SARS-CoV-2 fitness and transmissibility (226). Later studies showed that D614G increases the stability and infectivity of SARS-CoV-2 virions, thus enhancing its capacity of replicating in human airway cells and tissues, and leading to higher infectious titers in the trachea of infected hamsters (227-229). Recent animal studies have also shown 614G to increase transmission in 20% compared to wild-type and thus representing a competitive advantage in the transmission process (230, 231). Phylodynamic analysis using 25,000 SARS-CoV-2 sequences from the United Kingdom concluded that the overtake of 614G

variants is consistent with a selective advantage and that 614G is associated to lower Ct levels and younger age, but not with increased disease severity (232). D614G figures as the first observation of an evolutionary selectively adapted mutation in SARS-CoV-2 pandemic genomes.

Since then, other mutations in the Spike protein have been associated to the emergence of lineages of potential public health interest. Such lineages are currently separated into three different categories according to their epidemiological characteristics: variant of concern (VOC), variant of interest (VOI) and variant under monitoring (VUM) (233). A VOI is a new SARS-CoV-2 lineage presenting (1) mutations predicted or proven to alter epidemiologically relevant virus characteristics (e.g., transmissibility, severity, diagnostic or escape to therapy), with (2) preliminary data suggesting to increasing prevalence across multiple populations (233, 234). A VOC presents the same characteristics of a VOI with the addition that one or more altered characteristics have been confirmed through comparative studies to happen at the level of global public health significance: increased transmissibility, or increased virulence or altered disease presentation, or decreased effectiveness of public health measures, diagnostics, therapeutics or vaccines. In turn, VUMs are variants harbouring mutations which might potentially pose a public health risk in the future, but lack current evidence of its impact (233). SARS-CoV-2 VOC, VOI and VUMs are currently named according to a WHO-defined Greek alphabet nomenclature system considering the order in which these new variants have been first identified. This new nomenclature was developed to provide easy-to-pronounce and non-stigmatising variant names for the general public, and complement three other nomenclatures frequently used by the scientific community (Pango lineage, GISAID clades, NextStrain clades) (222, 233, 235).

Currently, the WHO recognizes 5 VOCs: Alpha (B.1.1.7, GRY, 20I/V1), Beta (B.1.351, GH/501Y.V2, 20H/V2, Gamma (P.1, GR/501Y.V3, 20J/V3), Delta (B.1.617.2, G/478K.V1/ 21A, 21I, 21J), and Omicron (B.1.1.529/ GRA/ 21K, 21L, 21M). The first variant of concern, Alpha, was described on the 18th of December 2020 by the COVID-19 Genomics UK Consortium (COG-UK) (236) and designated as a VOC by Public Health England (PHE). Retrospective investigations were able to trace the earliest Alpha cases back to the 20th of September 2020 in samples from Kent, south-eastern England, and Greater London areas. By December 2020, Alpha had spread across different regions of the UK and was rapidly increasing in frequency (236), even during the UK's second lockdown, suggesting increased selective advantage and transmissibility fitness. Considering SARS-CoV-2's relatively slow evolutionary rate, Alpha called attention for harbouring an array of lineage defining mutations: 14 non-synonymous mutations, 3 nucleotide deletions and 6 synonymous mutations (236). It has been hypothesized that the fast accumulation of lineage-defining mutations may be associated to chronically-infected SARS-CoV-2 patients who are immunosuppressed or immunodeficient. Such patients would undergo extensive therapy with convalescent sera, which can increase selective pressures over specific variants through direct selection and genetic hitchhiking (236-239). Most of Alpha's non-synonymous mutations and deletions, 47% (8/17), were present in the Spike protein. The most important of these mutations is N501Y, which lies in the receptor binding domain (RBD) and has been shown to increase the binding affinity of SARS-CoV-2 spike protein to hACE2 receptors, potentially affecting virus transmissibility and neutralization (240). Studies have estimated Alpha to be 50-100% more transmissible compared to wild-type and the hazard of death associated to an infection with an alpha strain virus to be 61% (42-82%) higher than that of previously circulating lineages (241, 242). Although reduced neutralization of Alpha has been shown, no evidence for significant impact in immune protection, naturally or vaccine-

mediated, has been found (234, 240). Given its increased transmissibility, Alpha swiftly spread outside the UK and became the main circulating lineage, especially in Europe, North America and Middle East, causing large second/third waves of cases and deaths (243, 244).

Shortly after the description of Alpha by the UK government, 2 new VOCs were identified in South Africa and in Brazil, both presenting a constellation of lineage-defining mutations, including N501Y. Beta was also first described in December 2020. It was first detected in samples from 8th October 2020 and followed by a rapid second wave in South Africa characterized by an exponential increase in the number of cases and deaths, which would later on surpass the burden of South Africa's COVID-19 first wave (245). Its emergence was dated back to early August 2020 in Nelson Mandela Bay (245). Beta's Spike protein harbours 9 amino acid changes: 8 non-synonymous mutations and 1 deletion. Three of the non-synonymous mutations are located in the RBD region of spike protein (N501Y, E484K and K417N). Mutations N501Y and E484K have been shown to increase spike's binding affinity to hACE2 receptors, individually and further when combined (246). Epidemiological studies have shown that Beta is 1.50 times (95% CrI: 1.20-2.13) more transmissible and is also associated with increased severity compared to previously circulating lineages (247, 248). However, differently from Alpha, Beta's mutation array, especially E484K, has been associated with significant evasion of both natural and vaccine-mediated immunity (249, 250).

The Gamma VOC was first reported on the 10th of January 2021 after its detection in samples from Japanese travellers returning from Manaus, Brazil (251). This initial report was followed by two preliminary studies pointing to the circulation of Gamma in Manaus and to its exponential increase in just a few weeks after its first detection (252, 253). Due to Gamma's emergence, Manaus experienced a second wave of an unprecedent magnitude of cases and deaths (254). Similarly to Alpha and Beta, Gamma also acquired a constellation

of lineage-defining mutations compared to its precursor, Pango lineage B.1.1.28: 15 non-synonymous mutations, 1 insertion, 1 deletion, and 6 synonymous mutations. Gamma's spike protein harbours 10 amino acid changes, from which 3 are located in the RBD: N501Y, E484K and K417T. As previously mentioned, mutations in these sites are associated to increased transmissibility, severity and immune evasion (252). Gamma has since been shown to present increased transmissibility, to evade partially evade natural and vaccine-mediated immune response and to lead to more severe disease (247, 255, 256). Since its emergence in Manaus, Gamma has spread to all other Brazilian states, leading to daily records of cases and deaths of 115,228 and 4,249, respectively (10). Gamma has also spread to several other countries, becoming the dominant circulating lineage in the region in South America (244). The early detection and transmission of Gamma is described in more detail in Chapter 3 of this thesis.

In March 2021, India saw a swift increase in the number of COVID-19 cases leading to a hard-hitting second wave in the following weeks. Initially thought to be caused by multiple newly circulating lineages, including Alpha introduced from the UK, the new lineage Delta, first detected in Maharashtra, rapidly surpassed the number of cases caused by other circulating lineages (257). This observation suggested that Delta was more transmissible than other lineages, including other VOCs. Delta is characterised by spike mutations T19R, Δ157-158, L452R, T478K, D614G, P681R, and D950N. Some of these mutations have been shown to impact virus replication and/or transmission, such as P681R which is located at the S1-S2 spike subunits cleave site and lead to increased replication and transmission (258, 259). Delta has been estimated to be 40-60% more transmissible than Alpha (260), and hospitalization is twice as likely for individuals infected with Delta when compared to Alpha (261). Although reduced, effectiveness of two doses of SARS-CoV-2 vaccines against Delta is almost the same as that against Alpha. However, immunity

mediated by one dose-only is dramatically reduced from 48.7% to 30.7% (Alpha vs Delta) (262). These findings led many governments to reduce the interval between the first and second dose of COVID-19 from 12 to 8 weeks with the aim of slowing the spread of Delta and avoiding the increase in hospitalisations and deaths. Given these characteristics, delta rapidly spread globally and became the dominant lineage even in countries where other VOCs were already established (263-265). As of 4th January 2022, Delta has spread to at least 180 countries worldwide (265).

Omicron was first reported by a South African scientist on the 24th November 2021 and classified as a VOC by the WHO only 2 days later (266). The new variant drew attention for its rapid spread after its first detection in Tshwane, Gauteng Province, South Africa on 9 November 2021, and for the extensive set of newly acquired mutations, especially in the Spike protein (267). Omicron presents 45-52 amino acid across its genome, 26-32 of which are located in the Spike protein, including the Δ69-70 deletion, which leads to failure of some PCR tests targeting the Spike gene (S-gene target failure). Several Spike mutations can be found in immunogenic regions such as the RBD and the N-terminal Domain (NTD), as well as mutations around the furin cleavage site (267). This array of new mutations has been shown to decrease vaccine-mediated efficacy to approximately 30% against symptomatic disease, resulting in breakthrough infections (268, 269). However, vaccine-mediated immunity seems to be restored by a booster dose to over 70% (269). There is currently a large debate on the severity of Omicron cases as early data seems to point to lower risk for hospitalisation and severe disease (270, 271). However, Omicron transmits much faster than previous variants, potentially associated with a combination of its immune evasion capacity and of an increased ability to preferably infect the upper rather than the lower respiratory tract (272, 273). Omicron's global spread has been met with the highest daily number of cases in many countries and also globally (10). Such combination of

transmissibility and disease severity is likely to lead to a significant number of deaths and straining of public health systems. As of 4th January 2021, Omicron has been detected in at least 91 countries worldwide (265).

1.4 Brazil in the Context of Infectious Diseases

With an estimated total population of over 213 million people and a territory of over 8.5 million km^2 (274), Brazil is the largest country in Latin America and the 5th largest in the world both in population and in territory. It has the 12th largest economy in the world, the 4th amongst developing nations (275). It has a Human Development Index (HDI) of 0.765, 84th in the world. It is geopolitically organized into 5 regions (North, Northeast, Center-West, South and Southeast), 26 federal states and 1 federal district, the country's capital Brasília. Brazilian states are further subdivided into a total of 5.568 municipalities (equivalent to towns or cities) (274). With a national Gini index of 53.4 (276), Brazil is one of the countries with the most unequal income distribution worldwide. Brazilian regions, states and municipalities are highly unequal in terms of population size, economy, education, poverty, access to health care and other socioeconomic indicators (277, 278). Human development index for Brazilian States in 2017 varied between 0.683 and 0.850, but generally following a North/South divide.

The Southeast and South are Brazil's most developed regions (278). Southeast Brazil is home to the 3 largest metropolitan areas in the country (São Paulo, Rio de Janeiro and Belo Horizonte), and 4 of the country's 5 busiest airports, making it the most connected region nationally and internationally (279). In 2018, São Paulo state had the highest HDI in the country, 0.783, and its GDP accounted for 32% of Brazil's GDP, making it the wealthiest unity in the Federation (280). The São Paulo Metropolitan Area is one the largest conurbations in the world with an estimate population of approximately 22 million people

and a population density of 2,714,45 hab./km² (274). It is the largest economic centre in the country, home to the largest hospital complex in Latin America (281), the Hospital das Clínicas at the Faculty of Medicine from the University of Sao Paulo (HC-FMUSP), and two of the busiest airports in Brazil (Guarulhos and Congonhas), making it the main human mobility hub in the country and in Latin America (279).

On the other hand, the Brazilian North and Northeast regions are the poorest regions in the country with comparably some of the worst socioeconomic indicators, especially the North, where the Brazilian Amazon can be found. Amazonas is territorially the largest state is Brazil accounting for 18.3% of its territory (274). It is home to the largest indigenous population in Brazil, to most of the Amazon tropical rainforest and to the largest river in the world, Amazonas (274, 282). Human mobility patterns in the Amazon are complex and mostly driven by fluvial movement and air travel (282). The Amazonas state has an estimated population of approximately 4.3, 52.8% of which resides in its capital, Manaus, the largest urban centre in North Brazil and the 7th in the county (274). Most of its economy is driven by the creation of the Manaus Free Zone in 1957 "with the objective of creating in the Amazon Region an industrial, commercial and agricultural centre under economic conditions that allow its development" (283, 284). For this reason, Manaus has the only high HDI, 0.737, in a state dominated by municipalities with medium and low HDI (285).

1.4.1 The Brazilian Public Health System

According to article 196 of the Brazilian Constitution in 1988, "Health is a right of all and a duty of the State and shall be guaranteed by means of social and economic policies (…), universal and equal access to actions and services for [health] promotion, protection and recovery." This same constitution introduced great public health reforms and formally created the *Sistema Único de Saúde* (SUS– Brazilian Unified Health System), the largest of

its kind in the world (286). The SUS has 3 major principles: universal, equal and integral access to healthcare. Some of the SUS's main public health achievements have been the creation of a *Programa Nacional de Imunização* (PNI–National immunization Programme), which provides universal and free access to vaccines to the Brazilian population since 1973, and the creation of the *Estratégia Saúde da Família* (ESF–Family Health Strategy). Together with the restructuring of the health system and its universality, such programs have led to powerful achievements such as the reduction of the burden of communicable diseases in the country (287) and the reduction of child mortality, with a shift increase in the proportion of deaths towards older ages and non-communicable diseases (288). However, SUS has recently faced major challenges amid the political crisis and democratic instabilities, and an increasingly strong focus on privatization of healthcare.

The HIV/AIDS pandemic was the first major infectious disease threat requiring an urgent response by SUS, and the Brazilian National AIDS programme is internationally recognized for its successful outcomes in a developing country (289, 290). A report from the WHO in 2004 estimated HIV rates of infection to be much lower than what had been projected for the country, with a reduction of 50% in deaths due to HIV infection and reduction of 70% in duration of hospitalization (291). Such accomplishments were made possible by funding from the Brazilian Federal government and the WHO, and a response rooted in the use of scientific knowledge to launch an integrated response aiming to increase disease awareness, prevention, diagnosis and treatment, with free access to antiretrovirals starting from 1996. The Brazilian response also included major campaigns focused on the use of condoms and the de-stigmatization of HIV/AIDS (290) and the creation of a nationwide effort for genome sequencing of the HIV-1 strains circulating in the country, the Brazilian Network for HIV-1 Isolation and Characterization (292).

1.4.2 Brazil's Genome Sequencing Capacity

To contextualise the capacity of viral genome sequencing in Brazil, I performed an analysis on data available on NCBI GenBank. As SARS-CoV-2 genomes are mostly deposited on GISAID, SARS-CoV-2 metadata is not included in this analysis. As of 5th December 2021, there were 56,074 virus genetic sequence entries available on GenBank collected in Brazil (Figure 1A). Most sequences are viral genetic fragments shorter than 5,000 bp (96.4%), with a median length of 903 bp (range 53 bp to 1.5Mb). Only 2.6% of viral sequences from Brazil have a length above 8Kb. Encouragingly, there has been a substantial increase in the number of virus genetic sequences with length above 8Kb from Brazil in the last decade, with most data (72.6%) being submitted from 2016 onwards (Figure 1A). The median turnaround times between sample collection and GenBank submission for Brazilian genomes is 3.5 years (range 0.05-89.4 years) when considering all entries (Figure 1B). This median is reduced to 2.7 years (range 0.05-9.98 years) when time gaps >10 years are excluded to decrease potential biases introduced by sequencing of historical samples. However, there has been a considerable reduction of such time gaps across the years and since 2017 the median time between collection and submission has ranged between 1.3 and 2.6 years. The increase in median length and decrease in turnaround times coincides with the CHIKV, ZIKV and YFV epidemics and DENV outbreaks in the country (Figure 1C) and the first use of portable genome sequencing technologies in Brazil.

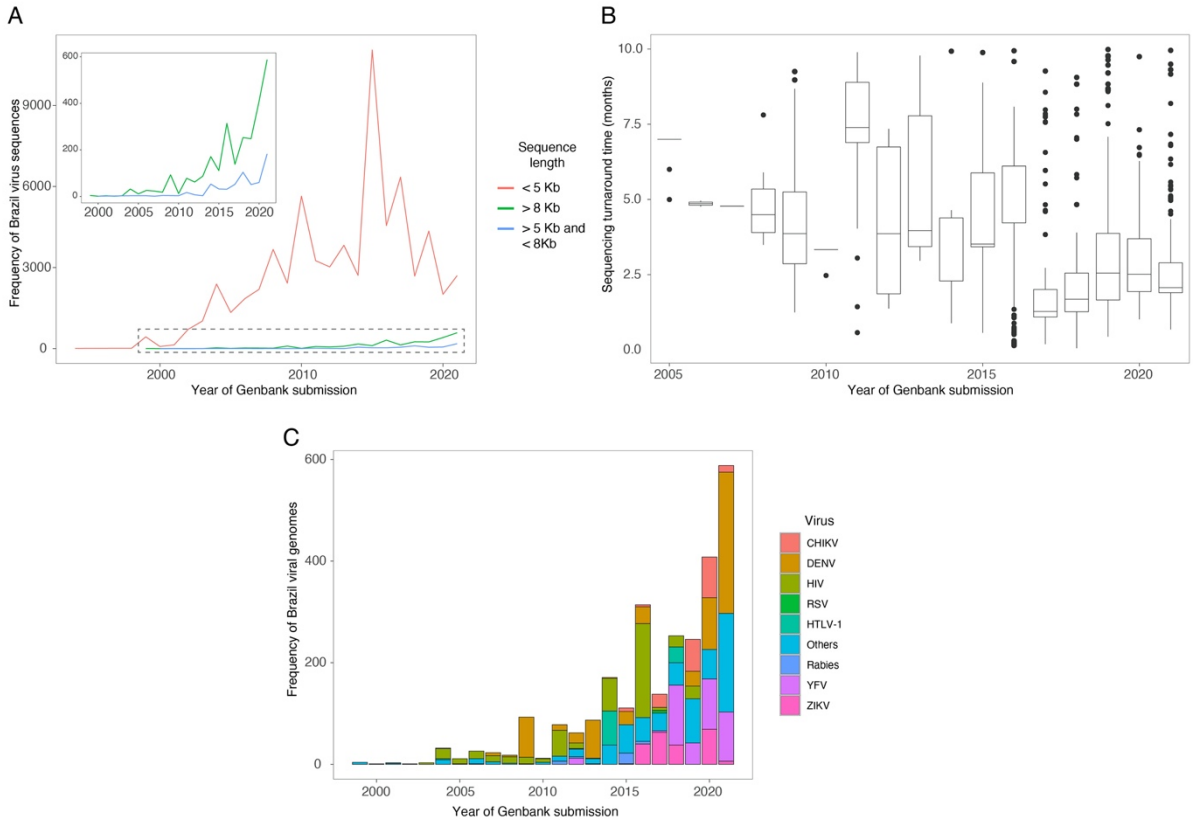


Figure 1.1. Overview of Brazil viral genome submissions on GenBank across time. All metadata for viral genetic sequence entries from Brazil (n=56,074), regardless of their size, were downloaded from NCBI GenBank using rentrez and gbfetch R packages. (A) Timeseries plot of all Brazil viral genetic sequences in NCBI GenBank stratified by sequence length: <5Kb (red), >5Kb and <8Kb (blue), and >8Kb (green). Sequence length groups were determined based on a bimodal distribution of sequence lengths identified on histogram plot analysis, with most submissions falling into <5Kb and >8Kb. Inset shows an expanded view of the area around the dotted line: groups >5Kb and <8Kb (blue), and >8Kb (green). (B) Time between date of sample collection and date of NCBI GenBank submission. As historical sequences have commonly been submitted, time gap has been restricted to up to 10 years in plot B. (C) Time series of viral genome submission from Brazil with length >8Kb (segmented viruses not included) coloured according to organism: chikungunya virus (CHIKV), dengue virus (DENV), human immunodeficiency virus (HIV), respiratory syncytial virus (RSV), human T-lymphotropic virus type 1 (HTLV-1), rabies lyssavirus (Rabies), yellow fever virus (YFV), zika virus (ZIKV).

Out of all virus submissions to GenBank, arboviruses are the largest group of viruses with near-complete/complete genomes from Brazil (Figure 2C). Figure 3 gives a context of sequencing capacity in Brazil when compared to the other top 20 countries in terms of arbovirus sequencing. Since 2015 Brazil has been facing a triple epidemic of DENV, ZIKV and CHIKV with occasional outbreaks of YFV (293). It is estimated that Brazil accounts for

55% of all DENV cases reported in the Americas in the period between 1995-2015 (294), and is also the country reporting the highest number of DENV cases worldwide (295). However, Brazil is only the 6th country in terms of DENV sequence submissions to GenBank and the 4th when considering only sequences >8Kb, with irregular submission patterns. For CHIKV, although Brazil figures as the third country with regards to the number of total viral sequences deposited in GenBank, Brazil leads in the total number of available genomes with length >8Kb. Finally, Brazil also leads when it comes to ZIKV and YFV genomes. However, even though Brazil might be doing better when compared to other developing nations experiencing similar epidemics in absolute terms, the median number of sequences with length >8Kb submitted per year is still very low for a country reporting millions of arbovirus cases a year (296, 297). In addition, turnaround times of around two years is remarkably slow specially when considering recent improvements in genomic capacity during the SARS-CoV-2 pandemic, with turnaround times of only 48h reported by our team (298).

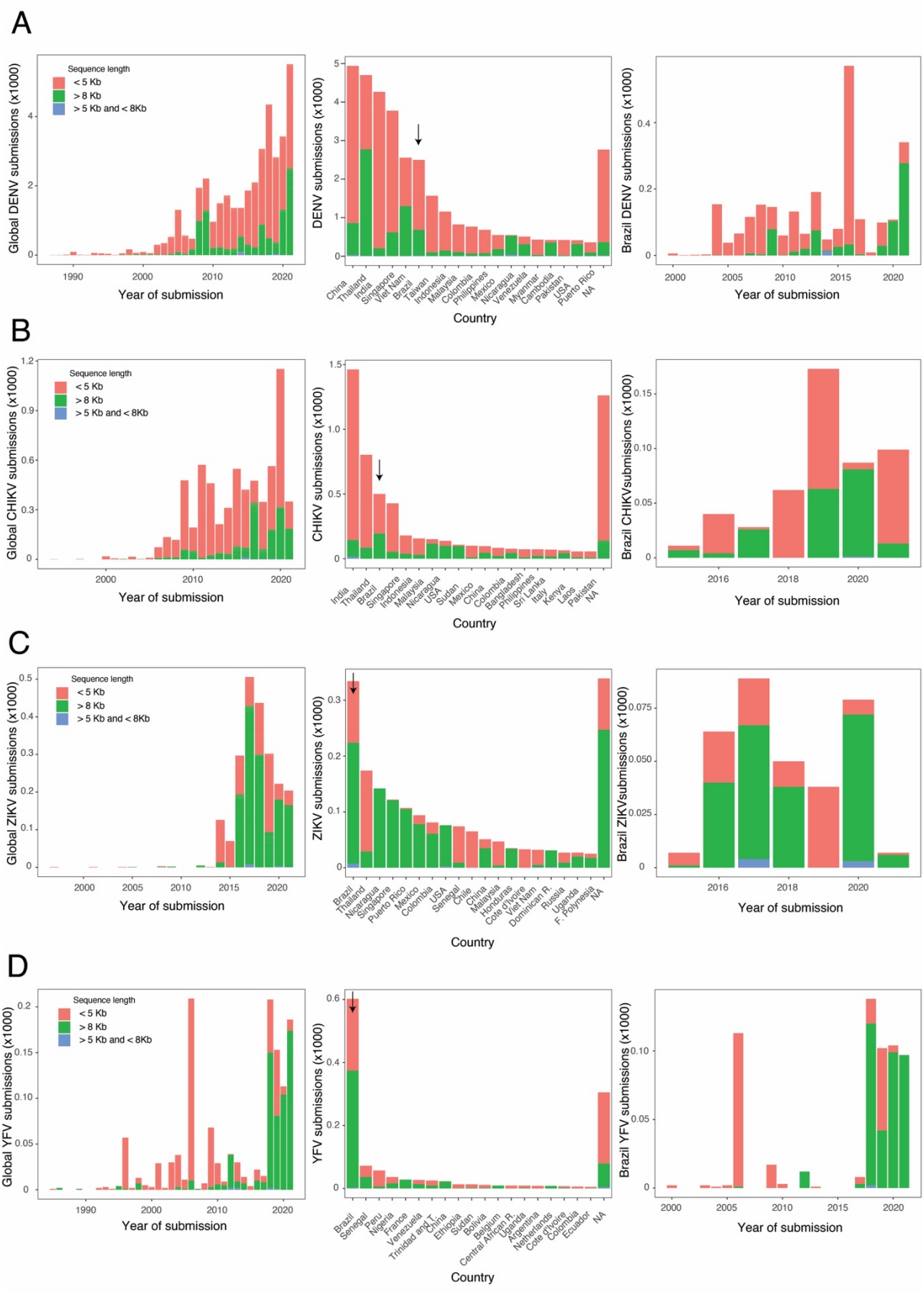


Figure 1.2. Overview of GenBank submissions for the main circulating arboviruses in Brazil. Each panel presents data for a different arbovirus, as follows, (A) DENV, (B) CHIKV, (C) ZIKV and (D) YFV. Panels include one timeseries of all global sequences

submitted to GenBank, top 20 countries in absolute number of submissions and timeseries of genomes submitted from Brazil. Entries have been stratified according to sequence length: <5Kb (red), >5Kb and <8Kb (blue), and >8Kb (green). Arrows on sequence distribution per country plots highlight Brazilian data.

1.5 Thesis Outline

In this thesis, I explore the different applications of genomic epidemiology to guide our understanding during the first year of the SARS-CoV-2 pandemic in Brazil. I describe the SARS-CoV-2 introduction in the country, its initial nationwide spread, identify and describe a new SARS-CoV-2 VOC and investigate the within-hospital transmission in the largest hospital complex in Latin America. This thesis travels through different timings of analyses (real-time, near real-time and retrospective) and different geographic scales (nationwide, state, city and hospital complex). It also draws on traditional epidemiology analysis and multiple sources of data to maximise our understanding of SARS-CoV-2 spread in Brazil and the country's response to it.

In **Chapter 2**, I describe the early stages of SARS-CoV-2 in Brazil, including the country's early response to the pandemic. I start by anticipating which air travel journeys would be the main roots for SARS-CoV-2 importation in the country by using flight data and SARS-CoV-2 incidence. Next, I use real-time portable genome sequencing and phylogenetic analysis to show that the first imported SARS-CoV-2 cases in Brazil likely came from Italy. I then use traditional epidemiology approaches to contextualise SARS-CoV-2 initial spread by estimating the reproduction number of the 5 most affected Brazilian states and compare them to the reproduction number of other countries. Finally, I use data on 72% of all Brazilian municipalities to uncover the temporal history of adoption and easing of non-pharmaceutical interventions (NPIs) and reveal the fragmentation of the Brazilian response.

In **Chapter 3**, I analyse 427 newly generated genomes from 18 Brazilian states. I use discrete and continuous trait phylogeographic approaches to reconstruct SARS-CoV-2's introduction into Brazil and its geographical and temporal nationwide spread during its first 2 months of circulation. I also use nationwide human air travel mobility to explore the different stages of SARS-CoV-2 spread in the country, human mobile phone data and case counts to investigate the impact of NPIs in mitigating the SARS-CoV-2 spread at the city level. This is, to date, the largest study investigating the early stages of SARS-CoV-2 spread in Brazil.

Chapter 4 explores the importance of real-time genomic epidemiology and surveillance to track virus evolution and identify newly emergent and circulating lineages. This chapter is composed by my contributions to two published articles. I start by providing the first evidence for the circulation of a new VOC in Manaus, P.1/gamma, and for a faster evolutionary rate leading to its emergence. This investigation was performed in real-time while Manaus was experiencing a critical second wave of COVID-19 cases, despite estimates of high attack rates in the city. I also describe the first two introductions of VOC B.1.1.7/alpha in Brazil.

While I explore large geographical scales in chapters 2, 3 and 4, in **Chapter 5**, I focus on understanding within-hospital SARS-CoV-2 transmission in the largest hospital complex in Latin America. In this chapter, I use traditional epidemiology and phylogenetic analysis to uncover SARS-CoV-2 nosocomial transmission and reveal that SARS-CoV-2 transmission was higher in non-COVID hospitals than in the one COVID-only hospital in the complex, despite universal masking. Such results might be linked to risk perception and reveal important aspects of healthcare worker (HCW) behaviour, which can be very essential for the hospital-management of future outbreaks.

1.6 References

1. Woolhouse M, Gaunt E. Ecological origins of novel human pathogens. Crit Rev Microbiol. 2007;33(4):231-42.
2. Grillet ME, Hernández-Villena JV, Llewellyn MS, Paniz-Mondolfi AE, Tami A, Vincenti-Gonzalez MF, et al. Venezuela's humanitarian crisis, resurgence of vector-borne diseases, and implications for spillover in the region. Lancet Infect Dis. 2019.
3. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol. 2019;4(1):10-9.
4. Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. Nature. 2007;447(7142):279-83.
5. Beyrer C. A pandemic anniversary: 40 years of HIV/AIDS. Lancet. 2021;397(10290):2142-3.
6. Yactayo S, Staples JE, Millot V, Cibrelus L, Ramon-Pardo P. Epidemiology of Chikungunya in the Americas. J Infect Dis. 2016;214(suppl 5):S441-S5.
7. Fauci AS, Morens DM. Zika Virus in the Americas--Yet Another Arbovirus Threat. N Engl J Med. 2016;374(7):601-4.
8. Freitas LP, Cruz OG, Lowe R, Sá Carvalho M. Space-time dynamics of a triple epidemic: dengue, chikungunya and Zika clusters in the city of Rio de Janeiro. Proc Biol Sci. 2019;286(1912):20191867.
9. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33.
10. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20(5):533-4.
11. Bonita R, Beaglehole R, Kjellström T. Basic epidemiology: World Health Organization; 2006.
12. Burrell CJ, Howard CR, Murphy FA. Epidemiology of Viral Infections. Fenner and White's Medical Virology. 2017:185-203.
13. Louten J. Virus Transmission and Epidemiology. Essential Human Virology. 2016:71-92.
14. Rainwater-Lovett K, Rodriguez-Barraquer I, Moss WJ. Viral Epidemiology: Tracking Viruses with Smartphones and Social Media. Viral Pathogenesis. 2016:241-52.
15. Rife BD, Mavian C, Chen X, Ciccozzi M, Salemi M, Min J, et al. Phylodynamic applications in 21. Glob Health Res Policy. 2017;2:13.
16. Naveca FG, Claro I, Giovanetti M, de Jesus JG, Xavier J, Iani FCM, et al. Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon. PLoS Negl Trop Dis. 2019;13(3):e0007065.
17. Tolles J, Luong T. Modeling Epidemics With Compartmental Models. JAMA. 2020;323(24):2515-6.
18. van den Driessche P. Reproduction numbers of infectious disease models. Infect Dis Model. 2017;2(3):288-303.
19. Kraemer MUG, Pybus OG, Fraser C, Cauchemez S, Rambaut A, Cowling BJ. Monitoring key epidemiological parameters of SARS-CoV-2 transmission. Nat Med. 2021;27(11):1854-5.
20. Anderson RM, May RM. Infectious diseases of humans: dynamics and control: Oxford university press; 1992.
21. Herd immunity in the epidemiology and control of COVID-19 The Royal Society: royalsociety.org; 2020 [Available from: https://royalsociety.org/-

/media/policy/projects/set-c/set-c-herd-immunity.pdf?la=en-GB&hash=2B8F6255CF6FD83CE71FC510596FBCFB.
22. Guerra FM, Bolotin S, Lim G, Heffernan J, Deeks SL, Li Y, et al. The basic reproduction number (R. Lancet Infect Dis. 2017;17(12):e420-e8.
23. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the Basic Reproduction Number (R. Emerg Infect Dis. 2019;25(1):1-4.
24. Breban R, Vardavas R, Blower S. Theory versus data: how to calculate R0? PLoS One. 2007;2(3):e282.
25. Dietz K. The estimation of the basic reproduction number for infectious diseases. Stat Methods Med Res. 1993;2(1):23-41.
26. Keeling MJ, Grenfell BT. Individual-based perspectives on R(0). J Theor Biol. 2000;203(1):51-61.
27. Adam D. A guide to R - the pandemic's misunderstood metric. Nature. 2020;583(7816):346-8.
28. Rubin D, Huang J, Fisher BT, Gasparrini A, Tam V, Song L, et al. Association of Social Distancing, Population Density, and Temperature With the Instantaneous Reproduction Number of SARS-CoV-2 in Counties Across the United States. JAMA Netw Open. 2020;3(7):e2016099.
29. Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One. 2007;2(8):e758.
30. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355-9.
31. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science. 2003;300(5627):1961-6.
32. Kaplan EH, Craft DL, Wein LM. Emergency response to a smallpox attack: the case for mass vaccination. Proc Natl Acad Sci U S A. 2002;99(16):10935-40.
33. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, et al. Transmission dynamics and control of severe acute respiratory syndrome. Science. 2003;300(5627):1966-70.
34. Inglesby TV. Public Health Measures and the Reproduction Number of SARS-CoV-2. JAMA. 2020;323(21):2186-7.
35. Cowling BJ, Ali ST, Ng TWY, Tsang TK, Li JCM, Fong MW, et al. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. Lancet Public Health. 2020;5(5):e279-e88.
36. Nishiura H, Satou K. Potential effectiveness of public health interventions during the equine influenza outbreak in racehorse facilities in Japan, 2007. Transbound Emerg Dis. 2010;57(3):162-70.
37. Linka K, Peirlinck M, Kuhl E. The reproduction number of COVID-19 and its correlation with public health interventions. Comput Mech. 2020:1-16.
38. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020;584(7820):257-61.
39. Haug N, Geyrhofer L, Londei A, Dervic E, Desvars-Larrive A, Loreto V, et al. Ranking the effectiveness of worldwide COVID-19 government interventions. Nat Hum Behav. 2020;4(12):1303-12.
40. Abat C, Chaudet H, Rolain JM, Colson P, Raoult D. Traditional and syndromic surveillance of infectious diseases and pathogens. Int J Infect Dis. 2016;48:22-8.

41. Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. Progress and challenges in virus genomic epidemiology. Trends Parasitol. 2021;37(12):1038-49.
42. Pace NR, Sapp J, Goldenfeld N. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proc Natl Acad Sci U S A. 2012;109(4):1011-8.
43. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013;9(3):e1002947.
44. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004;303(5656):327-32.
45. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009;10(8):540-50.
46. Holmes EC. The phylogeography of human viruses. Mol Ecol. 2004;13(4):745-56.
47. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520.
48. Pybus OG. Model selection and the molecular clock. PLoS Biol. 2006;4(5):e151.
49. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16(2):111-20.
50. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences. 1986;17(2):57-86.
51. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9(8):772.
52. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol. 2004;53(5):793-808.
53. Pybus OG, Rambaut A. GENIE: estimating demographic history from molecular phylogenies. Bioinformatics. 2002;18(10):1404-5.
54. Wakeley J. Coalescent theory: an introduction2009.
55. Holmes EC, Nee S, Rambaut A, Garnett GP, Harvey PH. Revealing the history of infectious disease epidemics through phylogenetic trees. Philos Trans R Soc Lond B Biol Sci. 1995;349(1327):33-40.
56. Nee S, Holmes EC, Rambaut A, Harvey PH. Inferring population history from molecular phylogenies. Philos Trans R Soc Lond B Biol Sci. 1995;349(1327):25-31.
57. Kingman JFC. The coalescent. Stochastic Processes and their Applications. 1982;13(3):235-48.
58. Kuhner MK, Yamato J, Felsenstein J. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics. 1998;149(1):429-34.
59. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 2000;155(3):1429-37.
60. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 2002;161(3):1307-20.
61. Hill V, Baele G. Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. Mol Biol Evol. 2019.
62. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol. 2008;25(7):1459-71.
63. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2013;30(3):713-24.

64. Gill MS, Lemey P, Bennett SN, Biek R, Suchard MA. Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates. Syst Biol. 2016;65(6):1041-56.
65. Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, Bürgisser P, et al. Estimating the basic reproductive number from viral sequence data. Mol Biol Evol. 2012;29(1):347-57.
66. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A. 2013;110(1):228-33.
67. Vasylyeva TI, du Plessis L, Pineda-Peña AC, Kühnert D, Lemey P, Vandamme AM, et al. Tracing the Impact of Public Health Interventions on HIV-1 Transmission in Portugal Using Molecular Epidemiology. J Infect Dis. 2019;220(2):233-43.
68. Baele G, Suchard MA, Rambaut A, Lemey P. Emerging Concepts of Data Integration in Pathogen Phylodynamics. Syst Biol. 2017;66(1):e47-e65.
69. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, et al. Emergence of Zaire Ebola virus disease in Guinea. N Engl J Med. 2014;371(15):1418-25.
70. Dudas G, Rambaut A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. PLoS Curr. 2014;6.
71. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546(7658):406-10.
72. Rambaut A, Holmes E. The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. PLoS Curr. 2009;1:RRN1003.
73. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544(7650):309-15.
74. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Elife. 2018;7.
75. Faria NR, Kraemer MUG, Hill SC, Goes de Jesus J, Aguiar RS, Iani FCM, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science. 2018;361(6405):894-9.
76. Vasylyeva TI, Liulchuk M, Friedman SR, Sazonova I, Faria NR, Katzourakis A, et al. Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine. Proc Natl Acad Sci U S A. 2018;115(5):1051-6.
77. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014;10(2):e1003932.
78. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, et al. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Comput Biol. 2014;10(4):e1003505.
79. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. Science. 2009;324(5934):1557-61.
80. Hedge J, Lycett SJ, Rambaut A. Real-time characterization of the molecular epidemiology of an influenza pandemic. Biol Lett. 2013;9(5):20130331.
81. dos Reis M, Donoghue PC, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet. 2016;17(2):71-80.
82. Britten RJ. Rates of DNA Sequence Evolution Differ between Taxonomic Groups. Science. 1986;231(4744):1393-8.
83. Hasegawa M, Kishino H. Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. Jpn J Genet. 1989;64(4):243-58.

84. Felsenstein J. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Systematic Zoology. 1978;27(4):401-10.
85. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4(5):e88.
86. Yoder AD, Yang Z. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol. 2000;17(7):1081-90.
87. Rambaut A, Bromham L. Estimating divergence dates from molecular sequences. Mol Biol Evol. 1998;15(4):442-8.
88. Hasegawa M, Kishino H, Yano T-a. Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. Journal of Human Evolution. 1989;18(5):461-76.
89. Thorne JL, Kishino H, Painter IS. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol. 1998;15(12):1647-57.
90. Aris-Brosou S, Yang Z. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst Biol. 2002;51(5):703-14.
91. Drummond AJ, Suchard MA. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 2010;8:114.
92. Holmes EC. Evolutionary history and phylogeography of human viruses. Annu Rev Microbiol. 2008;62:307-28.
93. Faria NR, Suchard MA, Rambaut A, Lemey P. Toward a quantitative understanding of viral phylogeography. Curr Opin Virol. 2011;1(5):423-9.
94. Lemmon AR, Lemmon EM. A likelihood framework for estimating phylogeographic history on a continuous landscape. Syst Biol. 2008;57(4):544-61.
95. Pagel M. The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies. Systematic Biology. 1999;48(3):612-22.
96. Pagel M, Meade A, Barker D. Bayesian estimation of ancestral character states on phylogenies. Syst Biol. 2004;53(5):673-84.
97. Minin VN, Suchard MA. Fast, accurate and simulation-free stochastic mapping. Philos Trans R Soc Lond B Biol Sci. 2008;363(1512):3985-95.
98. O'Brien JD, Minin VN, Suchard MA. Learning to count: robust estimates for labeled distances between molecular sequences. Mol Biol Evol. 2009;26(4):801-14.
99. Minin VN, Suchard MA. Counting labeled transitions in continuous-time Markov models of evolution. J Math Biol. 2008;56(3):391-412.
100. Nunes MR, Faria NR, Vasconcelos HB, Medeiros DB, Silva de Lima CP, Carvalho VL, et al. Phylogeography of dengue virus serotype 4, Brazil, 2010-2011. Emerg Infect Dis. 2012;18(11):1858-64.
101. Lemey P, Ruktanonchai N, Hong SL, Colizza V, Poletto C, Van den Broeck F, et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. Nature. 2021;595(7869):713-7.
102. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Mol Biol Evol. 2010;27(8):1877-85.
103. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol Biol Evol. 2016;33(8):2167-9.
104. Dellicour S, Rose R, Faria NR, Lemey P, Pybus OG. SERAPHIM: studying environmental rasters and phylogenetically informed movements. Bioinformatics. 2016;32(20):3204-6.

105. WHO. Genomic Sequencing of SARS-CoV-2: a Guide to Implementation for Maximum Impact on Public Health. In: WHO, editor. Geneva2021.
106. Skowronski DM, Astell C, Brunham RC, Low DE, Petric M, Roper RL, et al. Severe acute respiratory syndrome (SARS): a year in review. Annu Rev Med. 2005;56:357-81.
107. Hu J, Wang J, Xu J, Li W, Han Y, Li Y, et al. Evolution and variation of the SARS-CoV genome. Genomics Proteomics Bioinformatics. 2003;1(3):216-25.
108. Qin E, He X, Tian W, Liu Y, Li W, Wen J, et al. A genome sequence of novel SARS-CoV isolates: the genotype, GD-Ins29, leads to a hypothesis of viral transmission in South China. Genomics Proteomics Bioinformatics. 2003;1(2):101-7.
109. Stadler K, Masignani V, Eickmann M, Becker S, Abrignani S, Klenk HD, et al. SARS--beginning to understand a new virus. Nat Rev Microbiol. 2003;1(3):209-18.
110. Vijayanand P, Wilkins E, Woodhead M. Severe acute respiratory syndrome (SARS): a review. Clin Med (Lond). 2004;4(2):152-60.
111. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459(7250):1122-5.
112. Mena I, Nelson MI, Quezada-Monroy F, Dutta J, Cortes-Fernández R, Lara-Puente JH, et al. Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. Elife. 2016;5.
113. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22(13).
114. Butler D. Swine flu goes global. Nature. 2009;458(7242):1082-3.
115. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. Nature. 2016;538(7624):193-200.
116. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345(6202):1369-72.
117. Holmes EC. The Evolution and Emergence of RNA Viruses: Oxford University Press; 2009.
118. Ladner JT, Wiley MR, Mate S, Dudas G, Prieto K, Lovett S, et al. Evolution and Spread of Ebola Virus in Liberia, 2014-2015. Cell Host Microbe. 2015;18(6):659-69.
119. Fels JM, Bortz RH, Alkutkar T, Mittler E, Jangra RK, Spence JS, et al. A Glycoprotein Mutation That Emerged during the 2013-2016 Ebola Virus Epidemic Alters Proteolysis and Accelerates Membrane Fusion. mBio. 2021;12(1).
120. Deen GF, Broutet N, Xu W, Knust B, Sesay FR, McDonald SLR, et al. Ebola RNA Persistence in Semen of Ebola Virus Disease Survivors - Final Report. N Engl J Med. 2017;377(15):1428-37.
121. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, et al. Molecular Evidence of Sexual Transmission of Ebola Virus. N Engl J Med. 2015;373(25):2448-54.
122. Christie A, Davies-Wayne GJ, Cordier-Lassalle T, Cordier-Lasalle T, Blackley DJ, Laney AS, et al. Possible sexual transmission of Ebola virus - Liberia, 2015. MMWR Morb Mortal Wkly Rep. 2015;64(17):479-81.
123. Fischer RJ, Judson S, Miazgowicz K, Bushmaker T, Munster VJ. Ebola Virus Persistence in Semen Ex Vivo. Emerg Infect Dis. 2016;22(2):289-91.
124. Thorson A, Formenty P, Lofthouse C, Broutet N. Systematic review of the literature on viral persistence and sexual transmission from recovered Ebola survivors: evidence and recommendations. BMJ Open. 2016;6(1):e008859.
125. Tong YG, Shi WF, Liu D, Qian J, Liang L, Bo XC, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature. 2015;524(7563):93-6.

126. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228-32.
127. Arias A, Watson SJ, Asogun D, Tobin EA, Lu J, Phan MVT, et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. Virus Evol. 2016;2(1):vew016.
128. Stadler T, Kühnert D, Rasmussen DA, du Plessis L. Insights into the early epidemic spread of ebola in sierra leone provided by viral sequence data. PLoS Curr. 2014;6.
129. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017;12(6):1261-76.
130. Xavier J, Giovanetti M, Fonseca V, Thézé J, Gräf T, Fabri A, et al. Circulation of chikungunya virus East/Central/South African lineage in Rio de Janeiro, Brazil. PLoS One. 2019;14(6):e0217871.
131. de Lima STS, de Souza WM, Cavalcante JW, da Silva Candido D, Fumagalli MJ, Carrera JP, et al. Fatal Outcome of Chikungunya Virus Infection in Brazil. Clin Infect Dis. 2021;73(7):e2436-e43.
132. de Jesus JG, Dutra KR, Sales FCDS, Claro IM, Terzian AC, Candido DDS, et al. Genomic detection of a virus lineage replacement event of dengue virus serotype 2 in Brazil, 2019. Mem Inst Oswaldo Cruz. 2020;115:e190423.
133. Faria NR, Sabino EC, Nunes MR, Alcantara LC, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. Genome Med. 2016;8(1):97.
134. Viruses CSGotICoTo. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5(4):536-44.
135. Payne S. Family Coronaviridae. Viruses. 2017:149-58.
136. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology. 2019;17(3):181-92.
137. Yin Y, Wunderink RG. MERS, SARS and other coronaviruses as causes of pneumonia. Respirology. 2018;23(2):130-7.
138. Zhao X, Ding Y, Du J, Fan Y. 2020 update on human coronaviruses: One health, one world. Med Nov Technol Devices. 2020;8:100043.
139. Woo PC, Lau SK, Yuen KY. Clinical features and molecular epidemiology of coronavirus-HKU1-associated community-acquired pneumonia. Hong Kong Med J. 2009;15 Suppl 9:46-7.
140. Chen B, Tian EK, He B, Tian L, Han R, Wang S, et al. Overview of lethal human coronaviruses. Signal Transduct Target Ther. 2020;5(1):89.
141. Peiris JS, Yuen KY, Osterhaus AD, Stöhr K. The severe acute respiratory syndrome. N Engl J Med. 2003;349(25):2431-41.
142. Breiman RF, Evans MR, Preiser W, Maguire J, Schnur A, Li A, et al. Role of China in the quest to define and control severe acute respiratory syndrome. Emerg Infect Dis. 2003;9(9):1037-41.
143. Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science. 2003;302(5643):276-8.
144. Xu RH, He JF, Evans MR, Peng GW, Field HE, Yu DW, et al. Epidemiologic clues to SARS origin in China. Emerg Infect Dis. 2004;10(6):1030-7.
145. Tsang KW, Ho PL, Ooi GC, Yee WK, Wang T, Chan-Yeung M, et al. A cluster of cases of severe acute respiratory syndrome in Hong Kong. N Engl J Med. 2003;348(20):1977-85.

146. WHO. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. WHO; 2003.
147. Lee N, Hui D, Wu A, Chan P, Cameron P, Joynt GM, et al. A major outbreak of severe acute respiratory syndrome in Hong Kong. N Engl J Med. 2003;348(20):1986-94.
148. Pessôa R, Patriota JV, Lourdes de Souza M, Felix AC, Mamede N, Sanabani SS. Investigation Into an Outbreak of Dengue-like Illness in Pernambuco, Brazil, Revealed a Cocirculation of Zika, Chikungunya, and Dengue Virus Type 1. Medicine (Baltimore). 2016;95(12):e3201.
149. Scales DC, Green K, Chan AK, Poutanen SM, Foster D, Nowak K, et al. Illness in intensive care staff after brief exposure to severe acute respiratory syndrome. Emerg Infect Dis. 2003;9(10):1205-10.
150. Wong G, Liu W, Liu Y, Zhou B, Bi Y, Gao GF. MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease. Cell Host Microbe. 2015;18(4):398-401.
151. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012;367(19):1814-20.
152. WHO. MERS situation update September 2019. WHO; 2019.
153. WHO. MERS situation report August 2021. WHO; 2021.
154. Azhar EI, El-Kafrawy SA, Farraj SA, Hassan AM, Al-Saeed MS, Hashem AM, et al. Evidence for camel-to-human transmission of MERS coronavirus. N Engl J Med. 2014;370(26):2499-505.
155. Conzade R, Grant R, Malik MR, Elkholy A, Elhakim M, Samhouri D, et al. Reported Direct and Indirect Contact with Dromedary Camels among Laboratory-Confirmed MERS-CoV Cases. Viruses. 2018;10(8).
156. Perera RA, Wang P, Gomaa MR, El-Shesheny R, Kandeil A, Bagato O, et al. Seroepidemiology for MERS coronavirus using microneutralisation and pseudoparticle virus neutralisation assays reveal a high prevalence of antibody in dromedary camels in Egypt, June 2013. Euro Surveill. 2013;18(36):pii=20574.
157. Reusken CB, Haagmans BL, Müller MA, Gutierrez C, Godeke GJ, Meyer B, et al. Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. Lancet Infect Dis. 2013;13(10):859-66.
158. Müller MA, Corman VM, Jores J, Meyer B, Younan M, Liljander A, et al. MERS coronavirus neutralizing antibodies in camels, Eastern Africa, 1983-1997. Emerg Infect Dis. 2014;20(12):2093-5.
159. Donnelly CA, Malik MR, Elkholy A, Cauchemez S, Van Kerkhove MD. Worldwide Reduction in MERS Cases and Deaths since 2016. Emerg Infect Dis. 2019;25(9):1758-60.
160. Ball P. The lightning-fast quest for COVID vaccines - and what it means for other diseases. Nature. 2021;589(7840):16-8.
161. Kim J-H, Ah Reum An J, Oh SJ, Oh J, Lee J-K. Emerging COVID-19 success story: South Korea learned the lessons of MERS: Exemplars in Global Health; 2021 [Available from: https://ourworldindata.org/covid-exemplar-south-korea#licence.
162. Hilgenfeld R, Peiris M. From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses. Antiviral Res. 2013;100(1):286-95.
163. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26(4):450-2.
164. WHO. Pneumonia of unknown cause – China 2020 [Available from: https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229.

165. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-3.
166. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.
167. WHO. Archived: WHO Timeline - COVID-19: WHO; 2020 [Available from: https://www.who.int/news/item/27-04-2020-who-timeline---covid-19.
168. Holmes EC. Novel 2019 coronavirus genome: Virological; 2020 [Available from: https://virological.org/t/novel-2019-coronavirus-genome/319.
169. Yu P, Zhu J, Zhang Z, Han Y. A Familial Cluster of Infection Associated With the 2019 Novel Coronavirus Indicating Possible Person-to-Person Transmission During the Incubation Period. J Infect Dis. 2020;221(11):1757-61.
170. Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol. 2021;19(3):141-54.
171. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507-13.
172. WHO. WHO statement on novel coronavirus in Thailand: WHO; 2020 [Available from: https://www.who.int/news/item/13-01-2020-who-statement-on-novel-coronavirus-in-thailand.
173. CDC. First Travel-related Case of 2019 Novel Coronavirus Detected in United States: CDC; 2020 [Available from: https://www.cdc.gov/media/releases/2020/p0121-novel-coronavirus-travel-case.html.
174. Tian H, Liu Y, Li Y, Wu CH, Chen B, Kraemer MUG, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science. 2020;368(6491):638-42.
175. WHO. Novel Coronavirus (2019-nCoV) Situation Report -10 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2.
176. WHO. Novel Coronavirus (2019-nCoV) Situation Report - 13 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2.
177. WHO. Novel Coronavirus (2019-nCoV) Situation Report - 25 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2.
178. WHO. Novel Coronavirus (2019-nCoV) Situation Report - 27 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2.
179. WHO. Novel Coronavirus (2019-nCoV) Situation Report - 51 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2.
180. WHO. Novel Coronavirus (2019-nCoV) Situation Report - 54 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2.
181. Cena L, Rota M, Calza S, Massardi B, Trainini A, Stefana A. Estimating the Impact of the COVID-19 Pandemic on Maternal and Perinatal Health Care Services in Italy: Results of a Self-Administered Survey. Front Public Health. 2021;9:701638.
182. WHO. Coronavirus disease (Covid-19) Situation Report - 105: WHO; 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200504-covid-19-sitrep-105.pdf?sfvrsn=4cdda8af_2.

183. Thompson CN, Baumgartner J, Pichardo C, Toro B, Li L, Arciuolo R, et al. COVID-19 Outbreak - New York City, February 29-June 1, 2020. MMWR Morb Mortal Wkly Rep. 2020;69(46):1725-9.
184. Omer SB, Malani P, Del Rio C. The COVID-19 Pandemic in the US: A Clinical Update. JAMA. 2020;323(18):1767-8.
185. Altman D. Understanding the US failure on coronavirus-an essay by Drew Altman. BMJ. 2020;370:m3417.
186. WHO. Coronavirus disease 2019 (COVID-19) Situation Report - 38: WHO; 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200227-sitrep-38-covid-19.pdf?sfvrsn=2db7a09b_4.
187. WHO. Coronavirus disease 2019 (COVID-19) Situation Report - 40: WHO; 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200227-sitrep-38-covid-19.pdf?sfvrsn=2db7a09b_4.
188. WHO. Coronavirus disease 2019 (COVID-19) Situation Report - 41: WHO; 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200227-sitrep-38-covid-19.pdf?sfvrsn=2db7a09b_4.
189. WHO. Coronavirus disease 2019 (COVID-19) Situation Report - 50: WHO; 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200227-sitrep-38-covid-19.pdf?sfvrsn=2db7a09b_4.
190. WHO. Coronavirus disease 2019 (COVID-19) Situation Report - 47: WHO; 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200227-sitrep-38-covid-19.pdf?sfvrsn=2db7a09b_4.
191. Dávila-Cervantes CA, Agudelo-Botero M. Health inequalities in Latin America: persistent gaps in life expectancy. Lancet Planet Health. 2019;3(12):e492-e3.
192. Taylor L. Uruguay is winning against covid-19. This is how. BMJ. 2020;370:m3575.
193. Bojorquez I, Cabieses B, Arósquipa C, Arroyo J, Novella AC, Knipper M, et al. Migration and health in Latin America during the COVID-19 pandemic and beyond. Lancet. 2021;397(10281):1243-5.
194. Garcia PJ, Alarcón A, Bayer A, Buss P, Guerra G, Ribeiro H, et al. COVID-19 Response in Latin America. Am J Trop Med Hyg. 2020;103(5):1765-72.
195. Rubin R, Abbasi J, Voelker R. Latin America and Its Global Partners Toil to Procure Medical Supplies as COVID-19 Pushes the Region to Its Limit. JAMA. 2020;324(3):217-9.
196. ECLAC. Restrictions on the export of medical products hamper efforts to contain coronavirus disease (COVID-19) in Latin America and the Caribbean: ECLAC; 2020 [
197. Burki T. COVID-19 in Latin America. Lancet Infect Dis. 2020;20(5):547-8.
198. Faiola, Anthony, Herrero, Vanessa A. Bodies lie in the streets of Guayaquil, Ecuador, emerging epicenter of the coronavirus in Latin America. The Washington Post. 2020.
199. Boadle A. WHO says the Americas are new COVID-19 epicenter as deaths surge in Latin America: Reuters; 2020 [Available from: https://www.reuters.com/article/us-health-coronavirus-latam-idUSKBN2322G6.
200. WHO. Coronavirus disease (COVID-19) Situation Report - 123: WHO; 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200522-covid-19-sitrep-123.pdf?sfvrsn=5ad1bc3_4.
201. Whittaker C, Walker PGT, Alhaffar M, Hamlet A, Djaafara BA, Ghani A, et al. Under-reporting of deaths limits our understanding of true burden of covid-19. BMJ. 2021;375:n2239.

202. Albani V, Loria J, Massad E, Zubelli J. COVID-19 underreporting and its impact on vaccination strategies. BMC Infect Dis. 2021;21(1):1111.
203. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9(4):267-76.
204. Wain-Hobson S. Retrovirus Evolution. Origin and Evolution of Viruses. 2008:259-77.
205. Smith EC, Sexton NR, Denison MR. Thinking Outside the Triangle: Replication Fidelity of the Largest RNA Viruses. Annu Rev Virol. 2014;1(1):111-32.
206. Lee CH, Gilbertson DL, Novella IS, Huerta R, Domingo E, Holland JJ. Negative effects of chemical mutagenesis on the adaptive behavior of vesicular stomatitis virus. J Virol. 1997;71(5):3636-40.
207. Pfeiffer JK, Kirkegaard K. Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. PLoS Pathog. 2005;1(2):e11.
208. Pfeiffer JK, Kirkegaard K. A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity. Proc Natl Acad Sci U S A. 2003;100(12):7289-94.
209. Vignuzzi M, Wendt E, Andino R. Engineering attenuated virus vaccines by controlling replication fidelity. Nat Med. 2008;14(2):154-61.
210. Nga PT, Parquet MeC, Lauber C, Parida M, Nabeshima T, Yu F, et al. Discovery of the first insect nidovirus, a missing evolutionary link in the emergence of the largest RNA virus genomes. PLoS Pathog. 2011;7(9):e1002215.
211. Smith EC, Denison MR. Coronaviruses as DNA wannabes: a new model for the regulation of RNA virus replication fidelity. PLoS Pathog. 2013;9(12):e1003760.
212. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. Nidovirales: evolving the largest RNA virus genome. Virus Res. 2006;117(1):17-37.
213. Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, et al. Discovery of an RNA virus 3'->5' exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc Natl Acad Sci U S A. 2006;103(13):5108-13.
214. Ivanov KA, Hertzig T, Rozanov M, Bayer S, Thiel V, Gorbalenya AE, et al. Major genetic marker of nidoviruses encodes a replicative endoribonuclease. Proc Natl Acad Sci U S A. 2004;101(34):12694-9.
215. Ferron F, Subissi L, Silveira De Morais AT, Le NTT, Sevajol M, Gluais L, et al. Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. Proc Natl Acad Sci U S A. 2018;115(2):E162-E71.
216. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, et al. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog. 2010;6(5):e1000896.
217. Tao K, Tzou PL, Nouhin J, Gupta RK, de Oliveira T, Kosakovsky Pond SL, et al. The biological and clinical significance of emerging SARS-CoV-2 variants. Nat Rev Genet. 2021;22(12):757-73.
218. Gribble J, Stevens LJ, Agostini ML, Anderson-Daniels J, Chappell JD, Lu X, et al. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. PLoS Pathog. 2021;17(1):e1009226.
219. Rambaut A. Phylogenetic analysis of nCoV-2019 genomes: Virological; 2020 [Available from: https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356.
220. MacLean OA, Orton RJ, Singer JB, Robertson DL. No evidence for distinct types in the evolution of SARS-CoV-2. Virus Evol. 2020;6(1):veaa034.

221. Pereson MJ, Flichman DM, Martínez AP, Baré P, Garcia GH, Di Lello FA. Evolutionary analysis of SARS-CoV-2 spike protein for its different clades. J Med Virol. 2021;93(5):3000-6.
222. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5(11):1403-7.
223. Zhang L, Yang J-R, Zhang Z, Lin Z. Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission. medRxiv. 2020:2020.02.25.20027953.
224. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395(10224):565-74.
225. Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, et al. The origins of SARS-CoV-2: A critical review. Cell. 2021;184(19):4848-56.
226. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020;182(4):812-27.e19.
227. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2020.
228. Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun. 2020;11(1):6013.
229. Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, tenOever BR, et al. The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. Elife. 2021;10.
230. Zhou B, Thao TTN, Hoffmann D, Taddeo A, Ebert N, Labroussaa F, et al. SARS-CoV-2 spike D614G change enhances replication and transmission. Nature. 2021;592(7852):122-7.
231. Burioni R, Topol EJ. Has SARS-CoV-2 reached peak fitness? Nat Med. 2021;27(8):1323-4.
232. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell. 2021;184(1):64-75.e11.
233. WHO. Tracking SARS-CoV-2 variants: WHO; 2021 [Available from: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.
234. ECDC. SARS-CoV-2 variants of concern: ECDC; 2021 [Available from: https://www.ecdc.europa.eu/en/covid-19/variants-concern.
235. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software.6(67):3773.
236. Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations: Virological; 2020 [Available from: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
237. Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. N Engl J Med. 2020;383(23):2291-3.

238. Avanzato VA, Matson MJ, Seifert SN, Pryce R, Williamson BN, Anzick SL, et al. Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. Cell. 2020;183(7):1901-12.e9.
239. Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature. 2021;592(7853):277-82.
240. Supasa P, Zhou D, Dejnirattisai W, Liu C, Mentzer AJ, Ginn HM, et al. Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. Cell. 2021;184(8):2201-11.e7.
241. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science. 2021;372(6538).
242. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 2021;593(7858):266-9.
243. Mishra S, Mindermann S, Sharma M, Whittaker C, Mellan TA, Wilton T, et al. Changing composition of SARS-CoV-2 lineages and rise of Delta variant in England. EClinicalMedicine. 2021;39:101064.
244. Hodcroft EB. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. 2021 [Available from: https://covariants.org/.
245. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021;592(7854):438-43.
246. Barton MI, MacGowan SA, Kutuzov MA, Dushek O, Barton GJ, van der Merwe PA. Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. Elife. 2021;10.
247. Funk T, Pharris A, Spiteri G, Bundle N, Melidou A, Carr M, et al. Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. Euro Surveill. 2021;26(16).
248. Pearson CAB, Russell TW, Davies N, Kucharski AJ, Edmunds J, Eggo RM. Estimates of severity and transmissibility of novel SARS-CoV-2 variant 501Y.V2 in South Africa: CMMID/LSHTM; 2021 [Available from: https://cmmid.github.io/topics/covid19/sa-novel-variant.html.
249. Madhi SA, Baillie V, Cutland CL, Voysey M, Koen AL, Fairlie L, et al. Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. N Engl J Med. 2021;384(20):1885-98.
250. Cele S, Gazy I, Jackson L, Hwa SH, Tegally H, Lustig G, et al. Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. Nature. 2021;593(7857):142-6.
251. NIID-Japan. Brief report: New Variant Strainof SARS-CoV-2Identified in Travelers from Brazil: NIID-Japan; 2021 [Available from: https://www.niid.go.jp/niid/images/epi/corona/covid19-33-en-210112.pdf.
252. Faria N, Claro I, Candido D, Moyses Franco L, Andrade P, Coletti T, et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. 2021. 2021.
253. Naveca F, Nascimento V, Souza V, Corado A, Nascimento F, Silva G, et al. Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein. Virological org. 2021;1:1-8.

254. Sabino EC, Buss LF, Carvalho MPS, Prete CA, Jr., Crispim MAE, Fraiji NA, et al. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. Lancet. 2021;397(10273):452-5.
255. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DDS, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 2021;372(6544):815-21.
256. Dejnirattisai W, Zhou D, Supasa P, Liu C, Mentzer AJ, Ginn HM, et al. Antibody evasion by the P.1 strain of SARS-CoV-2. Cell. 2021;184(11):2939-54.e9.
257. Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, et al. SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. Microorganisms. 2021;9(7).
258. Liu Y, Liu J, Johnson BA, Xia H, Ku Z, Schindewolf C, et al. Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. bioRxiv. 2021:2021.08.12.456173.
259. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. Cell. 2020;182(5):1284-94.e9.
260. SPI-M-O. SPI-M-O: Consensus statement on COVID-19, 12 May 2021: Public Health England; 2021 [Available from: https://www.gov.uk/government/publications/spi-m-o-consensus-statement-on-covid-19-12-may-2021.
261. Sheikh A, McMenamin J, Taylor B, Robertson C, Collaborators PHSatEI. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. Lancet. 2021;397(10293):2461-2.
262. Lopez Bernal J, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. N Engl J Med. 2021;385(7):585-94.
263. Naveca FG, Nascimento V, Souza V, de Lima Corado A, Nascimento F, Mejía M, et al. The SARS-CoV-2 variant Delta displaced the variants Gamma and Gamma plus in Amazonas, Brazil.
264. Torjesen I. Covid-19: Delta variant is now UK's most dominant strain and spreading through schools. BMJ. 2021;373:n1445.
265. Mullen JL, Tsueng G, Latif AA, Alkuzweny M, Cano M, Haag E, et al. outbreak.info. 2020.
266. Scott L, Hsiao NY, Moyo S, Singh L, Tegally H, Dor G, et al. Track Omicron's spread with molecular data. Science. 2021;374(6574):1454-5.
267. NGS-SA. SARS-CoV-2 Sequencing Update26 November 2021 https://www.nicd.ac.za: Network for Genomic Surveillance in South Africa (NGS-SA); 2021 [Available from: https://www.nicd.ac.za/wp-content/uploads/2021/11/Update-of-SA-sequencing-data-from-GISAID-26-Nov_Final.pdf.
268. Buchan SA, Chung H, Brown KA, Austin PC, Fell DB, Gubbay JB, et al. Effectiveness of COVID-19 vaccines against Omicron or Delta infection. medRxiv. 2022:2021.12.30.21268565.
269. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Effectiveness of COVID-19 vaccines against the Omicron (B.1.1.529) variant of concern. medRxiv. 2021:2021.12.14.21267615.
270. Christensen PA, Olsen RJ, Long SW, Snehal R, Davis JJ, Saavedra MO, et al. Early signals of significantly increased vaccine breakthrough, decreased hospitalization rates, and less severe disease in patients with COVID-19 caused by the Omicron variant of SARS-CoV-2 in Houston, Texas. medRxiv. 2022:2021.12.30.21268560.

271. Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, Groome M, et al. Early assessment of the clinical severity of the SARS-CoV-2 Omicron variant in South Africa. medRxiv. 2021:2021.12.21.21268116.
272. Abdelnabi R, Foo CS, Zhang X, Lemmens V, Maes P, Slechten B, et al. The omicron (B.1.1.529) SARS-CoV-2 variant of concern does not readily infect Syrian hamsters. bioRxiv. 2021:2021.12.24.474086.
273. Bentley EG, Kirby A, Sharma P, Kipar A, Mega DF, Bramwell C, et al. SARS-CoV-2 Omicron-B.1.1.529 Variant leads to less severe disease than Pango B and Delta variants strains in a mouse model of severe COVID-19. bioRxiv. 2021:2021.12.26.474085.
274. IBGE. Instituto Brasileiro de Geografia e Estatística: IBGE - Instituto Brasileiro de Geografia e Estatística; [Available from: https://www.ibge.gov.br/.
275. Bank W. GDP (current US$). The World Bank Group.
276. Bank W. Gini index (World Bank estimate): The World Bank Group; [Available from: https://data.worldbank.org/indicator/SI.POV.GINI.
277. Mourao P, Junqueira A. Through the irregular paths of inequality: an analysis of the evolution of socioeconomic inequality in Brazilian states since 1976. Sustainability. 2021;13(4):2356.
278. Silva SAd. Regional inequalities in Brazil: Divergent readings on their origin and public policy design. EchoGéo. 2017(41).
279. ANAC. Consulta Interativa – Indicadores do Mercado de Transporte Aéreo https://www.gov.br/anac/: Agência Nacional de Aviação Civil; 2019 [Available from: https://www.gov.br/anac/pt-br/assuntos/dados-e-estatisticas/mercado-de-transporte-aereo/consulta-interativa/demanda-e-oferta-origem-destino.
280. Deak M, Cerqueira C. Boletim# 1 –Março de 2019 saopaulo.sp.gov.br: Governo do estado de São Paulo; 2019 [Available from: https://www.desenvolvimentoeconomico.sp.gov.br/Content/uploads/Boletim%20diagnostico%20SP.pdf.
281. Hospital das Clínicas é destaque em rankings de Saúde https://eephcfmusp.org.br: Hospital da Clinicas da Faculdade de Medicina da Universidade de São Paulo; 2020 [Available from: https://eephcfmusp.org.br/portal/online/hospital-das-clinicas-saude/.
282. Amazonas. Dados http://www.amazonas.am.gov.br: Governo do Estado do Amazonas; [Available from: http://www.amazonas.am.gov.br/o-amazonas/dados/.
283. Amazonas. Economia http://www.amazonas.am.gov.br/: Governo do estado do Amazonas; [Available from: http://www.amazonas.am.gov.br/o-amazonas/economia/.
284. Rivas AAF, Kahn JR. Industrial Policy as an Environmental Policy: Forest Preservation and the Industrialization of Manaus. Oxford University Press; 2021.
285. PNUD. IDHM Municípios 2010 https://www.br.undp.org/: PNUD Brasil; 2010 [Available from: https://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.html.
286. Ferigato S, Fernandez M, Amorim M, Ambrogi I, Fernandes LMM, Pacheco R. The Brazilian Government's mistakes in responding to the COVID-19 pandemic. Lancet. 2020;396(10263):1636.
287. Teixeira MG, Costa MDCN, Paixão ESD, Carmo EH, Barreto FR, Penna GO. The achievements of the SUS in tackling the communicable diseases. Cien Saude Colet. 2018;23(6):1819-28.
288. Collaborators GCoD. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1151-210.

289. Gómez EJ. What the United States can learn from Brazil in response to HIV/AIDS: international reputation and strategic centralization in a context of health policy devolution. Health Policy Plan. 2010;25(6):529-41.
290. Okie S. Fighting HIV--lessons from Brazil. N Engl J Med. 2006;354(19):1977-81.
291. Berkman A, Garcia J, Muñoz-Laboy M, Paiva V, Parker R. A critical analysis of the Brazilian response to HIV/AIDS: lessons learned for controlling and mitigating the epidemic in developing countries. Am J Public Health. 2005;95(7):1162-72.
292. Galvão-Castro B, Couto-Fernandez JC, Mello MA, Linhares-de-Carvalho MI, Castello-Branco LR, Bongertz V, et al. A nationwide effort to sistematically monitor HIV-1 diversity in Brazil: preliminary results. Brazilian Network for the HIV-1 Isolation and Characterization. Mem Inst Oswaldo Cruz. 1996;91(3):335-8.
293. Silva SJRD, Magalhães JJF, Pena L. Simultaneous Circulation of DENV, CHIKV, ZIKV and SARS-CoV-2 in Brazil: an Inconvenient Truth. One Health. 2021;12:100205.
294. Nunes PCG, Daumas RP, Sánchez-Arcila JC, Nogueira RMR, Horta MAP, Dos Santos FB. 30 years of fatal dengue cases in Brazil: a review. BMC Public Health. 2019;19(1):329.
295. Brazil https://www.worldmosquitoprogram.org: World Mosquito Program; 2021 [Available from: https://www.worldmosquitoprogram.org/en/global-progress/brazil.
296. PAHO. Dengue Cases https://www3.paho.org/2021 [Available from: https://www3.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en/dengue-nacional-en/252-dengue-pais-ano-en.html.
297. PAHO. Cases of Chikungunya Virus Disease: PAHO; 2021 [Available from: https://www3.paho.org/data/index.php/en/mnu-topics/chikv-en/550-chikv-weekly-en.html.
298. Goes de Jesus J, Sacchi C, Claro I, Salles F, Manulli E, da Silva D, et al. First cases of coronavirus disease (COVID-19) in Brazil, South America (2 genomes, 3rd March 2020) Virological.: Virological.org; 2020 [

Chapter 2

Overview of SARS-COV-2 importation, initial spread and response in Brazil

When the first cases of a new pneumonia outbreak in Wuhan were reported to the WHO in 31st December 2019, later on spreading to other countries, our genome sequencing work in Brazil was focused on arboviruses and occasionally febrile disease of unknown aetiology. However, we soon started preparing for the likely importation of SARS-CoV-2 into Brazil by updating techniques and protocols, stocking reagents and capacitating our human resources. Soon, our team was engaging in daily meetings with scientists from Brazil, the UK and across the globe.

This chapter focuses on my contributions to some of the publications which resulted from the articulation of this scientific taskforce to rapidly respond to Brazil's public health crisis within the first pandemic of the 21st century. Here, I provide an overview to some of the key aspects of the SARS-CoV-2 pandemic in Brazil, serving as a background to what is presented in chapters 3, 4 and 5. As the findings presented here have been published in different studies, this chapter has been organized into four subchapters.

I started working on **Chapter 2.1**, "Routes for COVID-19 importation in Brazil", just before the first COVID-19 cases were confirmed in Brazil. This chapter's conception lies in the need to inform Brazilian public Health authorities, scientific and health communities of the likely routes through which SARS-CoV-2 would enter the country. At

that point, there were no such analyses integrating case counts and air travel flow to inform public health responses in Brazil. I was responsible for all aspects of this work and it is included in this thesis in full. This work was made available as a preprint on MedRxiv on the 18th of March 2020 and published in the *Journal of Travel Medicine*.

***Candido DDS**, Watts A, Abade L, Kraemer MUG, Pybus OG, Croda J, et al. Routes for COVID-19 importation in Brazil. J Travel Med. 2020 May 18;27(3).*

Chapter 2.2, "Importation and early local transmission of COVID-19 in Brazil, 2020", resulted from an expansion of CADDE's 48-hour report on the sequencing of the first detected SARS-CoV-2 cases in Brazil and Latin America. It presents an epidemiological and genomic overview of the first imported cases and highlights the first cases of local transmission in Brazil. This work was made available published in the Revista do Instituto de Medicina Tropical de São Paulo and is also included in full in this thesis.

Jesus JG de, Sacchi C*, **Candido D da S***, Claro IM, Sales FCS, Manuli ER, et al. Importation and early local transmission of COVID-19 in Brazil, 2020. Rev Inst Med Trop Sao Paulo. 2020 May 11;62:e30.*

Chapter 2.3, "Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil", is a result of a major collaboration between CADDE researchers during the first few months of SARS-CoV-2 epidemic spread in Brazil. Led by William de Souza, this works is composed by the contributions of multiple shared-first authors. For this reason, this work is only partially included in this thesis, as my main contribution to this work was the estimation of the basic reproduction numbers for four Brazilian states (Amazonas, Ceará, Rio de Janeiro and São Paulo) and to five countries (Brazil, Italy, France, Spain and UK),

together with Alexander Zarebski. These estimates are interpreted in the context of the NPIs taken in different location and figures as an assessment of the early spread of SARS-CoV-2 in Brazil. This work as made available as a preprint on MedRxiv on the 29th April 2020 and was subsequently published in *Nature Human Behaviour*.

de Souza WM, Buss LF*, **Candido D da S***, Carrera J-P*, Li S*, Zarebski AE, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat Hum Behav. 2020 Aug;4(8):856–65.*

Finally, **Chapter 2.4**, "Dataset on SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities", was generated from a survey independently conducted by the Brazilian Confederation of Municipalities (Confederação Nacional de Municípios – CNM). This is the largest dataset to date on NPI's in Brazil and it figures as an important overview of the epi-political aspects of the Brazilian response to SARS-CoV-2 spread. This manuscript is presented here in full. It was first made available as a report and it was published in *Scientific Data*.

de Souza Santos AA, **Candido D da S***, de Souza WM*, Buss L, Li SL, Pereira RHM, et al. Dataset on SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities. Sci Data. 2021 Mar 4;8(1):73.*

"This pandemic has magnified every existing inequality in our society – like systemic racism, gender inequality, and poverty"

Melinda Gates

Journal of Travel Medicine

International Society of Travel Medicine
Established 1991
Promoting healthy travel worldwide

Journal of Travel Medicine, 2020, 1–3
doi: 10.1093/jtm/taaa042
Advance Access Publication Date: 23 March 2020
Rapid Communication

Rapid Communication

Routes for COVID-19 importation in Brazil

Darlan Da S. Candido, MSc[1], Alexander Watts, PhD[2,3], Leandro Abade, DPhil[1], Moritz U. G. Kraemer, DPhil[1,4,5], Oliver G. Pybus, DPhil[1,6], Julio Croda, MD, PhD[7,8,9], Wanderson de Oliveira, PhD[7], Kamran Khan, MD, MPH[2,3], Ester C. Sabino, PhD[10] and Nuno R. Faria, PhD[1,10,11]

[1]Department of Zoology, University of Oxford, Oxford, UK, [2]Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada, [3]Division of Infectious Diseases, Department of Medicine, University of Toronto, Toronto, ON, Canada, [4]Harvard Medical School, Harvard University, Boston, MA, USA, [5]Computational Epidemiology Group, Boston Children's Hospital, Boston, MA, USA, [6]Department of Pathobiology and Population Sciences, The Royal Veterinary College, London, UK, [7]Secretaria de Vigilância em Saúde, Coordenação Geral de Laboratórios de Saúde Pública, Ministério da Saúde, Brasília, Brazil, [8]Laboratório de Pesquisa em Ciências da Saúde, Universidade Federal da Grande Dourados, Dourados Mato Grosso do Sul, Brazil, [9]Fundação Oswaldo Cruz Campo Grande, Campo Grande, Brazil, [10]Instituto de Medicina Tropical, University of São Paulo, São Paulo, Brazil and [11]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London

*To whom correspondence should be addressed. Email: nuno.faria@zoo.ox.ac.uk

Submitted 12 March 2020; Revised 13 March 2020; Editorial Decision 16 March 2020; Accepted 16 March 2020

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV2) was first detected in Wuhan, Hubei province, China, on 8 December 2019. SARS-CoV-2 infection can cause coronavirus disease (COVID-19) and can lead to acute respiratory syndrome, hospitalization and death.[1] As of 12 March 2020, the global SARS-CoV-2 outbreak has been declared a pandemic, with 125 048 cases, and 4613 deaths have been notified by the World Health Organization (WHO) in 117 countries/territories or areas worldwide (who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports). The first case in Latin America was confirmed on 26 February 2020, in the São Paulo metropolis, the most populous city in the Southern hemisphere (~11 million people, Instituto Brasileiro de Geografia e Estatística, www.ibge.gov.br). Self-declared travel history and subsequent genetic analyses confirmed that the first detected infection was acquired via importation of the virus from Northern Italy.[2] Since then, Brazil has reported the largest number of cases in Latin America ($n = 34$, as of 10 March 2020). SARS-CoV-2 has been now detected in 7 (26%) of the 27 federal states of Brazil. So far, the transmission of SARS-CoV-2 appears to be primarily sporadic (85.3%, 29/34 are imported cases). Here, we analyze data on airline travellers to Brazil in 2019, who departed from countries that had reported local cases of COVID-19 transmission by 5 March 2020. This information provides insights into which Brazilian cities are most at risk for SARS-CoV-2 importation.

We used travel data on all air journeys that had a Brazilian city as their final destination during February and March 2019 as a proxy for flight density during the 2020 COVID-2019 outbreak (see Supplementary data). We focused on the data for 29 countries that had reported SARS-CoV-2 cases by 5 March 2020. We collated the total number of passengers flying to Brazilian airports during this period, country population size for 2019 from the United Nations World Population Prospects 2019 database, and the WHO-reported number of COVID-19 cases (as of 5 March 2020). We used these values to estimate the proportion of infected travellers potentially arriving in Brazilian cities from each country and for each route (additional information can be found in Supplementary data). No air passenger data from Iran and Portugal to Brazil were available for our analysis.

Between February and March 2019, Brazil received 841 302 international passengers in a total of 84 cities across the country (Figure 1). São Paulo, the largest city in the country, was the final destination of nearly half (46.1%) of the passengers arriving to Brazil, followed by Rio de Janeiro (21%) and Belo Horizonte (4.1%). More than half of the international passengers started their journey in the USA (50.8%) followed by France (7.9%) and

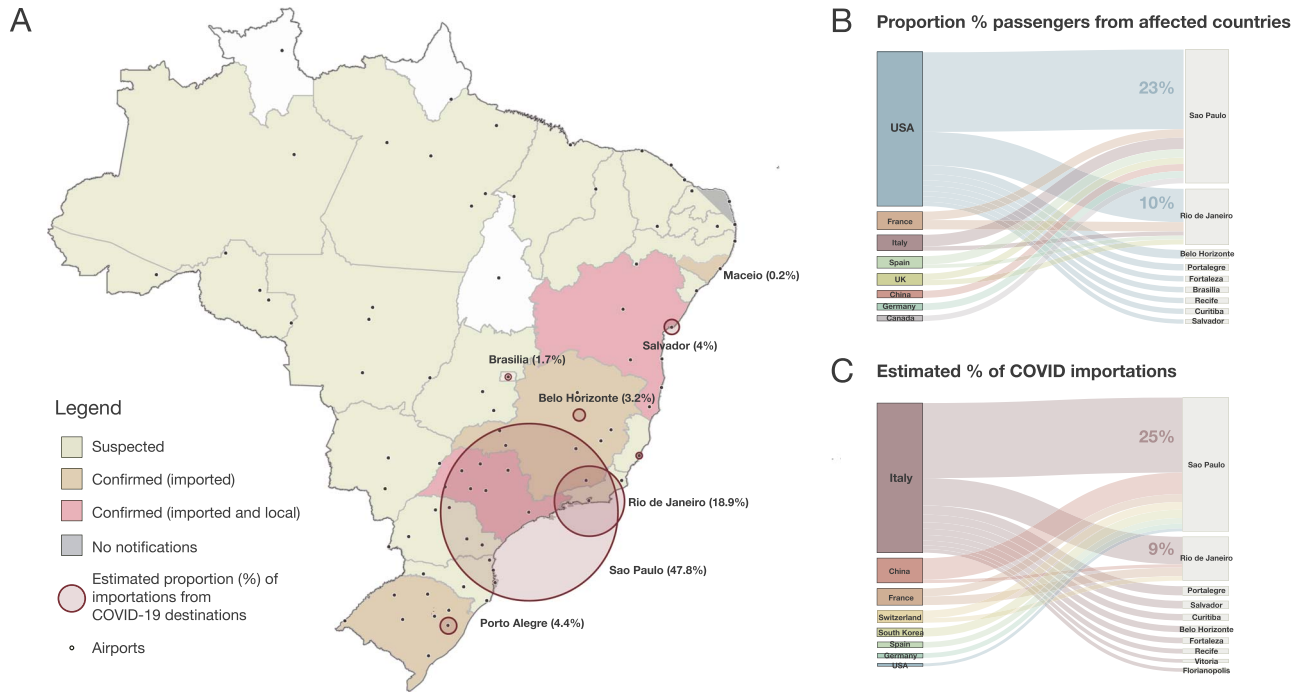


Figure 1. Potential for COVID-19 importation in Brazil. (A) Map of Brazilian federal states and federal district coloured according to COVID-19 notification status (as of 10 March 2020). Circles correspond to the estimated proportion of arrivals from the top 29 destinations (except Iran and Portugal) that had reported local COVID-19 by 5 March 2020. (B) Percentage of passengers for the top-20 routes to Brazilian airports from countries that had reported COVID-19 cases by 5 March 2020. (C) Estimated percentage of importations for the top-20 routes from countries that had reported local COVID-19 by 5 March 2020.

Italy (7.5%). The air-travel routes to airports in Brazil with most passengers were USA–São Paulo (23.3%), USA–Rio de Janeiro (9.8%) and Italy–São Paulo (3.4%).

To better understand the potential for SARS-CoV-2 introductions to Brazil, we estimate the relative risk of COVID-19 introduction to Brazilian cities by taking into account SARS-CoV-2 incidence per international traveller arriving at an airport in Brazil. We estimate that 54.8% of all imported cases would be expected to come from travellers infected in Italy and 9.3% and 8.3% of the cases would be from travellers infected in China and France, respectively. The route Italy–São Paulo was estimated to comprise 24.9% of total infected travellers flying to Brazil during this period. Moreover, we estimate that Italy has been the source location for five of the top 10 importation routes for infected travellers into Brazil based on the current epidemiological scenario (Supplementary data). Consistent with this, at least 48% ($n = 14/29$) of the reported imported cases in Brazil have a history of travelling to Italy prior to onset of symptoms, as of 9 March 2020. Six (23.1%) of the confirmed cases that acquired the virus in Italy have been identified in São Paulo (Supplementary data).

We found that the proportion of estimated imported cases by city of destination is highly correlated with the proportion of detected imported cases. Our study has several limitations. Unfortunately, data from Iran and Portugal were not available for this analysis. Moreover, our analysis relies on incidence data, and thus, the risk of importation will follow changes in epidemic sizes at source locations. In fact, with the reduction in the number of flights leaving from Italy and 51% of flights to Brazil departing from airports in the USA, we should anticipate an increasing proportion of infected travellers arriving from the USA. Moreover, the estimated risk of importation from China is likely an overestimate as recent measures have extensively decreased the flights to Brazil.

At a time when the number of SARS-CoV-2 cases is steadily growing in Brazil, our findings highlight the high potential for the introduction of new cases in several cities of Brazil, especially in São Paulo and Rio de Janeiro metropolises. Rapid identification of locations where clusters of local transmission might first ignite is critical to better coordinate preparedness, readiness and response actions.[3,4] There is a critical need for epidemiological, human mobility and genetic data[5] to understand virus transmission dynamics at local, regional and global scales. Continued integration of these data streams should help guide the deployment of resources to mitigate COVID-19 transmission.

Supplementary data

Supplementary data are available at *JTM* online.

Funding

This work was supported by a Medical Research Council and Fundação de Amparo à Pesquisa do Estado de São Paulo CADDE partnership award (MR/S0195/1) and a John Fell Research Fund (grant 005166). N.R.F. is supported by a Sir Henry Dale Fellowship (204311/Z/16/Z). D.D.S.C. is supported by the

Clarendon Fund and by the Oxford University Zoology Department.

Authors statements

K.K. is the Founder of BlueDot, a social enterprise that develops digital technologies for public health. K.K. and A.W. are employed at BlueDot. D.S.C., L.A., M.K., W.O., J.C., E.C.S., O.G.P. and N.R.F. have no conflicts of interest to declare.

Authors contributions

D.S.C., L.A. and N.R.F. conceived the idea and wrote the manuscript. D.S.C., L.A., N.R.F., K.K. and A.W. conducted data analysis. D.S.C., N.R.F., L.A., M.U.G.K., W.O., J.C., E.C.S., O.G.P., A.W. and K.K. interpreted data and contributed to writing.

References

1. Zhu N, Zhang D, Wang W *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020; **382**:727–33. doi: 10.1056/NEJMoa2001017.
2. de Jesus JG, Sacchi C, Claro I *et al.* First cases of coronavirus disease (COVID-19) in Brazil, South America (2 genomes, 3rd march 2020). http://virological.org/t/first-cases-of-coronavirus-disease-covid-19-in-brazil-south-america-2-genomes-3rd-march-2020/409, Virological, 2020.
3. Ministério da Saúde B. Brasil amplia monitoramento do coronavírus. 2020.
4. WHO. *Critical preparedness, readiness and response actions for COVID-19*. (Technical Guidance 2020, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/critical-preparedness-readiness-and-response-actions-for-covid-19).
5. Kraemer MUG, Cummings DAT, Funk S *et al.* Reconstruction and prediction of viral disease epidemics. *Epidemiol Infect* 2018; **147**: 1–7. doi: 10.1017/S0950268818002881.

ORIGINAL ARTICLE

http://doi.org/10.1590/S1678-9946202062030

REVISTA DO INSTITUTO DE MEDICINA TROPICAL DE SÃO PAULO

JOURNAL OF THE SÃO PAULO INSTITUTE OF TROPICAL MEDICINE

Importation and early local transmission of COVID-19 in Brazil, 2020

Jaqueline Goes de Jesus[1*], Claudio Sacchi[2*], Darlan da Silva Candido[3*], Ingra Morales Claro[1], Flávia Cristina Silva Sales[1], Erika Regina Manuli[1], Daniela Bernardes Borges da Silva[4], Terezinha Maria de Paiva[4], Margarete Aparecida Benega Pinho[4], Katia Correa de Oliveira Santos[4], Sarah Catherine Hill[3], Renato Santana Aguiar[5], Filipe Romero[5], Fabiana Cristina Pereira dos Santos[4], Claudia Regina Gonçalves[2], Maria do Carmo Timenetsky[4], Joshua Quick[6], Julio Henrique Rosa Croda[7,8,9], Wanderson de Oliveira[7], Andrew Rambaut[10], Oliver G. Pybus[3], Nicholas J. Loman[6], Ester Cerdeira Sabino[1*], Nuno Rodrigues Faria[1,3,11*]

[1]Universidade de São Paulo, Instituto de Medicina Tropical de São Paulo, São Paulo, São Paulo, Brazil

[2]Instituto Adolfo Lutz, Laboratório Estratégico, São Paulo, São Paulo, Brazil

[3]University of Oxford, Department of Zoology, Oxford, United Kingdom

[4]Instituto Adolfo Lutz, Centro de Virologia, São Paulo, São Paulo, Brazil

[5]Universidade Federal do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil

[6]University of Birmingham, Birmingham, United Kingdom

[7]Ministério da Saúde, Secretaria de Vigilância em Saúde, Coordenação Geral de Laboratórios de Saúde Pública, Brasília, DF, Brazil

[8]Universidade Federal da Grande Dourados, Laboratório de Pesquisa em Ciências da Saúde, Dourados, Mato Grosso do Sul, Brazil

[9]Fundação Osvaldo Cruz Campo Grande, Mato Grosso do Sul, Brazil

[10]University of Edinburgh, Institute of Evolutionary Biology, Edinburgh, United Kingdom

[11]Imperial College, School of Public Health, Department of Infectious Disease Epidemiology, London, United Kingdom

*This authors contributed equally

Correspondence to: Nuno Rodrigues Faria
University of Oxford, Department of Zoology, South Parks Rd, Oxford OX1 3SY, United Kingdom

E-mail: nuno.faria@zoo.ox.ac.uk

Received: 21 April 2020

Accepted: 22 April 2020

ABSTRACT

We conducted the genome sequencing and analysis of the first confirmed COVID-19 infections in Brazil. Rapid sequencing coupled with phylogenetic analyses in the context of travel history corroborate multiple independent importations from Italy and local spread during the initial stage of COVID-19 transmission in Brazil.

KEYWORDS: Public health surveillance. COVID-19. SARS-CoV-2.

INTRODUCTION

The severe acute respiratory syndrome 2 (SARS-CoV-2) was first identified in Wuhan, Central China, in early December 2019, and reported to the World Health Organization (WHO) country office in China on December 31, 2019[1]. SARS-CoV-2 infection causes coronavirus-associated acute respiratory disease in humans, a disease named corona virus disease 19 (COVID-19)[2]. COVID-19 is the third documented spill over of a coronavirus from an animal reservoir to humans in the last two decades to have caused a serious public health threat[3].

On January 30, 2020, COVID-19 was declared a Public Health Emergency of International Concern. On March 16, 2020, WHO reported 153,517 confirmed cases across the globe, and 5,735 confirmed deaths in 143 countries, territories or areas. Within Latin America, Brazil is the country with the largest number of confirmed cases. The country has reported 234 cases across 15 federal States; Sao Paulo, Rio de Janeiro and Bahia States have confirmed local transmission[4].

During the early stages of an epidemic disease, molecular surveillance can inform on the tracking and control of the virus spread across the global and at local scales. Moreover, viral genomes can help to design effective molecular diagnostics, improve vaccine design and complement the contact tracking[5,6]. However, the resolution of the transmission networks reconstructed from genetic data will depend on the rate at which genetic changes accumulate across viral genomes. Within outbreaks, short timescales mean that not all the observed changes will become fixed at the population level[7]. To investigate the early transmission dynamics of imported and local cases in

Brazil, we set up a genomic observatory in Sao Paulo where we sequenced and analysed two complete SARS-CoV-2 genomes in less than 48 h after the cases confirmation. Here we investigate the transmission patterns from phylogenetic analysis of the earliest six SARS-CoV-2 cases in Brazil.

MATERIALS AND METHODS

Samples from suspected SARS-CoV-2 cases underwent confirmatory diagnostic real-time RT-PCR testing[8] at the Instituto Adolfo Lutz (IAL), the regional reference laboratory for SARS-CoV-2 detection in Sao Paulo State, Southeast Brazil. Samples obtained from the Reference Centre for Arbovirus of Sao Paulo, Adolfo Lutz Institute (IAL) have been processed in agreement with routine surveillance activities from the Brazilian Ministry of Health.

We used the open COVID-19 sequencing available and the bioinformatics protocols developed by the ARTIC network. Sequencing protocols, multiplex PCR primers, and bioinformatic pipelines are described in detail at https://artic.network/ncov-2019. In brief, cDNA synthesis was conducted in duplicate for each sample and the concentration of PCR products was measured using a Qubit dsDNA High Sensitivity kit on a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, USA). Library preparation was conducted without a barcoding step and libraries were sequenced on an R9.4.1 flow cell using MinKNOW version 19.10.1 (Oxford Nanopore Technologies, Oxford, UK) for over 12 h. The open-source software RAMPART version 10.5 was used to assign and map reads in real-time. Raw files were base-called with Guppy, demultiplexed and trimmed with Porechop (https://github.com/rrwick/Porechop) and mapped against reference sequence Wuhan-Hu-1 (GenBank Accession Number MN908947). Variants were called using nanopolish 0.11.3. Low coverage regions were masked with N characters. Coverage for the SPBR1 and SPBR2 was 96.9 and 99.6%, with 552730 and 3461754 mapped reads, respectively (Table S1). Raw read data for SPBR1 is available for inspection from https://cadde.s3.climb.ac.uk/covid-19/BR1.sorted.bam.

We investigated the transmission dynamics of the early COVID-19 cases in Brazil (SPBR1 to SPBR6) by analysing genetic changes among the early genomes from Brazil belonging to imported and local cases, and by estimating a maximum likelihood phylogenetic tree together with a set of global reference sequences. We added the SBPBR1 and SPBR2 consensus sequences from Sao Paulo to a curated dataset of complete genomes available from GISAID that included four additional sequences from Brazil (available on GISAID of 15th March 2020). A multiple sequence alignment comprising 347 complete genomes from several countries was generated using MAFFT[9] and manually edited. A maximum likelihood (ML) phylogenetic tree was estimated using PhyML version 3.0[10] using a Hasegawa-Kishino-Yano nucleotide substitution model with a gamma-distributed rate variation across sites.

RESULTS

Four of the six patients self-reported travelling from European countries to Sao Paulo city (SPBR1 to SPBR4) (Table S1). Two patients (SPBR5 and SPBR6) reported direct contact with SPBR1 and no travel outside Brazil. Patient SPBR1 (60-65 year old male) self-reported arriving from Italy on the February 21, 2020he started symptoms on the February 24 and tested positive two days later. Patients SPBR5 and SPBR6 were in direct contact with patient SPBR1 on February 22, and tested positive on February 29, 2020. Figure 1 shows the clade containing Brazilian sequences along with location of infection (squares) and reporting (circles). Our analyses show that the SPBR1 genome is identical to the SPBR5 and SPBR6 contacts (illustrated by zero branch lengths in Figure 1; detailed tree with annotated tips and travel history information for the clade containing Brazilian sequences can be found in the Figure S1).

We found that the SPBR1, SPBR5 and SPBR6 are identical to several other genomes circulating in Italy and elsewhere collected between February 20 and March 2, 2020 (Figure 1). The lack of changes among SARS-CoV-2 genomes collected during this period is not surprising given the evolutionary rate of the virus that results in an average of 1 to 2 mutations per month[11]. These data highlight the critical importance of contextualizing phylogenetic information with travel history when investigating early transmission dynamics of SARS-CoV-2. As no epidemiological information was available for SPBR5 and SPBR6, one could not exclude an alternative scenario based on sequencing data alone that would suggest additional independent introductions from Italy or elsewhere.

Patients SPBR2, SPBR3 and SPBR4 all reported travelling to Italy, where the of incidence of COVID-19 has been the highest outside Wuhan in China[12]. Consistent with the travel history, sequences from these patients are found interspersed in the tree in agreement with the multiple independent introductions of SARS-CoV-2 to Sao Paulo from Italy. This finding highlights the key role of human mobility in the early stages of the pandemic and is in line with a recent analysis on the risk of importation of COVID-19 based on the history of air traveling data and the incidence data[13]. Given that the air traveling to Brazil from Italy has reduced, it is possible that the proportion

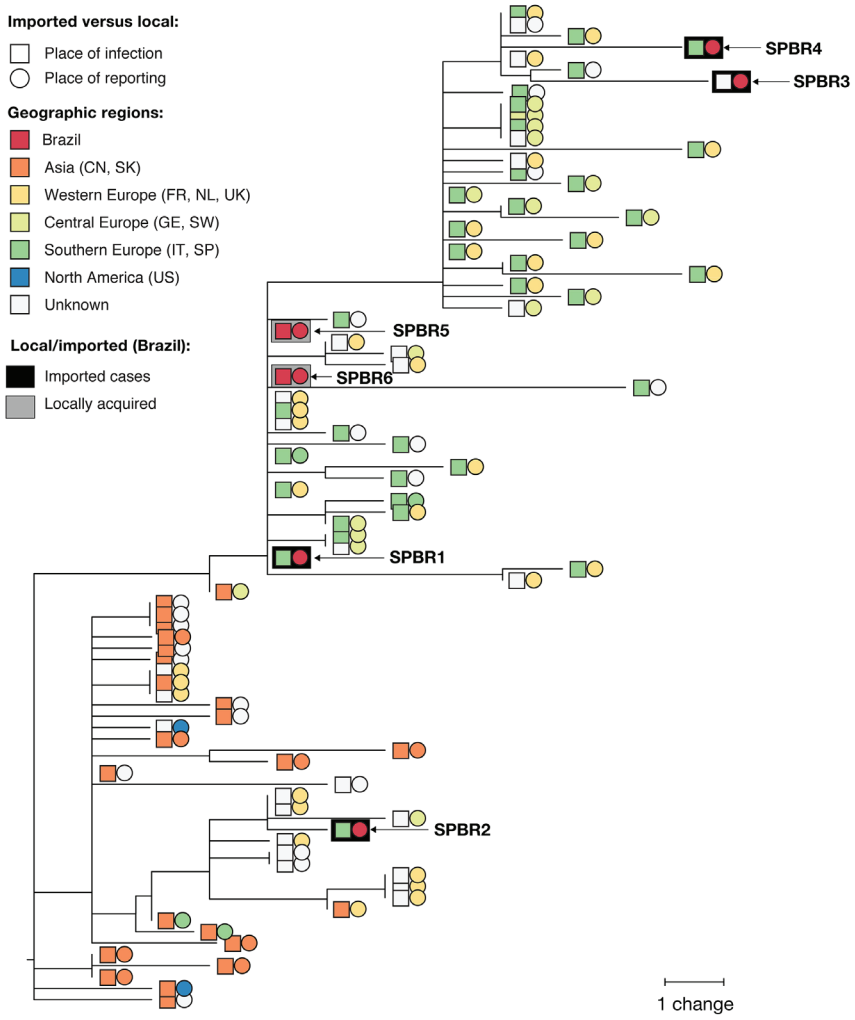


Figure 1 - Maximum likelihood phylogeny (n=88) including Brazilian SARS-CoV-2 genomes from the first confirmed cases in Brazil. Squares and circles are coloured according to the place of infection and the place of reporting, respectively. Local cases are highlighted with a grey background, imported cases are highlighted with a black background. A full tree (n=347) can be found in the Supplementary Material (Figure S1).

of SARS-CoV-2 imported cases from other countries, particularly the USA, may increase[13].

DISCUSSION

Our study provides a snapshot of the early establishment of the COVID-19 pandemic in Brazil, characterized by multiple independent introductions from Italy, followed by local transmission of the virus in Sao Paulo. Phylogenetic analyses are broadly consistent with the patients' self-reported traveling histories. We show that the two genomes associated with local transmission are linked to a patient infected in Italy and are identical to other Italian genomes collected in the same time window. Given the within-outbreak rate of evolutionary change estimated for SARS-CoV-2[11], we caution against inferring directionality of transmission based on genetic data alone. Such inferences can further be overshadowed by incomplete sampling due to delays, reflecting the lack of equitable access to diagnosis and genomic sequencing.

CONCLUSION

Given the findings of the present study, we conclude that phylogenetic data from the pandemic needs to be contextualized with appropriate metadata, including basic demographics, symptoms onset date, the sample collection date, the country of reporting and the self-reported travel history. Joint epidemiological and genomic surveillance of COVID-19 cases will be critical to rapidly identify possible clusters of local transmission in Brazil and in other countries, and to better understand and help mitigating the transmission in the community.

ACKNOWLEDGMENTS

We thank GISAID database for supporting rapid and open access real-time SARS-CoV-2 genomic data sharing. We would like to thank the authors who generated and shared data with the research community and all the patients involved, the staff and research groups who assisted with patient care, sample collection, genome sequencing and data sharing. We thank Josh Quick, Lucy Matkin and Julien Thezé for support.

AUTHORS' CONTRIBUTIONS

Conceptualization, Jaqueline de Jesus, Ester Sabino and Nuno Rodrigues Faria; data curation, Darlan da Silva Candido, Oliver G. Pybus and Nuno Rodrigues Faria; formal analysis, Darlan da Silva Candido, Sarah Catherine Hill, Filipe Romero, Andrew Rambaut, Nicholas James Loman and Nuno Rodrigues Faria; funding acquisition, Andrew Rambaut, Nicholas James Loman, Ester Cerdeira Sabino, Nuno Rodrigues Faria; investigation, Jaqueline de Jesus, Claudio Sacchi, Ingra Morales Claro, Flávia Cristina Silva Sales, Erika Regina Manuli, Daniela Bernardes Borges da Silva, Terezinha Maria de Paiva, Margarete Aparecida Benega, Katia Correa de Oliveira Santos, Renato Santana Aguiar, Fabiana Cristina Pereira Santos and Claudia Regina Gonçalves; methodology, Jaqueline de Jesus, Claudio Sacchi, Ingra Morales Claro, Fabiana Cristina Pereira dos Santos, Joshua Quick and Nicholas James Loman; project administration, Andrew Rambaut, Oliver G. Pybus, Nicholas James Loman, Ester Sabino and Nuno Rodrigues Faria; resources, Julio Henrique Rosa Croda and Wanderson de Oliveira, Oliver G Pybus, Andrew Rambaut, Ester Sabino and Nuno Rodrigues Faria; software, Filipe Romero, Andrew Rambaut, Nicholas J Loman; supervision, Maria do Carmo Timenetsky, Oliver G. Pybus, Ester Sabino and Nuno Rodrigues Faria; validation, Joshua Quick, Andrew Rambaut; writing – original draft, Jaqueline de Jesus, Darlan da Silva Candido and Nuno Rodrigues Faria; writing – review & editing, Andrew Rambaut, Oliver G. Pybus, Nicholas James Loman and Ester Sabino.

FUNDING

This work was supported by a Medical Research Council and FAPESP CADDE partnership award (MR/S0195/1), FAPESP grant Nº 2018/14389-0, and a John Fell Research Fund (grant Nº 005166). NRF is supported by a Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z). DDSC is supported by the Clarendon Fund and by the Oxford University Zoology Department.

REFERENCES

1. World Health Organization. Critical preparedness, readiness and response actions for COVID-19. [cited 2020 Apr 22]. Available from: https://www.who.int/publications-detail/critical-preparedness-readiness-and-response-actions-for-covid-19
2. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020 In Press.
3. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26:450-2.
4. Brasil. Ministério da Saúde. Notificação de casos de doença pelo coronavírus 2019 (COVID-19). [cited 2020 March 15]; Available from: https://www.saude.go.gov.br/component/content/article/34-page/9895-coronav%C3%ADrus.html?highlight=WyJjb3JvbmF2XHUwMGVkcnVzIl0=&Itemid=101
5. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SL, Andersen KG, et al. Ebola virus epidemiology, transmission, and evolution during seven nonths in Sierra Leone. Cell. 2015;161:1516-26.
6. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345:1369-72.
7. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: insights from the 2013-2016 epidemic. Nature. 2016;538:193-200.
8. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. 2020;25:2000045.
9. Katoh K, Standley DM. MAFFT: iterative refinement and additional methods. In: Russel DJ, editor. Multiple sequence alignment methods. New York: Humana; 2014. p.131-46.
10. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307-21.
11. Rambaut A. Phylogenetic analysis of nCoV-2019 genomes. [cited 2020 Apr 22]. Available from: http://virological.org/t/phylogenetic-analysis-of-23-ncov-2019-genomes-2020-01-23/335
12. World Health Organization. Coronavirus disease (COVID-2019) situation reports. [cited 2020 Apr 22]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
13. Candido DS, Watts A, Abade L, Kraemer MU, Pybus OG, Croda J, et al. Routes for COVID-19 importation in Brazil. J Travel Med. 2020 In press.

SUPPLEMENTARY MATERIAL

Table S1 - Sequencing statistics for the Brazilian SARS-COV-2 genomes from this study.

Isolate	Mapped Reads	Average depth coverage	Bases covered >10x	Bases covered >25x	Reference covered (%)
SPBR1	552730	3622.14	29426	29106	96.8966
SPBR2	3461754	5117.28	29849	29845	99.5954

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu™ Database on which this research is based. The list is detailed on Supplementary Table 2, which is available on GitHub (https://github.com/CADDE-CENTRE/REPOSITORY/blob/master/First%20genomes%20from%20Americas.docx). All submitters of data may be contacted directly via www.gisaid.org.

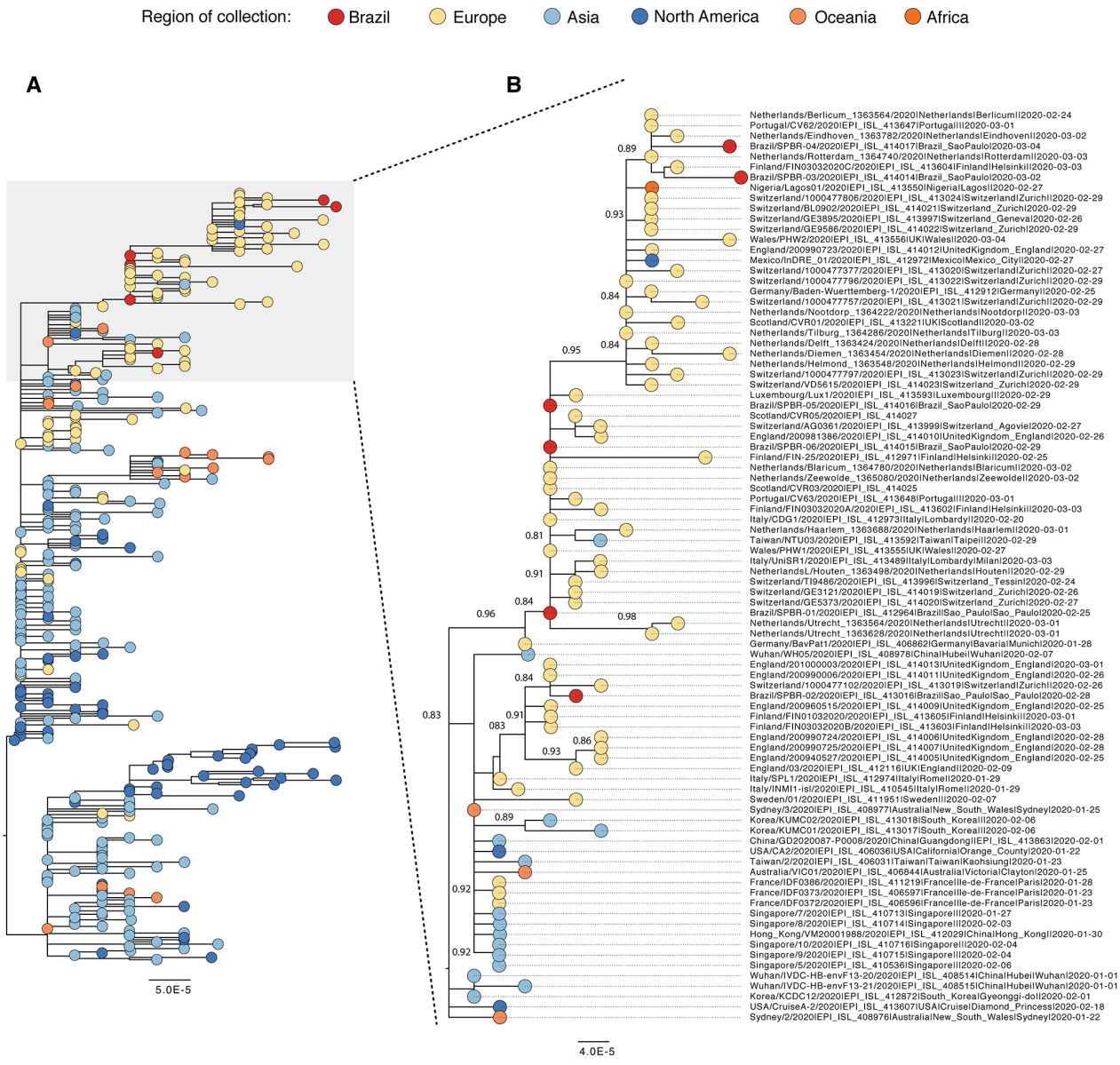


Figure S1 - SARS-CoV-2 phylogeny of the first confirmed Brazilian cases. A) Global SARS-CoV-2 Maximum Likelihood phylogeny including the six first genomes from Brazil. The tree was estimated using all available sequences in GISAID database (n=347) as of the 10th of March 2020. Tips are coloured according to the location of collection. B) Expansion of the clades containing the six Brazilian SARS-CoV-2 genomes (n=88). Tips are coloured according to the location of collection and only approximate likelihood ratio node supports > 0.80 are shown. Tips were labelled according to sequence name, GISAID accession number, place and date of collection.

ARTICLES
https://doi.org/10.1038/s41562-020-0928-4

nature human behaviour

Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil

William Marciel de Souza[1,29], Lewis Fletcher Buss[2,29], Darlan da Silva Candido[2,3,29], Jean-Paul Carrera[3,4,29], Sabrina Li[5,29], Alexander E. Zarebski[3], Rafael Henrique Moraes Pereira[6], Carlos A. Prete Jr[7], Andreza Aruska de Souza-Santos[8], Kris V. Parag[9], Maria Carolina T. D. Belotti[7], Maria F. Vincenti-Gonzalez[10], Janey Messina[5,11], Flavia Cristina da Silva Sales[2], Pamela dos Santos Andrade[2], Vítor Heloiz Nascimento[7], Fabio Ghilardi[2], Leandro Abade[3], Bernardo Gutierrez[3,12], Moritz U. G. Kraemer[3,13,14], Carlos K. V. Braga[6], Renato Santana Aguiar[15], Neal Alexander[16], Philippe Mayaud[17], Oliver J. Brady[15,18], Izabel Marcilio[19], Nelson Gouveia[20], Guangdi Li[21], Adriana Tami[10], Silvano Barbosa de Oliveira[22], Victor Bertollo Gomes Porto[22], Fabiana Ganem[22], Walquiria Aparecida Ferreira de Almeida[22], Francieli Fontana Sutile Tardetti Fantinato[22], Eduardo Marques Macário[23], Wanderson Kleber de Oliveira[23], Mauricio L. Nogueira[24], Oliver G. Pybus[3], Chieh-Hsi Wu[25,29], Julio Croda[23,26,27,28,29] ✉, Ester C. Sabino[2,29] and Nuno Rodrigues Faria[2,3,9,29] ✉

The first case of COVID-19 was detected in Brazil on 25 February 2020. We report and contextualize epidemiological, demographic and clinical findings for COVID-19 cases during the first 3 months of the epidemic. By 31 May 2020, 514,200 COVID-19 cases, including 29,314 deaths, had been reported in 75.3% (4,196 of 5,570) of municipalities across all five administrative regions of Brazil. The R_0 value for Brazil was estimated at 3.1 (95% Bayesian credible interval = 2.4–5.5), with a higher median but overlapping credible intervals compared with some other seriously affected countries. A positive association between higher per-capita income and COVID-19 diagnosis was identified. Furthermore, the severe acute respiratory infection cases with unknown aetiology were associated with lower per-capita income. Co-circulation of six respiratory viruses was detected but at very low levels. These findings provide a comprehensive description of the ongoing COVID-19 epidemic in Brazil and may help to guide subsequent measures to control virus transmission.

COVID-19 is a severe acute respiratory infection (SARI) that emerged in early December 2019 in Wuhan, China[1]. The outbreak was declared a public health emergency of international concern by the World Health Organization on 30 January 2020. COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), an enveloped, single-stranded positive-sense RNA virus that belongs to the *Betacoronavirus* genus and Coronaviridae family[2]. SARS-CoV-2 is closely related genetically to bat-derived SARS-like coronaviruses[3]. Human-to-human transmission occurs primarily via respiratory droplets and direct contact, similar to human influenza viruses, SARS-CoV and Middle East respiratory syndrome coronavirus[4]. The most commonly reported clinical symptoms are fever, dry cough, fatigue, dyspnoea, anosmia, ageusia, or some combination of these[1,4,5]. As of 16 June 2020, more than 7.9 million cases have been confirmed worldwide, resulting in 434,796 deaths[6].

Brazil declared COVID-19 a national public health emergency on 3 February 2020[7]. After the development of a national emergency plan and the early establishment of molecular diagnostic facilities across Brazil's network of public health laboratories, the country reported its first confirmed COVID-19 case on 25 February 2020, in a traveller returning to São Paulo from northern Italy[8]. São Paulo is the largest city in South America and no other Brazilian city receives a greater proportion of international flights[9]. Currently, Brazil has one of the fastest-growing COVID-19 epidemics in the world, now accounting for 1,864,681 cases and 72,100 deaths, comprising over 55% of the total number of reported cases in Latin America and the Caribbean (as of 14 July 2020)[6]. About 21% of Latin American and Caribbean populations are estimated to be at risk of severe COVID-19 illness[10]. The region has been experiencing large outbreaks, with growing epidemics in Brazil, Peru, Mexico, Chile, Colombia, Panama and possibly Venezuela and Nicaragua, amid growing concerns about testing capacity for COVID-19 (refs. [11–14]). Preparedness for laboratory surveillance of SARS-CoV-2 in Latin America is centred around a network of national reference influenza surveillance laboratories that is facing several challenges, including a shortage of reagents and equipment[15].

Conscious of the challenges associated with surveillance since the beginning of the epidemic in Brazil, here we focus on two main objectives. First, we contextualize the Brazilian SARS-CoV-2

A full list of affiliations appears at the end of the paper.

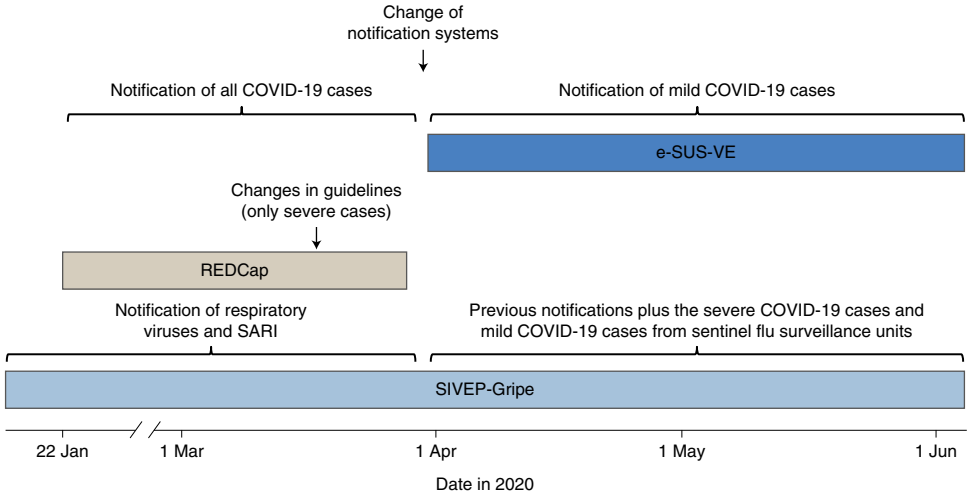


Fig. 1 | Timeline of national COVID-19 reporting systems in Brazil. The REDCap system operated between late January and 25 March 2020. Aggregated numbers from e-SUS-VE and SIVEP-Gripe data for mild and hospitalized COVID-19 cases, respectively, are updated on a daily basis on the Portal do COVID-19 website (https://covid.saude.gov.br/).

epidemic by comparing local transmission dynamics with those observed in other selected countries. Second, we use geospatial data related to confirmed COVID-19 cases and SARI cases with unknown aetiology to evaluate the relationship between socioeconomic factors and COVID-19 distribution.

Results

Contextualizing COVID-19 data reporting systems in Brazil. On 22 January 2020—more than 1 month before the first case in Brazil—the Brazilian Ministry of Health implemented the REDCap platform to report prospective suspected, probable and confirmed COVID-19 cases (see Methods for case definitions), as part of an early response to the pandemic[16]. By 27 March 2020, the REDCap system was discontinued (Fig. 1). Since then, mild COVID-19 cases started to be reported on e-SUS Vigilância Epidemiológica (e-SUS-VE), a new national COVID-19 reporting system, and hospitalized COVID-19 cases started to be recorded on a pre-existing Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe) system. The SIVEP-Gripe system has been in use since 2009 (having been implemented in response to the 2009 influenza H1N1 pandemic) and has since centralized the reporting of respiratory viruses and SARI for the Brazilian Ministry of Health (Fig. 1). Both e-SUS-VE and SIVEP-Gripe include suspected and confirmed COVID-19 cases as reported by public health and private services (primary and emergency care). These two reporting systems (e-SUS-VE and SIVEP-Gripe) are inter-related on the Portal do COVID-19 website (https://covid.saude.gov.br/), which summarizes daily the aggregated counts from both platforms.

SARS-CoV-2 reporting in Brazil. We analysed a total of 514,200 SARS-CoV-2 cases from the Portal do COVID-19 website (SIVEP-Gripe and e-SUS-VE databases combined) that were confirmed by molecular diagnostic and clinical epidemiological criteria by 31 May 2020 (see Methods). Cases were reported in 75.3% (4,196 of 5,570) of municipalities across all five administrative regions of Brazil and included 206,555 (40.2%) recovered patients and 29,314 fatal (17.5%) COVID-19 cases (Fig. 2a). We further analysed a total of 1,468 confirmed cases from the REDCap system, including 342 imported cases with associated travel history information. After excluding individuals who travelled to multiple countries before entering Brazil ($n=56$) and who had an unknown country of origin ($n=16$), the self-reported countries of infection for cases acquired abroad until 19 March 2020 were the United States (28.6%; $n=76$), Italy (24.4%; $n=65$), the United Kingdom (10.5%; $n=28$) and Spain (8.3%; $n=22$) (Extended Data Fig. 1). The first reported case (SPBR1) was reported on 25 February 2020 in the municipality of São Paulo, the fourth most populous urban area worldwide. Following the first reports of COVID-19 in Brazil's largest population centres, SARS-CoV-2 subsequently spread to municipalities with smaller population sizes (Fig. 2b). Until 31 May 2020, most confirmed cases and deaths were reported in the states of São Paulo (109,698 cases and 7,615 deaths), Rio de Janeiro (53,388 cases and 5,344 deaths), Ceará (48,489 cases and 3,010 deaths) and Amazonas (41,378 cases and 2,052 deaths), which together account for 49.2% of all cases and 61.5% of deaths in Brazil (Fig. 2c).

Basic reproduction number of SARS-CoV-2 in Brazil and comparison countries. To estimate the basic reproduction number (R_0) of SARS-CoV-2 in Brazil, daily confirmed cases in São Paulo, Rio de Janeiro, Ceará and Amazonas states were compiled from Ministry of Health data (for specification of the time windows used in the analyses, see Extended Data Fig. 2). For comparison, we compiled time series of confirmed cases in several European countries from the Johns Hopkins Coronavirus Resource Center (https://coronavirus.jhu.edu/; see also Extended Data Fig. 3). We found that São Paulo, Rio de Janeiro and Amazonas were characterized by similar R_0 values of 2.9 (95% Bayesian credible interval (BCI) = 2.2–5.1), 2.9 (95% BCI = 2.2–4.9) and 2.6 (95% BCI = 2.0–4.5), respectively. However, for Ceará, the estimated R_0 was considerably lower at 1.9 (95% BCI = 1.5–3.0) (Fig. 3 and Extended Data Fig. 1). This finding could be a result of the small window between the first reported cases and the early implementation of non-pharmaceutical interventions (NPIs) in this state (Supplementary Table 1 and Extended Data Fig. 2). On a national scale, the estimated R_0 for Brazil was slightly higher than that of the Brazilian states considered in this study, with a median of 3.1 (95% BCI = 2.4–5.5), and also slightly higher than R_0 values estimated for other severely affected countries: Spain (2.6; 95% BCI = 2.0–4.6); France (2.5; 95% BCI = 1.9–4.4); the United Kingdom (2.6; 95% BCI = 2.0–5.1); and Italy (2.5; 95% BCI = 2.0–4.4) (Fig. 3). While the incidence curves for European countries have consistently flattened and declined since the implementation of NPIs (suggesting that the R_0 value has fallen below 1), Brazil's daily incidence curve has continued to increase (Fig. 2a and Extended Data Fig. 4).

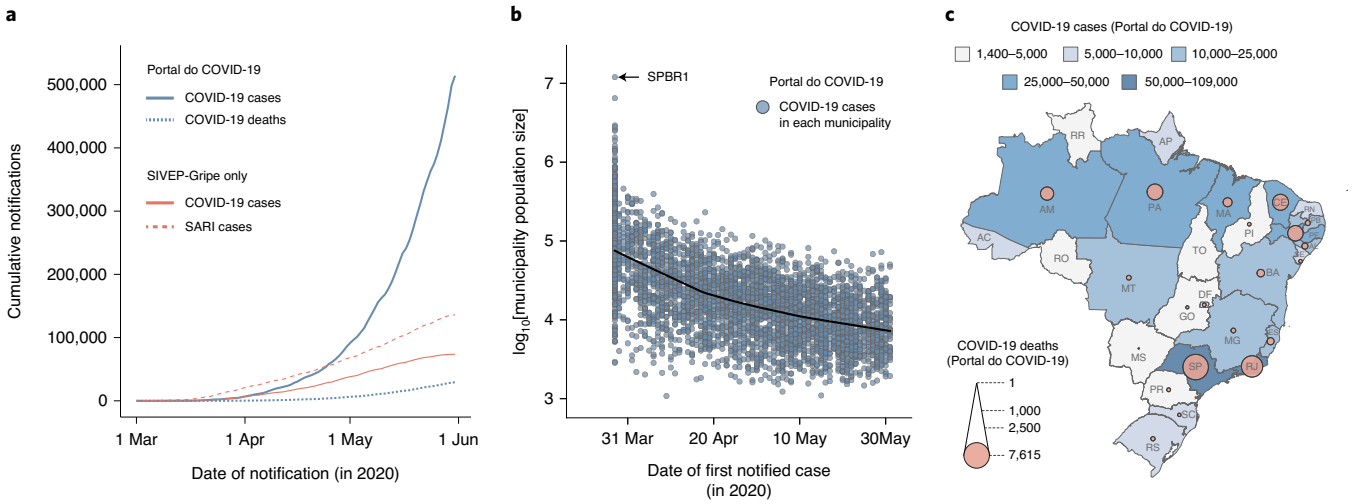


Fig. 2 | COVID-19 epidemiology in Brazil. a, Numbers of COVID-19 cases (blue solid line) and deaths (blue dashed line) reported to the Ministry of Health (Portal do COVID-19 website), along with numbers of COVID-19 confirmed cases (salmon solid line) and cases of SARI with unknown aetiology (salmon dashed line) reported to the SIVEP-Gripe database. **b**, First COVID-19 cases by date and Brazilian municipal population size based on the Ministry of Health data, from 28 March 2020. Each circle represents the first confirmed COVID-19 case in the municipality ($n = 4{,}196$ Brazilian municipalities). SPBR1 is the first detected SARS-CoV-2 infection in Brazil[8]. **c**, Map coloured according to the number of confirmed COVID-19 cases per state reported to the Ministry of Health (Portal do COVID-19 website). Circle sizes are proportional to the number of reported COVID-19 deaths in each federal unit. AC, Acre; AL, Alagoas; AM, Amazonas; AP, Amapá; BA, Bahia; CE, Ceará; DF, Distrito Federal; ES, Espírito Santo; GO, Goiás; MA, Maranhão; MG, Minas Gerais; MS, Mato Grosso do Sul; MT, Mato Grosso; PA, Pará; PB, Paraíba; PE, Pernambuco; PI, Piauí; PR, Paraná; RJ, Rio de Janeiro; RN, Rio Grande do Norte; RO, Rondônia; RR, Roraima; RS, Rio Grande do Sul; SC, Santa Catarina; SE, Sergipe; SP, São Paulo; TO, Tocantins.

SARIs mostly reflect COVID-19 cases. In the early phase of the COVID-19 epidemic in Brazil, we analysed the results for other respiratory pathogens tested in Brazil as part of a differential diagnosis by the Central Public Health Laboratories and National Influenza Centres (Brazilian Ministry of Health), obtained from a REDCap platform[17] designed for COVID-19. The respiratory viruses most frequently identified between 7 January 2020 and 27 March 2020, in patients with a suspected but negative diagnosis of COVID-19, were influenza A virus (347 (14.3%) of 2,429 tested cases), influenza B virus (251 (10.3%) of 2,429) and human rhinovirus (136 (5.6%) of 2,429). We found co-detection of SARS-CoV-2 with six other respiratory viruses, the most frequent of which were influenza A (11 (0.5%) of 2,429) and human rhinovirus (6 (0.2%) of 2,429) (Extended Data Fig. 7).

The SIVEP-Gripe system started reporting hospitalized COVID-19 cases in early March 2020 (epidemiological week 10) (Fig. 4). In this system, the number of tested cases is unavailable. We found that the peak of influenza confirmed cases ($n = 447$) occurred at epidemiological week 12 (15–21 March 2020). During the same week 12, we detected an 8.5-fold increase in total cases attributed to SARS-CoV-2 ($n = 3{,}789$) and a 9.9-fold increase in total cases reported as SARI with unknown aetiology ($n = 4{,}424$) (Fig. 4). From 2 January to 31 May 2020, a total of 2,136 influenza cases and 272 cases caused by other respiratory pathogens, including human respiratory syncytial virus, human rhinovirus, adenovirus and metapneumovirus, were reported in the SIVEP-Gripe database. The low observed incidence of influenza and other respiratory viruses may have been influenced by limited testing for these viruses during this period. Although NPIs may have an impact in reducing influenza virus transmission, this does not necessarily reflect a lower co-circulation of other respiratory viruses[18].

Socioeconomic differences are associated with COVID-19 diagnosis. Until 31 May 2020, a total of 73,648 COVID-19 confirmed cases and 168,001 SARI cases with unknown aetiology were reported in the SIVEP-Gripe system. We hypothesized that the 2.3-fold increase of SARI cases with unknown aetiology was associated with differential access to healthcare due to socioeconomic factors.

We focused on the Metropolitan Region of São Paulo (MRSP), which has a population of 23 million inhabitants across six sub-regions (Central, West, North, East, Southeast and Southwest) and 39 municipalities (Fig. 5a). To test this hypothesis, we obtained per-capita income at the census tract level (typically 150–300 households) in the MRSP, based on the residential address of each case. We then linked this information to each patient's final diagnosis outcome: confirmed case of COVID-19 or SARI with unknown aetiology. While the income distribution of SARI cases with unknown aetiology was similar to that of all residents of the MRSP over the whole period (Fig. 5b), we observed that the income distribution of individuals with COVID-19 confirmed by laboratory and clinical criteria was initially higher than that of all MRSP residents and decreased over time towards similar levels by epidemiological week 21 (Fig. 5b). Importantly, we found that the log odds of one or more confirmed COVID-19 cases per census tract increased with per-capita income in epidemiological weeks 12 and 22 (likelihood ratio test P value < 0.001; Fig. 5b and Supplementary Table 2). This provides statistical evidence of an association between confirmed COVID-19 diagnosis and per-capita income, suggesting a socioeconomic difference in access to COVID-19 diagnosis in the MRSP. For reference, we also provide a map of per-capita income (Fig. 5a) and population density in each census tract (Extended Data Fig. 8).

We conducted a geospatial analysis to understand the distribution of relative risk of observing a COVID-19 case or SARI case with unknown aetiology in the MRSP, using a Bayesian method and adjusted for spatial and non-spatial effects as defined by the Besag–York–Mollié model[19] (Fig. 5). Our estimates show an increase in the relative risk of COVID-19 diagnosis in higher-income census tracts between epidemiological weeks 12 and 21, especially in the central region of the MRSP (Fig. 5a,c). We observed a similar trend in the relative risk of SARI cases with unknown aetiology among residents of the central region. However, there was also an increased probability

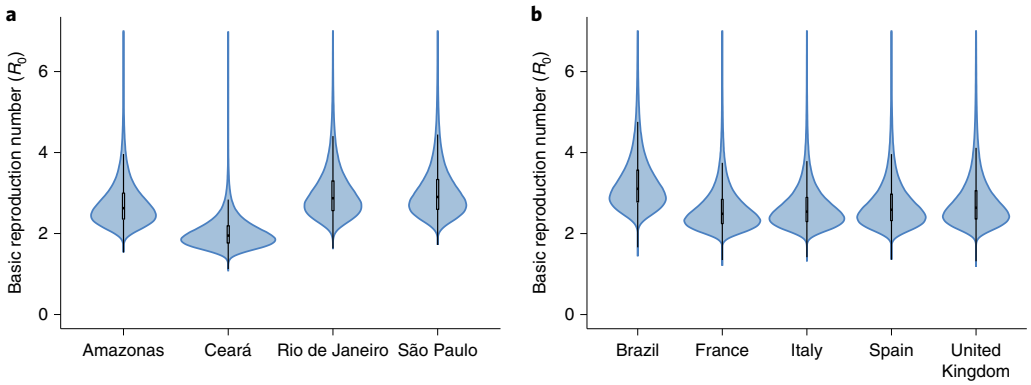


Fig. 3 | Estimated R_0 values for four Brazilian states and selected countries. Left: R_0 values for the Amazonas, Ceará, Rio de Janeiro and São Paulo states. Right: R_0 for Brazil, France, Italy, Spain and the United Kingdom. Violin plots of posterior samples for the basic reproduction number, the box plots show the median, first, and third quartiles. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. The daily numbers of infections used in each analysis can be found in Extended Data Figs. 3 and 4. Daily numbers of infections and prior distributions can be found in Extended Data Figs. 5 and 6.

of SARI cases with unknown aetiology in the southwest, west, north and south sub-regions, where income per capita is typically lower. Overall, the relative risk of SARI cases with unknown aetiology is more spatially widespread in the MRSP than that of confirmed COVID-19 cases (Fig. 5c).

The relative risk of SARI cases with unknown aetiology compared with confirmed COVID-19 cases in the central region of the MRSP decreased through time, probably as a response to several NPIs implemented throughout the state of São Paulo (see Supplementary Table 1). By week 16 (1 month after the start of the NPIs in São Paulo), we detected an increased risk particularly of SARI cases with unknown aetiology outside the central region of the MRSP, especially in the southwest region. SARI cases with unknown aetiology risk were also high in the east region. By week 21, the risk remained high throughout the central region and the risk of SARI cases with unknown aetiology decreased in the east region, possibly as a result of interventions targeting the reduction of SARS-CoV-2 transmission.

Demographics and characteristics of COVID-19 hospitalized and fatal cases in Brazil. Analysis of the age–sex structure of 67,180 confirmed COVID-19 cases reported on the SIVEP-Gripe system revealed a high proportion (44,027 (65.5%) of 67,180) of confirmed COVID-19 infections in middle- or older-aged individuals (≥50 years of age) and a lower proportion (1,454 (2.2%) of 67,180) in younger age groups (≤20 years of age) (Fig. 6a). The median age was 59 years (interquartile range = 44–72). The majority (38,654 (57.5%) of 67,180) were male. Similarly, 59% (14,498 of 24,519) of COVID-19 deaths were in men, and 85% (20,916 of 24,519) were in people aged ≥50 years. A total of 2.95% (1,983 of 67,180) cases were reported as nosocomial transmission, defined as a COVID-19 case acquired after hospitalization. Overall, 116 newborns (≤1 month old), 381 infants (≥1–12 months old), 518 children (≥1–12 years old) and 258 adolescents (≥12–17 years of age) were diagnosed with COVID-19. In addition, 740 patients were pregnant (61 in the first trimester, 172 in the second trimester, 447 in the third trimester and 60 with missing gestational age).

By 31 May 2020, 91% (67,042 of 73,649) of patients with COVID-19 reported in the SIVEP-Gripe system had been hospitalized. Of these, 30.3% (22,332 of 73,649) were admitted to an intensive care unit (ICU). The median length of ICU stay for patients with COVID-19 was 5 d (interquartile range = 2–10 d; range = 0–65 d), based on the ICU admission and discharge dates of 8,240 confirmed cases. Most symptoms reported by patients with COVID-19 were a cough (56,681 (85.2%) of 66,514 without missing data), fever (51,312 (79.6%) of 65,310) and dyspnoea (51,312 (76.6%) of 65,310) (Fig. 6b). These three symptoms comprise part of the case definition of SARI in Brazil. In addition, 68% (40,806 of 60,400) of individuals with COVID-19 were hypoxic (O_2 saturation < 95%), reflecting the overall severity of cases reported on SIVEP-Gripe (as shown in Fig. 1). The most prevalent comorbidities were cardiovascular disease (23,085 (66.5%) of 34,693 without missing data) and diabetes (17,271 (54.5%) of 31,672) (Fig. 6a). Among the patients with COVID-19, older age groups tended to have a higher proportion of comorbidities than younger age groups in different outcomes (Fig. 6c). The proportions of the general Brazilian population with cardiovascular disease and diabetes are 4.2 and 6.2%, respectively[20]. A total of 83.7% (17,921 of 21,414 with complete comorbidity information) of individuals with confirmed COVID-19 had at least one comorbidity (see Supplementary Table 2 for information on data completeness).

Discussion

While the COVID-19 epidemic in Brazil continues to grow, details of its transmission potential and clinical and epidemiological characteristics remains poorly understood. We estimate a higher median transmission potential (R_0) of SARS-CoV-2 of 3.1 (2.4–5.5) in Brazil compared with Italy, the United Kingdom, France, and Spain, which have point estimates of R_0 varying from 2.5–2.6; however, the credible intervals overlap substantially. We have also observed rapid spread of COVID-19 through the country, with more populated and better-connected municipalities being affected earlier, and less populated municipalities being affected at a later stage of the epidemic. In the São Paulo metropolitan region, we found a higher risk of diagnosed COVID-19 cases in census tracts with higher per-capita income during the early phase of the COVID-19 epidemic but also as the weeks progressed. This contrasts with the wider spread of SARI cases among sub-regions with lower per-capita income. Our results provide new insights into the Brazilian COVID-19 epidemic and highlight the high transmission potential of SARS-CoV-2 in the country, the role of its large urban centres and the lack of lockdown and the challenges in reporting and non-equitable access to testing/diagnostics as factors potentially contributing to the rapid and sustained spread of the epidemic in Brazil.

Recent estimates of R_0 at the beginning of the COVID-19 epidemic in Brazil have suggested that an infected individual would infect on average three or four others[21]. The credible intervals of our estimates broadly overlap with these observations and are lower compared with previously published estimates for Brazil[22]. As a

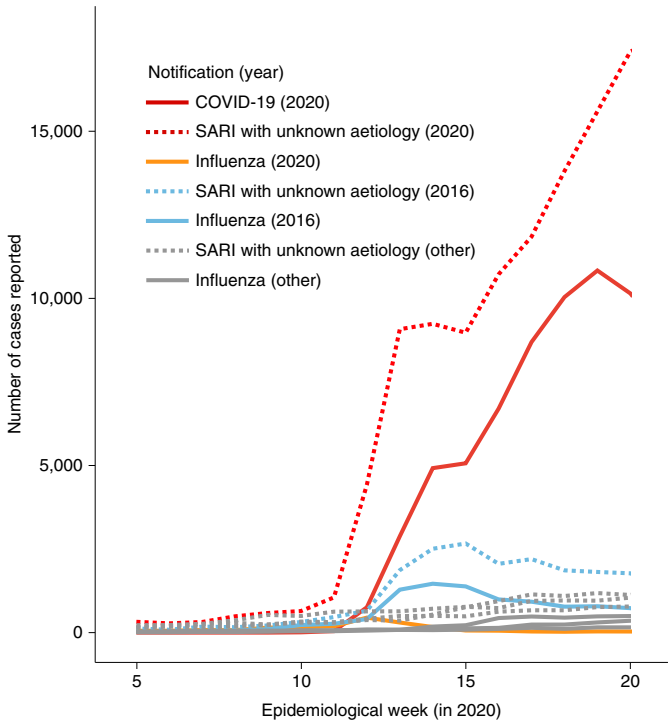


Fig. 4 | Reports of COVID-19 and SARI with unknown aetiology and influenza. The red and orange lines indicate cases reported in 2020 (solid red, COVID-19; solid orange, influenza; dashed red, SARI with unknown aetiology). The blue lines indicate cases reported in 2016 for influenza (solid blue line) and SARI with unknown aetiology (dashed blue line). Grey lines indicate influenza (solid line) and SARI cases with unknown aetiology (dashed line) for 2017, 2018 and 2019 combined.

comparison, the reproduction number in Peru has been estimated at around 2.3 (2.0–2.5)[23]. Since the start of the epidemic in Brazil, several types of NPI have been adopted with varied success by the country's 27 federal units and 5,596 municipalities. Virus transmission seems to have dropped substantially in most affected states[21] and also in the city of São Paulo[24]. However, the estimated reproduction number remains above 1 (refs. [21,24]). Thus, only mitigation (and not suppression) of the epidemic has been achieved so far, which has been linked to substantial excess deaths due to poorer healthcare available[25,26]. Closer surveillance of viral transmission at the local scales and an assessment of the impact of the different control measures on COVID-19 transmission will help to determine an optimal mitigation strategy to minimize infections and reduce healthcare demand in Brazil. Moreover, continued monitoring of the genetic diversity of the virus lineages circulating in Brazil[24] will be important, as recent data suggest that virus diversity may play a role in virus transmissibility[27,28].

We found that 65.5% of reports in the SIVEP-Gripe system, which includes most severe COVID-19 cases, are from patients aged ≥50 years of age. This observation is remarkably similar to current estimates for Latin America[10], where 65% of the individuals ≥50 years of age have been estimated to be at high risk of severe COVID-19, defined as individuals with at least one condition who would require hospitalization if infected. Moreover, we found that 57 and 59% of the severe COVID-19 cases and deaths (respectively) reported in SIVEP-Gripe were male, and that the most frequent comorbidities were cardiovascular disease and diabetes. Overall, 84% of SIVEP-Gripe reports had at least one underlying condition. Of these, 21% (n = 9,471/45,480) were included in the working age bracket (16–65 years of age). Moreover, only 2.6% (n = 1,892/73,673) of the COVID-19 confirmed cases reported in the SIVEP-Gripe system included occupation information. Information on socioeconomic determinants, as well as occupation and race/ethnicity, are critical[29] as this allows the prioritization of control efforts; for example, towards healthcare workers and patients attending hospitals[30] or work settings[31].

Our data uncover a socioeconomic bias in testing and diagnostics in current surveillance guidelines and suggest that the number of reported confirmed case counts may substantially underestimate the number of cases in the general population, particularly in regions of lower socioeconomic status. Socioeconomic differences are associated with access to healthcare[32] and should be taken into account when designing targeted interventions. We found that the proportion of SARI cases with unknown aetiology versus confirmed COVID-19 cases has increased across the entire country (as of 15 June 2020, the number of reported SARI cases with unknown aetiology was nearly twofold greater than the number of confirmed COVID-19 cases). Based on clinical and epidemiological grounds, it is likely that many SARI cases with unknown aetiology are caused by SARS-CoV-2. In order to rigorously establish the contribution of non-SARS-CoV-2 infections to the SARI cases, we would need additional denominator data to understand the level of testing for these viruses (that is, the negative test results). Our findings with regards to socioeconomic bias are likely to apply to other states and regions of Brazil and highlight the importance of scaling up surveillance and laboratory capacity within Latin America. Indeed, the largest Brazilian serosurvey conducted to date suggests that undetected cases may be seven times higher than reported cases[33].

We further show that SARI cases with unknown aetiology are associated with lower socioeconomic status in the MRSP. The socioeconomic disparities observed here were particularly evident at the beginning of the outbreak (Fig. 5b). This can be explained in part by: (1) the high proportion of early cases in returning travellers with higher income and better access to private laboratories for diagnostics; and (2) the more limited access to freely available diagnostic screening. For example, between 25 February and 18 March 2020, two-thirds (586 (66.9%) of 876) of diagnostic tests were performed in private medical laboratories where costs varied typically between 300 and 690 Brazilian Reais (for context, the current minimum monthly salary is 1,045 Brazilian Reais). Thus, the true burden of the epidemic in lower-income neighbourhoods is probably underestimated. In New York City, for example, poorer neighbourhoods have been found to have a higher disease burden, which is driven in part by the movement of essential workers using public transport during the pandemic[34]. Data-driven analyses are urgently needed to help tackle health inequities during the ongoing epidemic in Brazil. Strategies to evaluate and control transmission should consider differential assess to COVID-19 diagnosis for lower-income populations, changes in reporting systems and delays in reporting, which are key to accurately determining rates of epidemic growth[35]. Innovative infectious disease surveillance approaches such as those obtained from aggregated mobility data, when used properly, could help support public health actions across the COVID-19 epidemic[36–39].

Epidemics of COVID-19 and influenza seem to have occurred simultaneously in Brazil (Fig. 4 and Extended Data Fig. 7) and symptoms overlap between the two infections. We detected co-circulation of eight other respiratory viruses, the most common of which were influenza A and B and human rhinovirus. We also detected multiple co-detection of SARS-CoV-2 with other respiratory viruses, such as influenza A and B and human metapneumovirus, which have also been reported elsewhere[40,41]. Although, co-infections with other respiratory viruses have been reported in other countries[40,42,43], no difference in clinical disease severity between cases with and without viral co-infection has been observed thus far[44]. The co-circulation of other respiratory pathogens highlights the need to scale up laboratory and molecular screening of SARS-CoV-2 and other respiratory

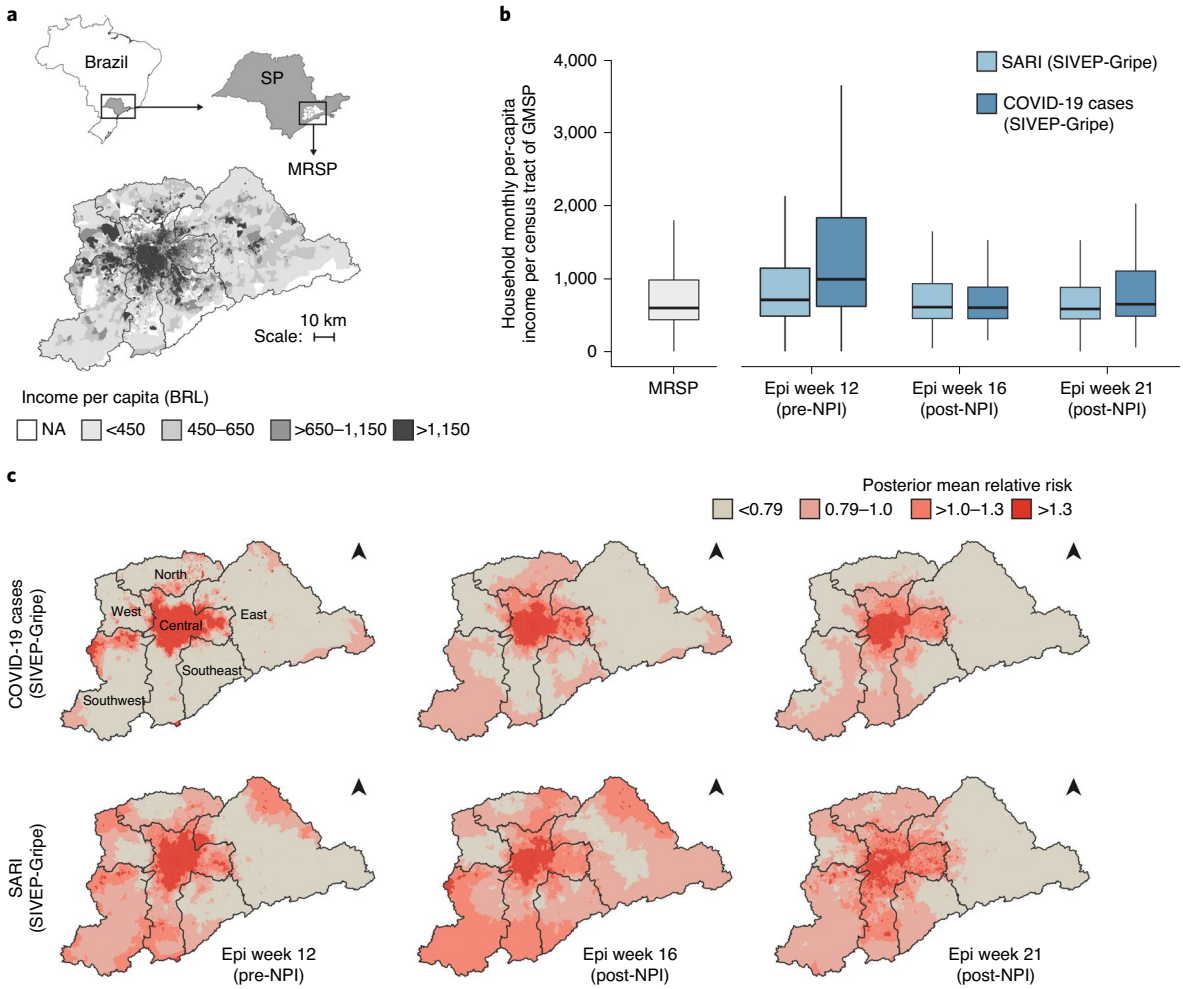


Fig. 5 | COVID-19 diagnosis and socioeconomic factors in the MRSP. a, Spatial distribution of income per capita of MRSP based on the census tract of residence. BRL, Brazilian reais. NA, not applicable. **b**, Distribution of household per-capita income based on the census tract of residence for COVID-19 cases and SARI cases with unknown aetiology. The distribution of average per-capita income for MRSP as a whole, weighted by population size, is shown on the left. Per-capita income distributions are presented in box plots, where the horizontal line inside the box represents the median per-capita income level, and the box edges show the per-capita income within the first and third interquartile range. The whiskers represent the per-capita income range. Epi, epidemiological. **c**, Posterior mean relative risk of COVID-19 confirmed diagnosis (top) and SARI cases with unknown aetiology (bottom) for epidemiological week 12 (before implementation of NPI in São Paulo state) and weeks 16 and 21 (after implementation of NPI in São Paulo state) (see Methods for details).

viruses in public laboratories across Brazil[15]. Continued molecular and genomic surveillance will be important to determine patterns of virus transmission and to guide public health measures in forthcoming phases of the epidemic[24,45–47].

There are several limitations to this study. First, detailed individual-level data were only available for the REDCap and SIVEP-Gripe systems, in which many cases had incomplete documentation, particularly regarding comorbidities. Second, our socioeconomic analysis was based partially on ecological inference, using the per-capita income in the census tract of residence (rather than the actual income of the patients), and assuming the same denominator for each census tract (~300 households). We emphasize that our spatial analysis is prone to methodological constraints caused by ecological fallacy and the modifiable areal unit problem. These constraints are inherent to any spatial analysis of aggregated data. Despite the above-mentioned limitation, census tracts correspond to small areas of analysis, of no more than 300 households but often fewer than that. Social science literature on Brazil not only highlights the country's socioeconomic inequality but also how it is spatially pronounced. For this reason, census tracts remain a useful tool with which to infer per-capita income in the absence of individual-level data. In addition, our databases were predominantly composed of hospitalized patients with COVID-19, and we were unable to evaluate the rate of hospitalization among the different socioeconomic statuses. In the future, robust modelling of the relationships between socioeconomic factors and disease severity will require a data collection system with detailed information on symptoms/signs and comorbidities both in severe and non-severe cases. Finally, our retrospective study focused predominantly on symptomatic patients who presented or were referred to health services for testing. Therefore, we are unable (and do not attempt) to describe the full spectrum of disease, nor can we describe the full epidemiological picture of this epidemic.

In conclusion, we have provided a comprehensive assessment of COVID-19 reporting and transmission in Brazil. Our findings provide important context for diagnostic screening and healthcare planning, and for future precision studies focusing on the impacts of non-pharmaceutical and pharmaceutical interventions, and the effects of social health determinants on COVID-19 transmission.

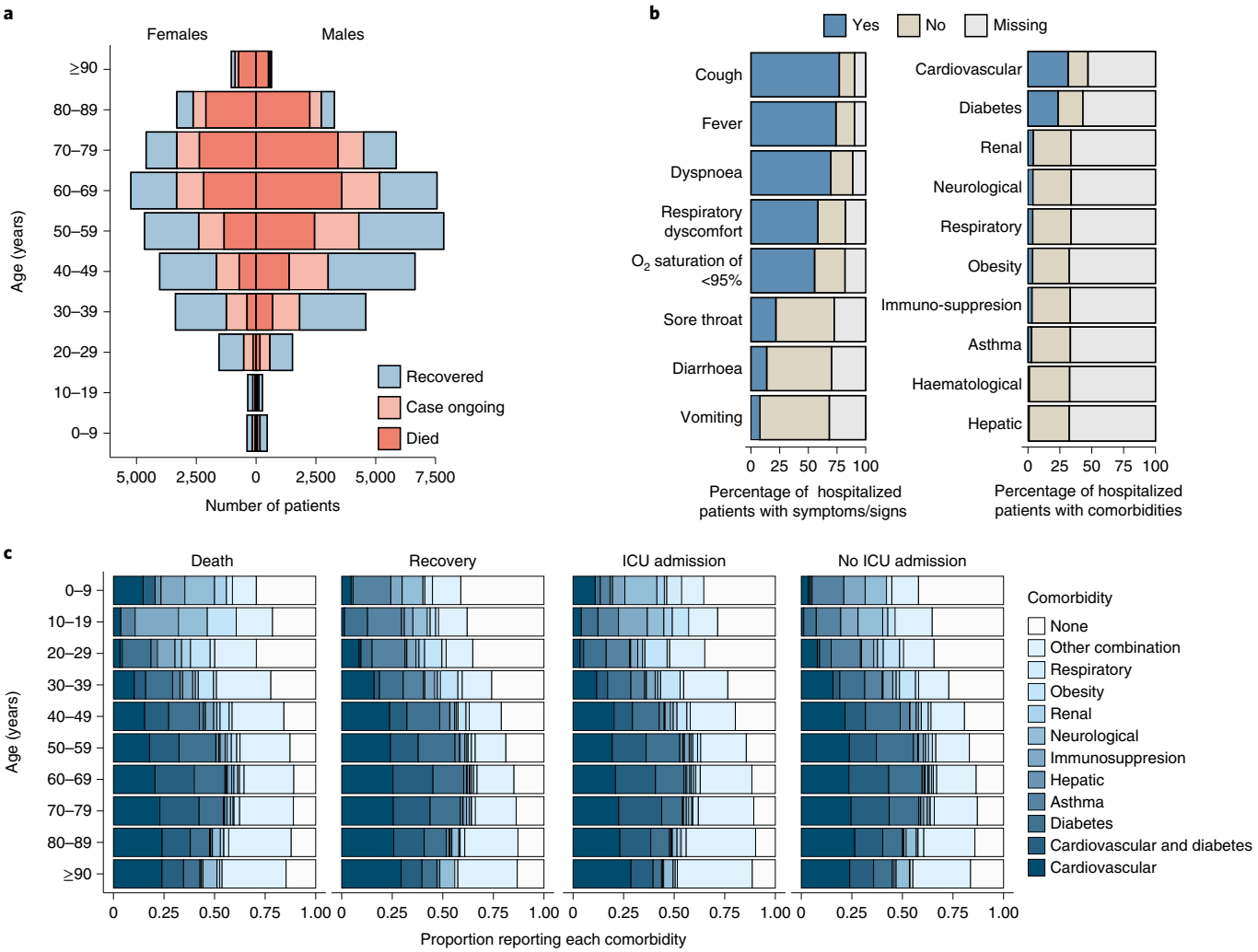


Fig. 6 | Age–sex structure and clinical features of confirmed COVID-19 cases reported in the SIVEP-Gripe system. a, Numbers of patients with ongoing COVID-19, or who have recovered or died from the disease, by age and sex. Ongoing cases were those still active on the SIVEP-Gripe database and without a recorded clinical outcome (death or recovered). **b**, Symptoms, signs and comorbidities of hospitalized individuals with confirmed COVID-19. **c**, Comorbidities among confirmed COVID-19 cases according to age and outcome ($n = 15{,}720$ confirmed COVID-19 cases with complete comorbidity and outcome (death or recovery) information; $n = 19{,}409$ confirmed COVID-19 cases with complete information on comorbidities and ICU admission). Horizontal axes show the proportion of patients in each age/outcome stratified for each of the comorbidities recorded.

Methods

Ethical approval and case definitions. This retrospective national study was supported by the Brazilian Ministry of Health and ethical approval was provided by the national ethical review board (Comissão Nacional de Ética em Pesquisa; protocol number CAAE 30127020.0.0000.0068).

A patient presenting with an acute respiratory syndrome (fever and at least one sign/symptom of respiratory illness) and: (1) a history of travel to a location with community transmission of COVID-19; or (2) contact with a confirmed or probable COVID-19 case in the 14 d preceding symptom onset; or (3) absence of an alternative diagnosis that completely explained the clinical presentation[6] was considered to have suspected COVID-19.

Initially, a traveller was suspected to have COVID-19 only when arriving from China, although the definition of suspected cases associated with travel later included Japan, Singapore, South Korea, North Korea, Thailand, Vietnam and Cambodia (21 February 2020), then also Italy, Germany, Australia, the United Arab Emirates, the Philippines, France, Iran and Malaysia (25 February 2020), then also the United States, Canada, Switzerland, the United Kingdom and four additional countries (3 March 2020). From 9 March 2020 onwards, the Ministry of Health decided to start testing all hospitalized patients with severe respiratory symptoms, regardless of their travel history.

Contact with a confirmed or probable COVID-19 case was defined as face-to-face or direct contact with someone known to have COVID-19, or direct contact in a healthcare setting. Moreover, patients reporting travel to an affected country in the preceding 14 d were considered imported cases. Cases not meeting this criterion were considered to be due to local transmission.

Suspected COVID-19 cases were confirmed by laboratory testing (that is, molecular diagnostics with real-time quantitative PCR), or by clinical epidemiological criteria. In the latter case, the classification was used when laboratory testing was inconclusive or unavailable, as recommended by the Brazilian Ministry of Health guidelines dated 6 April 2020[48], and by the World Health Organization interim guidance dated 25 March 2020[49].

Individual-level reporting of COVID-19 and SARI cases with unknown aetiology from Brazil. To investigate individual-level diagnostic and demographic data, self-reported travel history, place of residence and likely place of infection, differential diagnoses for other respiratory pathogens, as well as clinical details, including comorbidities, we collected three epidemiological data sources: (1) $n = 67{,}344$ suspected and $n = 1{,}468$ confirmed cases reported to the REDCap database from 25 February to 25 March 2020; (2) $n = 73{,}637$ confirmed SIVEP-Gripe cases from 1 March to 31 May 2020 (available at http://shiny.hmg.saude.gov.br/dataset); and (3) $n = 514{,}200$ confirmed cases from aggregated data released daily at the Portal do COVID-19 (Brazilian Health Ministry) from 25 February to 31 May 2020 (available at https://covid.saude.gov.br). The SIVEP-Gripe system reports cases of SARI, which can be defined as an acute respiratory infection with onset, within the past 10 d, of fever (≥38 °C) and cough, and typically requires hospitalization (see also Fig. 1a).

Basic reproduction number estimation. We estimated the basic reproduction number (R_0) for SARS-CoV-2 using time series of confirmed COVID-19 cases at the national and state (São Paulo, Rio de Janeiro, Ceará and Amazonas) level (Extended Data Fig. 1). To avoid the impact of NPIs on R_0 estimates, only data points up to 14 d after the implementation of the strictest interventions were used. As lockdown was not imposed in Brazil, the strictest measure was considered to be the closure of non-essential commerce. For European countries, the date of lockdown was used as the NPI date. NPI dates for Brazilian states were collected from state decrees. For Brazil as a whole, the NPI date for São Paulo state was used, as by that point most states in Brazil had already closed non-essential commerce. For the European countries, lockdown dates were collected from https://www.covid19healthsystem.org/mainpage.aspx.

To test the estimation routine and provide international context, this analysis was replicated on equivalent time series from Italy, Spain, France and the United Kingdom. Aggregated epidemiological data from the United States and China were not included due to possible heterogeneity within each country. Daily counts of confirmed cases were modelled with a negative binomial distribution with a mean equal to a fixed portion, ρ, of the total daily number of cases in an exponential model of incidence. The functional form of the incidence model is $\rho R_0 \gamma i_0 e^{(R_0-1)\gamma t}$, where ρ is the probability of an infection being counted in the time series, R_0 is the basic reproduction number, γ is the rate at which individuals cease to be infectious. and i_0 is the proportion of the population that was infectious at the start of the observations. We assume that the observed number of cases on day n was drawn from a negative binomial observation where the mean is $\mu(n)$ and the variance $\sigma = \mu + \mu 2/k$, with fixed size parameter k (dispersion parameter). The product of ρ and i_0 is denoted ξ. Since the probability of being observed and the initial condition only appear as the product ξ in the likelihood, there is an identifiability problem preventing the estimation of ρ and i_0 individually, and consequently we only consider their product, ξ. Although in this model it is theoretically possible to estimate both R_0 and γ, in practice this is difficult, so we use an informative prior to constrain γ to a priori plausible values. The factor of $\rho R_0 \gamma$ accounts for the partial observation of the incidence. In this analysis, the delay between infection and reporting was not accounted for.

Since ρ and i_0 only appear together, they were unidentifiable, and we combine them into a single parameter, ξ. This identifiability issue prevents us from estimating the prevalence without additional information to inform either i_0 or ρ. The analysis was carried out in a Bayesian framework with an uninformative prior distribution on R_0 and an informative prior on the removal rate. All other parameters had weakly informative prior distributions (see Supplementary Information). The informative prior ensures that an individual is infectious for an average of 5–14 d (ref. [50]) (Supplementary Information and Figs. 5 and 6). Standard diagnostics were used to check whether the Markov chain Monte Carlo samples were satisfactory. Full details of the model used, the estimation process and convergence of Markov chain Monte Carlo chains can be found in the Supplementary Information.

Geospatial analysis of COVID-19 cases and socioeconomic status. The average household per-capita income for the MRSP was retrieved at the census tract level from the 2010 census (https://censo2010.ibge.gov.br/). We geocoded 24,063 COVID-19 cases and 32,914 SARI cases with unknown aetiology from MRSP, which were reported until 28 May 2020. The geocoding was based on self-reported residential addresses or postal codes using the Galileo algorithm[51] and coordinates were confirmed using Google API.

To elucidate the distribution of COVID-19 cases and SARI cases with unknown aetiology, we mapped the mean relative risk of COVID-19 and SARI with unknown aetiology at the census tract level for MRSP for three epidemiological weeks (12, 16 and 21) (Extended Data Fig. 9). As the observation process was a confounding process and without additional assumptions (for example, covariates), we cannot disentangle an increase in prevalence from an increase in case ascertainment. The cumulative number of cases in each tract was modelled as a Poisson random variable with a mean specified by the expected number of cases under a null model adjusted by tract specific risk due to spatial and non-spatial effects: the Besag–York–Mollié model[19]. Estimates of the risk of COVID-19 diagnosis or SARI cases with unknown aetiology were obtained using approximate Bayesian methods (integrated nested Laplace approximation). A complete specification of the model and the computational methodology can be found in the Supplementary Information.

The association between final diagnostic category (COVID-19 or SARI with unknown aetiology) and socioeconomic status in the subset of cases in the MRSP with geocoded residential information was evaluated using logistic regression models. We focused on the cases in epidemiological weeks 12, 16 and 22. Within each of those weeks, if a census tract reported any COVID-19 or SARI with unknown aetiology, we calculated the proportion of the number of COVID-19 cases. Since most census tracts reported only one case each week, the proportion of COVID-19 cases for each census tract was mostly either 0 or 1 in a given week. For this reason, we defined two categories: (1) the census tract only reported SARI of unknown aetiology (that is, no COVID-19 cases); or (2) the census tract reported at least one COVID-19 case during the week. We used these two categories as the binary response, and applied logistic regression models to investigate whether income per capita was associated with this response. The analyses were adjusted by the logarithm of the population sizes and the longitude and latitude coordinates of the census tracts. The analyses were performed individually for each of epidemiological weeks 12, 16 and 22. Further analysis details can be found in the Supplementary Information.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Datasets of clinical and laboratory data presented in the current study from the SIVEP-Gripe/Portal do COVID-19 database are available at https://datadryad.org/stash/share/xj7kX8675lwvLzrnnPn9ebEfJNoOB38aXBTTQqfGBhE. The REDCap database and geolocation information are available from the corresponding authors upon request and ethical approval.

Code availability

The custom code used in this study is available at https://datadryad.org/stash/share/xj7kX8675lwvLzrnnPn9ebEfJNoOB38aXBTTQqfGBhE.

Received: 8 July 2020; Accepted: 15 July 2020;
Published online: 31 July 2020

References

1. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
2. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* https://doi.org/10.1038/s41564-020-0695-z (2020).
3. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
4. Guan, W.-J. et al. Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMoa2002032 (2020).
5. Livingston, E. & Bucher, K. Coronavirus disease 2019 (COVID-19) in Italy. *J. Am. Med. Assoc.* https://doi.org/10.1001/jama.2020.4344 (2020).
6. *Coronavirus Disease (COVID-2019) Situation Reports* (World Health Organization, 2020); https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
7. Croda, J. et al. COVID-19 in Brazil: advantages of a socialized unified health system and preparation to contain cases. *Rev. Soc. Bras. Med. Trop.* **53**, e20200167 (2020).
8. Jesus, J. G. et al. Importation and early local transmission of COVID-19 in Brazil, 2020. *Rev. Inst. Med. Trop. SP* **62**, e30 (2020).
9. Candido, D. S. et al. Routes for COVID-19 importation in Brazil. *J. Travel Med.* **27**, taaa042 (2020).
10. Clark, A. et al. Centre for the Mathematical Modelling of Infectious Diseases COVID-19 working group. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *Lancet Glob. Health* https://doi.org/10.1016/S2214-109X(20)30264-3 (2020).
11. Burki, T. COVID-19 in Latin America. *Lancet Infect. Dis.* **20**, 547–548 (2020).
12. Cimerman, S., Chebabo, A., Cunha, C. A. D. & Rodriguez-Morales, A. J. Deep impact of COVID-19 in the healthcare of Latin America: the case of Brazil. *Braz. J. Infect. Dis.* **24**, 93–95 (2020).
13. Ezequiel, G. E. et al. The COVID-19 pandemic: a call to action for health systems in Latin America to strengthen quality of care. *Int. J. Qual. Health Care* https://doi.org/10.1093/intqhc/mzaa062 (2020).
14. Miller, M. J., Loaiza, J. R., Takyar, A. & Gilman, R. H. COVID-19 in Latin America: novel transmission dynamics for a global pandemic? *PLoS Negl. Trop. Dis.* **14**, e0008265 (2020).
15. Andrus, J. K. et al. Perspectives on battling COVID-19 in countries of Latin America and the Caribbean. *Am. J. Trop. Med. Hyg.* https://doi.org/10.4269/ajtmh.20-0571 (2020).
16. Croda, J. H. R. & Garcia, L. P. Immediate health surveillance response to COVID-19 epidemic. *Epidemiol. Serv. Saude* **29**, e2020002 (2020).
17. Harris, P. A. et al. The REDCap consortium: building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
18. *Influenza Update* (WHO, 2020); https://www.who.int/influenza/surveillance_monitoring/updates/latest_update_GIP_surveillance/en/
19. Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**, 1–20 (1991).
20. *Pesquisa Nacional de Saúde 2013: Percepção do Estado de Saúde, Estilos de Vida e Doenças Crônicas. Brasil, Grandes Regiões e Unidades da Federação* (IBGE, 2015).
21. Mellan, T. A. et al. Report 21: estimating COVID-19 cases and reproduction number in Brazil. Preprint at *medRxiv* https://doi.org/10.1101/2020.05.09.20096701 (2020).

22. Caicedo-Ochoa, Y., Rebellon-Sanchez, D. E., Penaloza-Rallon, M., Cortes-Motta, H. F. & Mendez-Fandino, Y. R. Effective reproductive number estimation for initial stage of COVID-19 pandemic in Latin American countries. *Int. J. Infect. Dis.* **95**, 316–318 (2020).
23. Munayco, C. V. et al. Early transmission dynamics of COVID-19 in a Southern Hemisphere setting: Lima-Peru: February 29th–March 30th, 2020. *Infect. Dis. Model.* https://doi.org/10.1016/j.idm.2020.05.001 (2020).
24. Da Silva Candido, D. et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* https://doi.org/10.1126/science.abd2161 (2020).
25. Ferguson, N. et al. *Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID19 Mortality and Healthcare Demand* (Imperial College COVID-19 Response Team, 2020).
26. Walker, P. G. T. et al. The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *Science* https://doi.org/10.1126/science.abc0035 (2020).
27. Korber, B. et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.29.069054 (2020).
28. Zhang, L. et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.12.148726 (2020).
29. Khalatbari-Soltani, S., Cumming, R. G., Delpierre, C. & Kelly-Irving, M. Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. *J. Epidemiol. Commun. Health* https://doi.org/10.1136/jech-2020-214297 (2020).
30. Rivett, L. et al. Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. *eLife* **9**, https://doi.org/10.7554/eLife.58728 (2020).
31. Park, S. Y. et al. Coronavirus disease outbreak in call center, South Korea. *Emerg. Infect. Dis.* **26**, https://doi.org/10.3201/eid2608.201274 (2020).
32. Pereira, R. H. et al. *Mobilidade Urbana e o Acesso ao Sistema Único de Saúde para Casos Suspeitos e Graves de COVID-19 nas Vinte Maiores Cidades do Brasil* Nota Técnica No. 14 (Diretoria de Estudos e Politicas Regionais, Urbanas e Ambientais, IPEA, 2020).
33. Silveira, M. et al. Repeated population-based surveys of antibodies against SARS-CoV-2 in Southern Brazil. Preprint at *medRxiv* https://doi.org/10.1101/2020.05.01.20087205 (2020).
34. Sy, K. T. L., Martinez, M. E., Rader, B. & White, L. F. Socioeconomic disparities in subway use and COVID-19 outcomes in New York City. Preprint at *medRxiv* https://doi.org/10.1101/2020.05.28.20115949 (2020).
35. Dehning, J. et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* https://doi.org/10.1126/science.abb9789 (2020).
36. Buckee, C. O. et al. Aggregated mobility data could help fight COVID-19. *Science* **368**, 145–146 (2020).
37. De Oliveira, S. B. et al. Monitoring social distancing and SARS-CoV-2 transmission in Brazil using cell phone mobility data. Preprint at *medRxiv* https://doi.org/10.1101/2020.04.30.20082172 (2020).
38. Kraemer, M. U. G. et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
39. Nouvellet, P. et al. *Report 26: Reduction in Mobility and COVID-19 Transmission* (Imperial College COVID-19 Response Team, 2020).
40. Wu, X. et al. Co-infection with SARS-CoV-2 and influenza A virus in patient with pneumonia, China. *Emerg. Infect. Dis.* https://doi.org/10.3201/eid2606.200299 (2020).
41. Kim, D., Quinn, J., Pinsky, B., Shah, N. H. & Brown, I. Rates of co-infection between SARS-CoV-2 and other respiratory pathogens. *J. Am. Med. Assoc.* https://doi.org/10.1001/jama.2020.6266 (2020).
42. Cuadrado-Payan, E. et al. SARS-CoV-2 and influenza virus co-infection. *Lancet* **395**, e84 (2020).
43. Zheng, X. et al. Co-infection of SARS-CoV-2 and influenza virus in early stage of the COVID-19 epidemic in Wuhan, China. *J. Infect.* https://doi.org/10.1016/j.jinf.2020.05.041 (2020).
44. Asner, S. A. et al. Clinical disease severity of respiratory viral co-infection versus single viral infection: a systematic review and meta-analysis. *PLoS ONE* **9**, e99392 (2014).
45. Black, A., MacCannell, D. R., Sibley, T. R. & Bedford, T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat. Med.* https://doi.org/10.1038/s41591-020-0935-z (2020).
46. Deng, X. et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* https://doi.org/10.1126/science.abb9263 (2020).
47. Lu, J. et al. Genomic epidemiology of SARS-CoV-2 in Guangdong province, China. *Cell* **181**, 997–1003.e9 (2020).
48. *Coronavirus COVID-19 Diretrizes para Diagnostico e Tratamento da COVID-19* (Ministério da Saúde do Brasil, 2020).
49. *COVID-19 Coding in ICD-10* (WHO, 2020); https://www.who.int/classifications/icd/COVID-19-coding-icd10.pdf?ua=1
50. Wölfel, R. et al. Virological assessment of hospitalized patients with COVID-2019. *Nature* https://doi.org/10.1038/s41586-020-2196-x (2020).
51. Medel, C. H., Catalan, C. C., Vidou, M. A. F. & Perez, E. S. The Galileo ground segment integrity algorithms: design and performance. *Int. J. Navigation Observation* https://doi.org/10.1155/2008/178927 (2008).

Acknowledgements

We thank M. Gome, L. Bastos and L. M. Carvalho (MAVE) for useful discussions on SIVEP-Gripe, and we thank L. Matkin (Oxford) for technical support. This work was supported by a FAPESP (2018/14389-0) and Medical Research Council and CADDE partnership award (MR/S0195/1) (http://caddecentre.org/). W.M.S. is supported by the São Paulo Research Foundation, Brazil (2017/13981-0 and 2019/24251-9). N.R.F. is supported by a Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z). O.J.B. was funded by a Sir Henry Wellcome Fellowship funded by the Wellcome Trust (206471/Z/17/Z). V.H.N. and C.A.P. were supported by FAPESP (2018/12579-7). A.E.Z. and B.G. were supported by Oxford Martin School. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

W.M.S., L.F.B., D.d.S.C., R.H.M.P., C.A.P., J.C., J.-P.C., V.H.N., A.E.Z., J.M., F.C.S.S., P.d.S.A., F. Ghilardi, A.A.S.-S., B.G., C.-H.W., S.L., N.G., S.B.O., K.V.P., M.C.T.D.B., V.B.G.P., C.K.V.B., F. Ganem, W.A.F.A., F.F.S.T.F., E.M.M. and W.K.O. collected the epidemiological, spatial and clinical data and processed the statistical data. N.R.F., W.M.S., L.F.B., C.-H.W., J.-P.C., D.d.S.C., R.H.M.P., J.M., E.C.S., P.M., S.L., L.A., A.A.S.-S., G.L., A.T., M.F.V.-G., M.U.G.K., R.S.A., N.A., P.M., O.J.B., I.O.M.S., N.G., G.L., O.G.P., A.E.Z., M.L.N. and J.C. interpreted the results and wrote the manuscript. All authors read and revised the final manuscript. W.M.S., L.F.B., S. L., J.C., A. E. Z. and N.R.F. summarized the epidemiological and clinical data.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-020-0928-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-0928-4.

Correspondence and requests for materials should be addressed to J.C. or N.R.F.

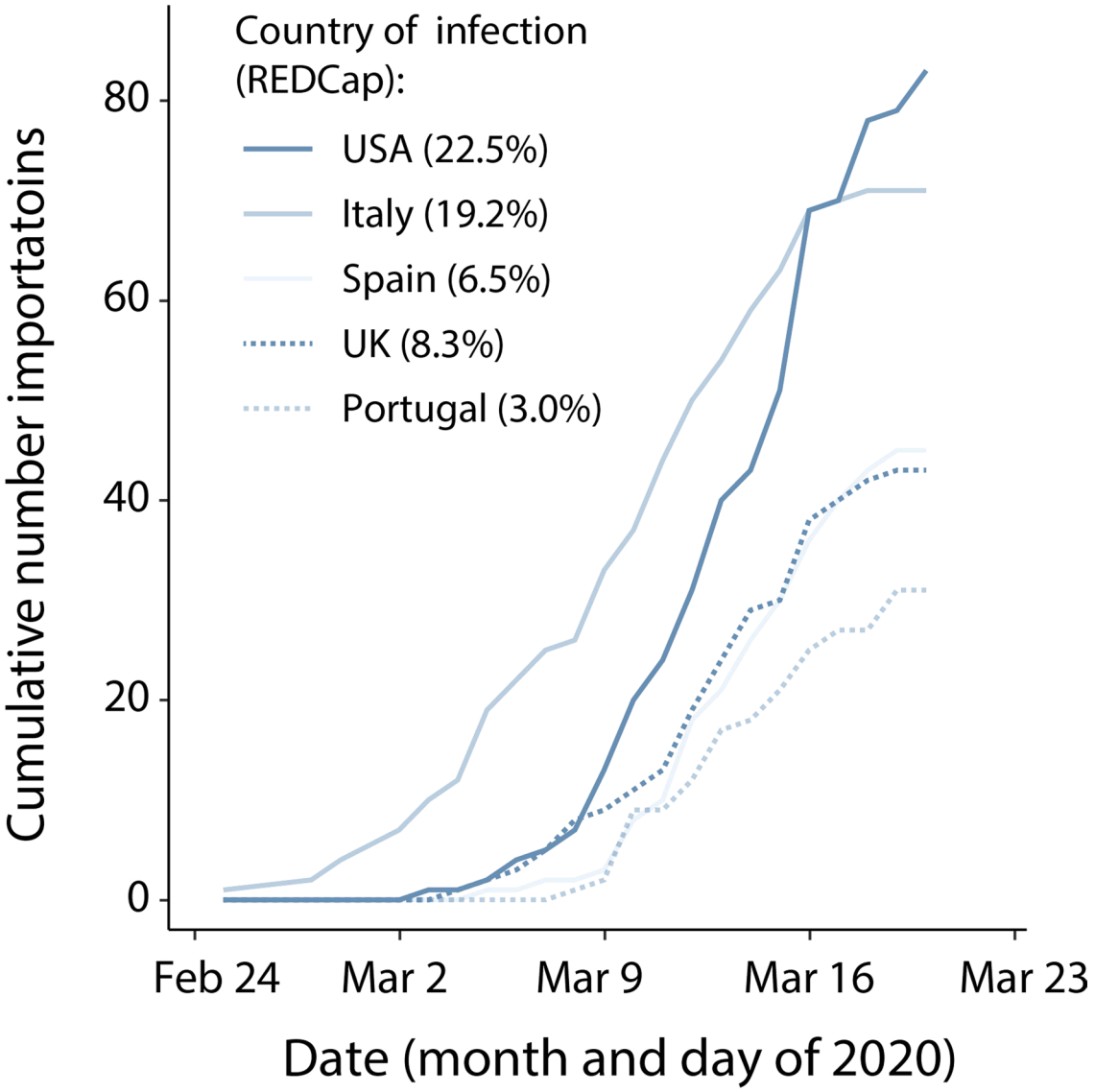
Peer review information Primary Handling Editor: Stavroula Kousta.

Reprints and permissions information is available at www.nature.com/reprints.

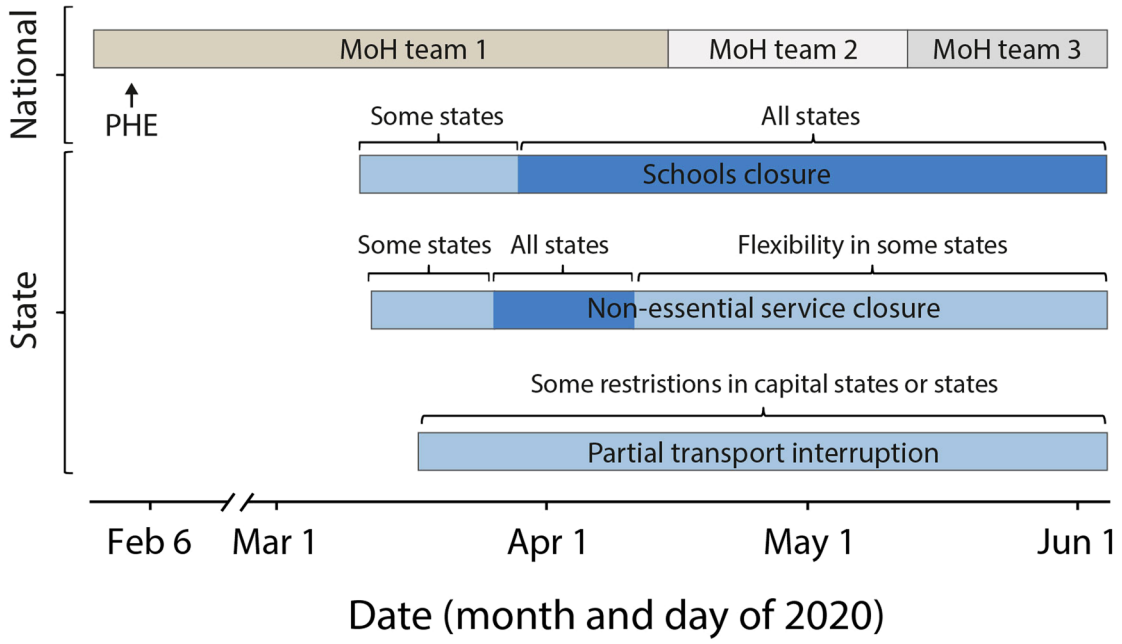
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Centro de Pesquisa em Virologia, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, Brazil. [2]Instituto de Medicina Tropical, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil. [3]Department of Zoology, University of Oxford, Oxford, UK. [4]Department of Research in Virology and Biotechnology, Gorgas Memorial Institute of Health Studies, Panama City, Panama. [5]School of Geography and the Environment, University of Oxford, Oxford, UK. [6]Institute for Applied Economic Research (IPEA), Brasília, Brazil. [7]Escola Politécnica, Universidade de São Paulo, São Paulo, Brazil. [8]Brazilian Studies Programme, Latin American Centre, University of Oxford, Oxford, UK. [9]MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, London, UK. [10]Department of Medical Microbiology and Infection Prevention, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [11]Oxford School of Global and Area Studies, University of Oxford, Oxford, UK. [12]School of Biological and Environmental Sciences, Universidad San Francisco de Quito (USFQ), Quito, Ecuador. [13]Harvard Medical School, Harvard University, Boston, MA, USA. [14]Boston Children's Hospital, Boston, MA, USA. [15]Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [16]MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology, Faculty of Epidemiology

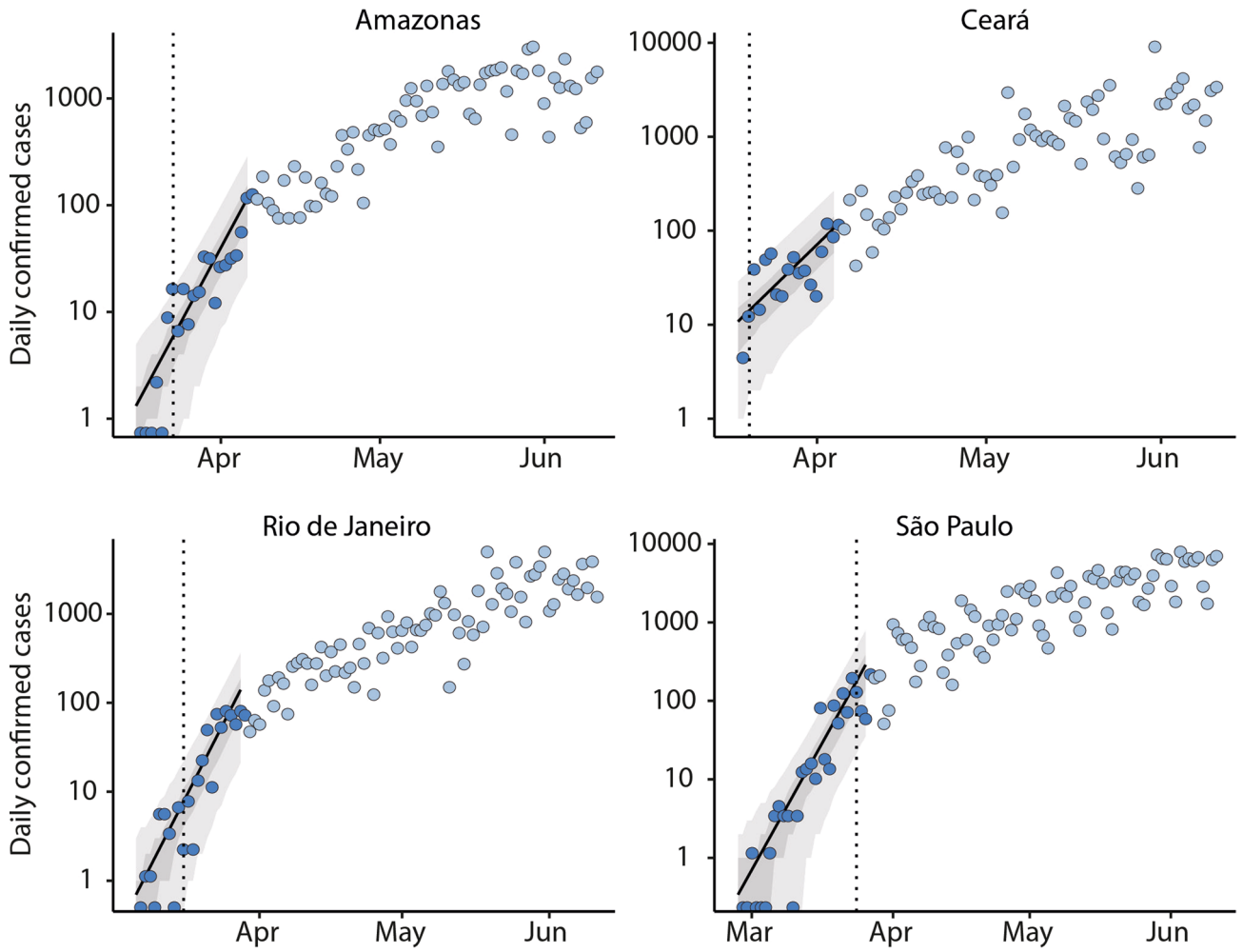
and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [17]Department of Clinical Research, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [18]Centre for the Mathematical Modelling of Infectious Diseases, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [19]Núcleo de Vigilância Epidemiológica do Hospital das Clínicas da Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil. [20]Departamento de Medicina Preventiva, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil. [21]Department of Epidemiology and Health Statistics, Xiangya School of Public Health, Central South University, Changsha, China. [22]Secretariat of Health Surveillance, Department of Immunization and Communicable Diseases, Brazilian Ministry of Health, Brasília, Brazil. [23]Secretariat of Health Surveillance, Brazilian Ministry of Health, Brasília, Brazil. [24]Faculdade de Medicina de São José do Rio Preto, São Jose do Rio Preto, Brazil. [25]Mathematical Sciences, University of Southampton, Southampton, UK. [26]Laboratório de Pesquisa em Ciências da Saúde, Universidade Federal da Grande Dourados, Dourados, Brazil. [27]Fundação Oswaldo Cruz, Campo Grande, Brazil. [28]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University, New Haven, CT, USA. [29]These authors contributed equally: William Marciel de Souza, Lewis Fletcher Buss, Darlan da Silva Candido, Jean-Paul Carrera, Sabrina Li, Chieh-Hsi Wu, Julio Croda, Ester C. Sabino, Nuno Rodrigues Faria.
✉e-mail: juliocroda@gmail.com; nfaria@ic.ac.uk



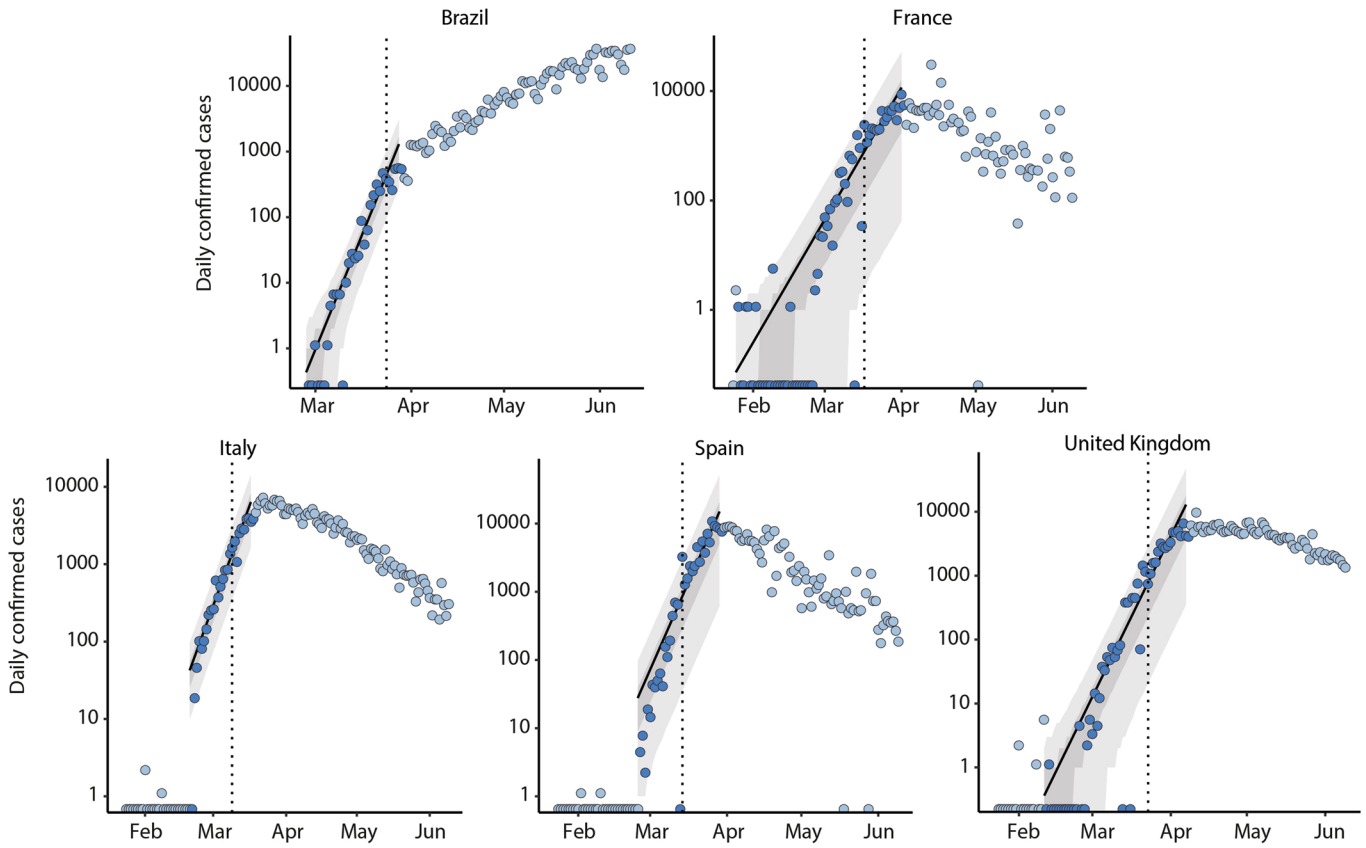
Extended Data Fig. 1 | Imported cases by self-reported country of infection from REDCap database. Percentage indicates proportion of cases acquired outside of Brazil between 25 February and 19 March (n = 342) by unambiguously identified country of infection as recorded in REDCap database (see also Fig. 1).



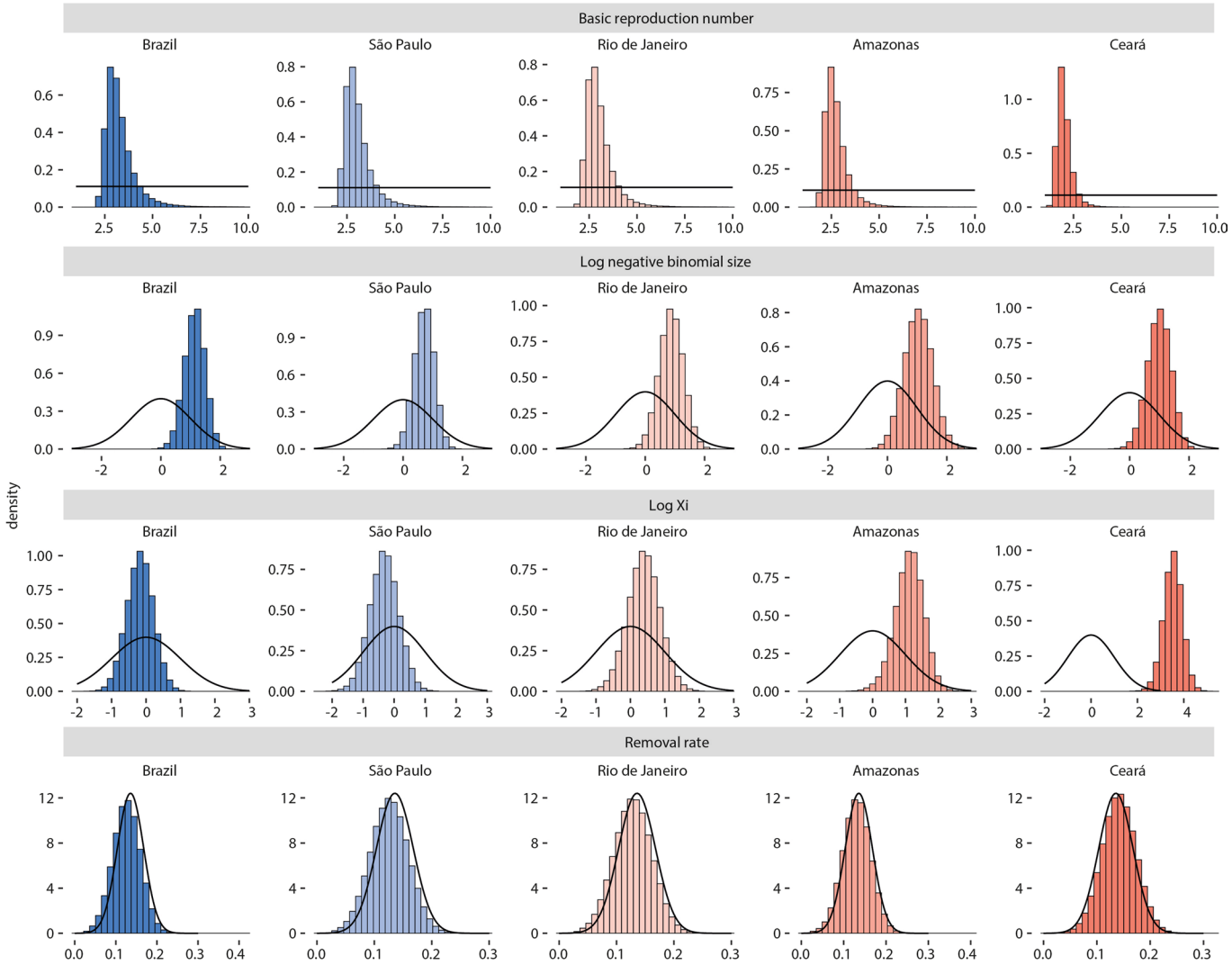
Extended Data Fig. 2 | Non-pharmaceutical interventions taken during the first three months of the epidemic in Brazil. Time of implementation of measures for COVID-19 control in Brazil. PHE = declaration of Public Health Emergency of International Concern. MoH=Ministry of Health. Data on non-pharmaceutical interventions compiled from state official decrees can be found in Supplementary Table 1.



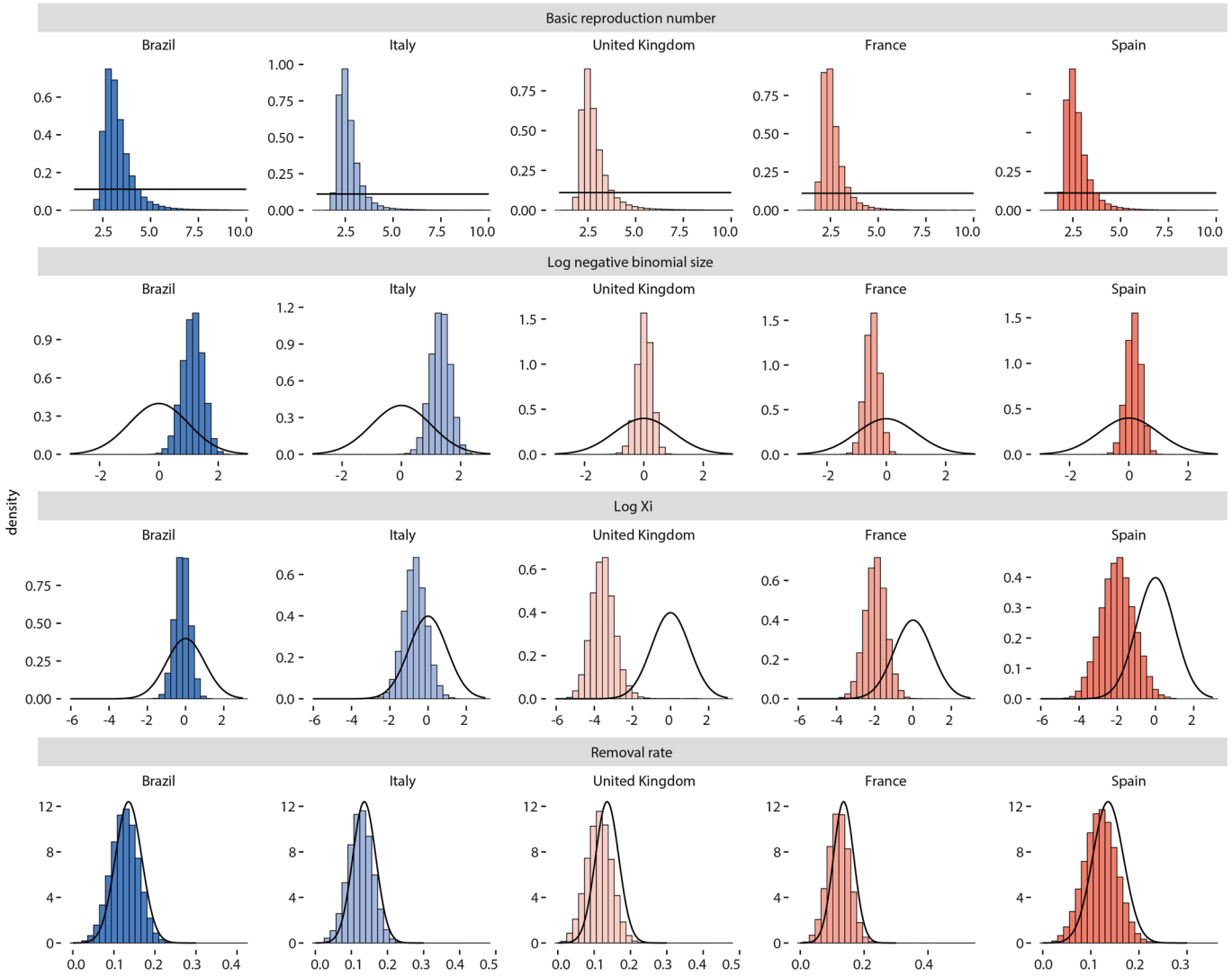
Extended Data Fig. 3 | Daily number of infections used for the R0 estimations of confirmed cases of Brazil and European countries (France, Italy, Spain, and United Kingdom). The dashed vertical line indicates when the non-pharmaceutical intervention (NPI) was implemented. The dark blue dots were used to estimate R_0. The shaded region is the model fit for those data points. The light blue dots included how the time series continued. They were included to show the effects of NPI.



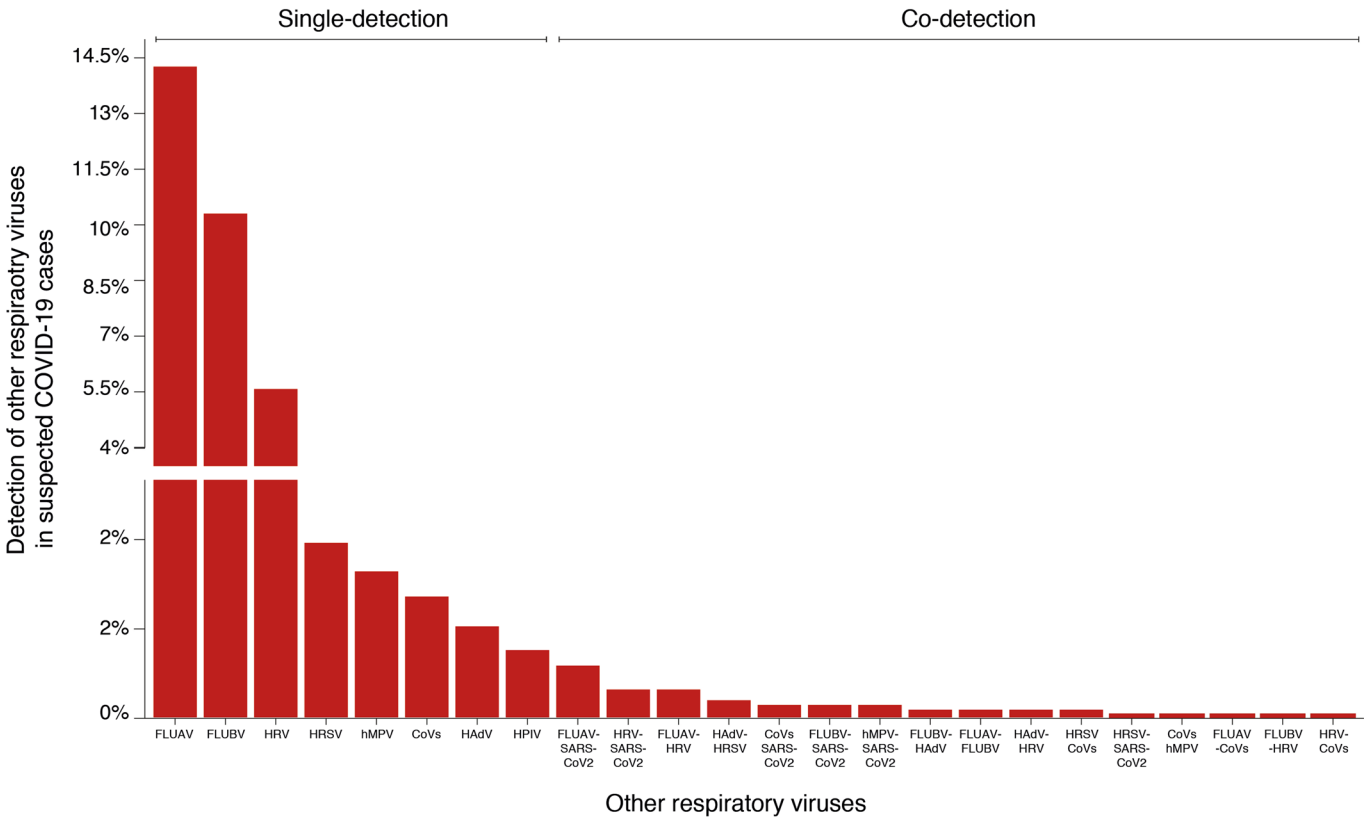
Extended Data Fig. 4 | Daily number of infections used for the R0 estimations of confirmed cases in states of Amazonas, Ceará, Rio de Janeiro, and São Paulo. The dashed vertical line indicates when the NPI was implemented. The dark blue dots were used to estimate R0. The shaded region is the model fit for those data points. The light blue dots included how the time series continued. They were included to show the effects of NPI.



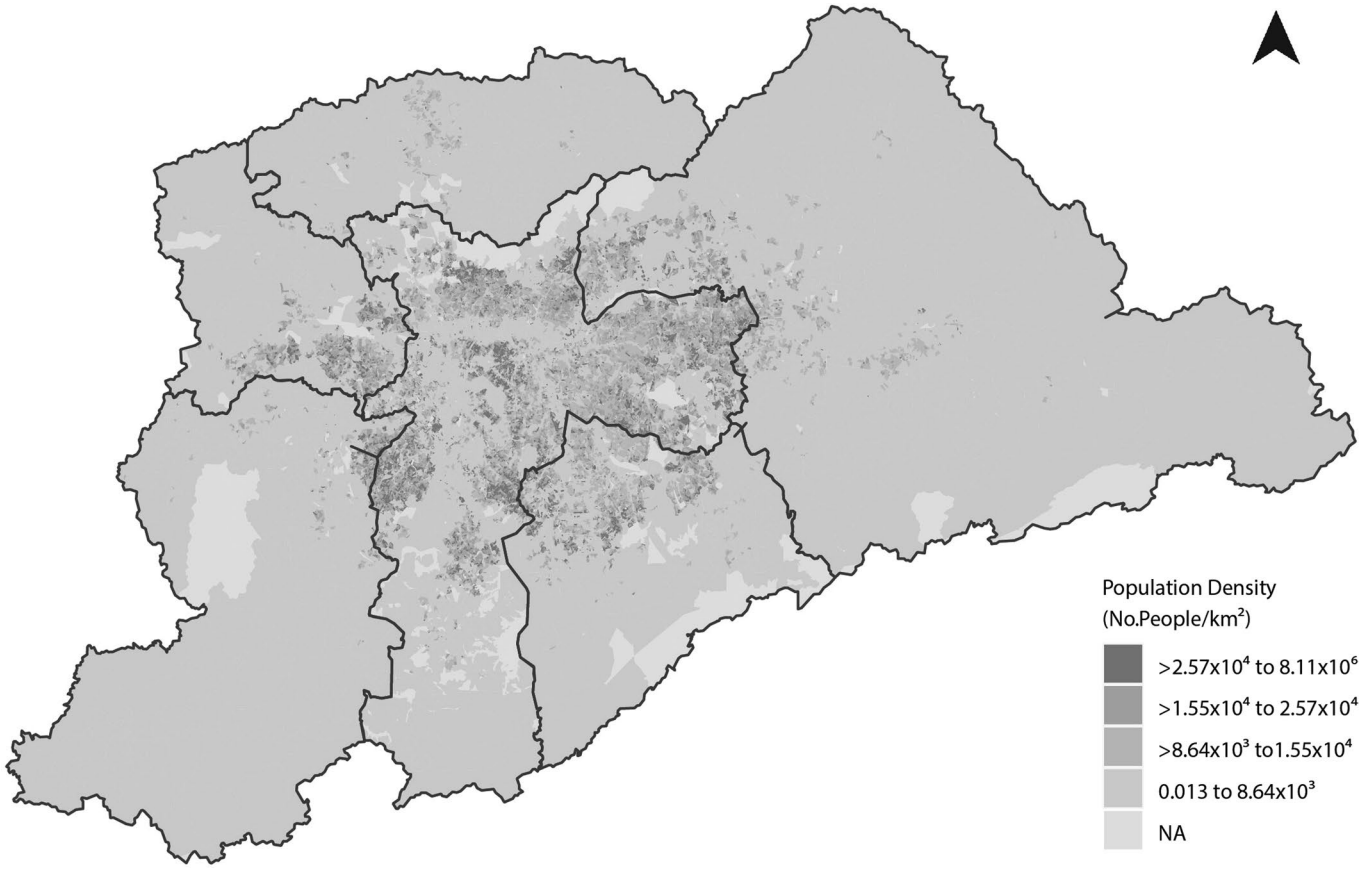
Extended Data Fig. 5 | The prior/posterior plots for the different parameters in the analysis of the time series from all of Brazil, and states of São Paulo, Rio de Janeiro, Amazonas, and Ceará. The histogram is of the posterior samples and the solid line shows the prior density about those values. From top to bottom, they are basic reproduction number, the log of the size of the negative binomial distribution, ξ, and removal rate.



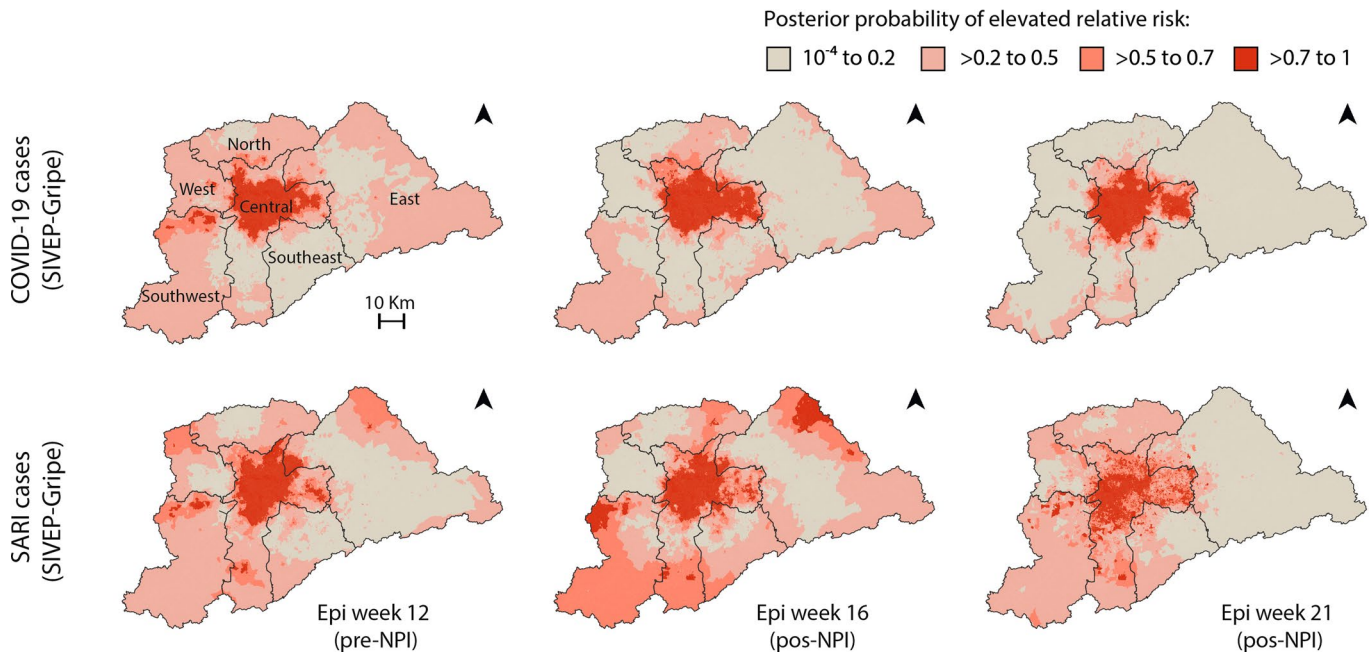
Extended Data Fig. 6 | The prior/posterior plots for the different parameters in the analysis of the time series of Brazil, Italy, the United Kingdom, France, and Spain. The histogram is of the posterior samples and the solid line shows the prior density about those values. From top to bottom, they are basic reproduction number, the log of the size of the negative binomial distribution, ξ, and removal rate.



Extended Data Fig. 7 | Diagnosis of other respiratory viruses in 2,429 suspected COVID-19 cases reported to Brazilian Ministry of Health between February 25 to March 25, 2020. influenza A virus (FLUAV), influenza B virus (FLUBV), human rhinovirus (HRV), human respiratory syncytial virus (HRSV), human metapneumovirus (hMPV), human adenovirus (HAdV), human parainfluenza viruses 1-4 (HPIV), and CoVs (that is, human coronavirus 229E, OC43, NL63 and HKU1).



Extended Data Fig. 8 | Map of the population density in each census tract in the Metropolitan Region of São Paulo. NA=not applicable.



Extended Data Fig. 9 | COVID-19 diagnosis and socio-economic factors in the Metropolitan Region of São Paulo. Posterior probability of elevated relative risk of COVID-19 for confirmed diagnosis (upper panels) and SARI cases with unknown aetiology (lower panels) for epidemiological weeks 12 (pre-implementation of non-pharmaceutical interventions in São Paulo state, and weeks 16 and 21 (post-implementation of non-pharmaceutical interventions in São Paulo state).

nature research

Corresponding author(s): Nuno Rodrigues Faria

Last updated by author(s): Jul 7, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Datasets of clinical and laboratory data presented in the current study will be made available from the corresponding authors upon request and ethical approval.
Data analysis	Custom code used in this study can found in GitHub repository (Link available upon acceptance of the publication).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Custom code used in this study can found in GitHub repository (Link available upon acceptance of the publication). Datasets of clinical and laboratory data presented in the current study will be made available from the corresponding authors upon request and ethical approval.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable as this is an observational study.
Data exclusions	We have not excluded data.
Replication	Not applicable as this is an observational study.
Randomization	Not applicable as this is an observational study.
Blinding	Not applicable as this is an observational study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	We have used data from COVID-19 cases notified to Brazilian Ministry of Health as suspected COVID-19 infections until the May 31, 2020. The characteristics of the dataset used in our study are described in detail in Materials and Methods (see also Data Availability statement above).
Recruitment	Not applicable.
Ethics oversight	National ethical review board (Comissão Nacional de Ética em Pesquisa), Brazil (CAAE 30127020.0.0000.0068).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Dataset on SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities

Andreza Aruska de Souza Santos[1]✉, Darlan da Silva Candido[2], William Marciel de Souza[2,3], Lewis Buss[4], Sabrina L. Li[5], Rafael H. M. Pereira[6], Chieh-Hsi Wu[7], Ester C. Sabino[4] & Nuno R. Faria[2,4,8]

Brazil has one of the fastest-growing COVID-19 epidemics worldwide. Non-pharmaceutical interventions (NPIs) have been adopted at the municipal level with asynchronous actions taken across 5,568 municipalities and the Federal District. This paper systematises the fragmented information on NPIs reporting on a novel dataset with survey responses from 4,027 mayors, covering 72.3% of all municipalities in the country. This dataset responds to the urgency to track and share findings on fragmented policies during the COVID-19 pandemic. Quantifying NPIs can help to assess the role of interventions in reducing transmission. We offer spatial and temporal details for a range of measures aimed at implementing social distancing and the dates when these measures were relaxed by local governments.

Background & Summary

Brazil has seen one of the highest case numbers of COVID-19 in the world. As of 6 December 2020, Brazil recorded over 6,577,177 million cases and more than 176,628 deaths (https://covid19.who.int). SARS-CoV-2 was introduced at least 100 times in Brazil[1]. Non-pharmaceutical interventions (NPIs), although unequal in date of implementation and duration, reduced virus transmission[1,2]. Several factors including changes to the national COVID-19 notification system[3] and uncoordinated implementation of public health measures may have contributed to rapid epidemic spread across the country. Here, we describe the complexity of asynchronous adoption and easing of NPIs in Brazilian municipalities.

Our data were gathered in a continuous municipal-level survey conducted by the Brazilian Confederation of Municipalities (*Confederação Nacional de Municípios* – CNM). Despite the existing examination of national and state-level NPI strategies[4], a city-level assessment of NPI's beyond capitals and second cities[5] is still missing. Local-level data collection is a challenge because of the number of municipalities in Brazil (5,568 municipalities and the Federal District), and each municipality passed a number of decrees related to COVID-19 control measures. This dataset offers a unique fine-grained understanding of local-level policies in Brazil, aiding future examination on the roles of NPIs on the increase, spread, and duration of local outbreaks.

Data Sources

The CNM interviewed 4,027 (72.3%) of 5,568 mayors and the Federal District's government on the implementation and relaxation of NPIs between 13 May and 31 July 2020. Response rates varied by region: North (29.1% of 450 municipalities), Northeast (50.5% of 1,793 municipalities), Centre-West (71.7% of 466 municipalities), Southeast (90.2% of 1,668 municipalities) and South (96.6% of 1,191 municipalities). This difference was attributed to municipal infrastructure and the starting region of the survey, moving South to North.

[1]Oxford School of Global and Area Studies, University of Oxford, Oxford, UK. [2]Department of Zoology, University of Oxford, Oxford, UK. [3]Virology Research Centre, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil. [4]Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [5]School of Geography and the Environment, University of Oxford, Oxford, UK. [6]Institute for Applied Economic Research, Brasília, Brazil. [7]Mathematical Sciences, University of Southampton, Southampton, UK. [8]MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, London, UK. ✉e-mail: andreza.desouzasantos@lac.ox.ac.uk

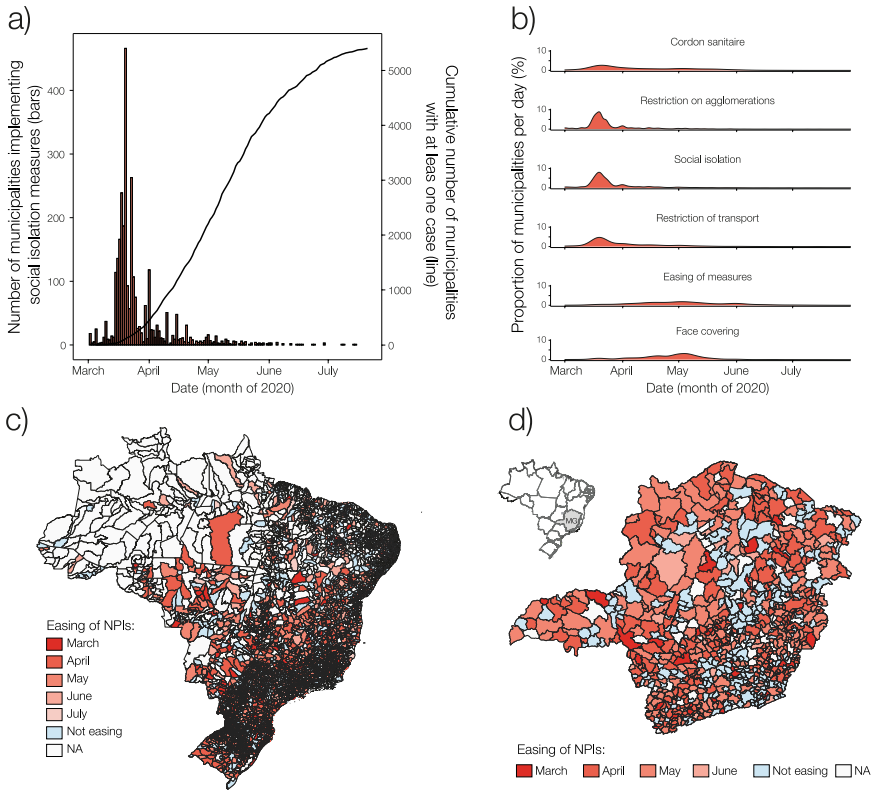


Fig. 1 (**a**) Prohibition of non-essential services in the country (red bars) and the cumulative number of municipalities reporting at least one case (black line). (**b**) Density plots showing dates of adoption and easing of NPIs by municipalities in Brazil. (**c**) Starting month for easing of NPIs across municipalities in Brazil. (**d**) Starting month for easing of NPIs in the state of Minas Gerais (MG). NA - not applicable.

In March, a number of municipalities closed non-essential services (2,237), prohibited large gatherings (2,932), reduced public transportation (999), and implemented cordon sanitaires (930), with a rapid uptake by mid-March (Fig. 1a). At that time, COVID-19 cases were restricted to a few highly populated state capitals, with cases mostly associated with overseas travel[6]. Of the total of 3,958 mayors that responded to the question on implementing social isolation (closure of all non-essential services), 3,062 municipalities adopted the measure and among those, 2,738 (89.4%) implemented the measure before the first reported case in their municipality (Fig. 1b). Despite an early uptake and comprehensive NPI adoption, in only two months, SARS-CoV-2 spread from 296 municipalities (5.3%), on 31 March 2020, to 4,196 municipalities (75.3%), as of 31 May 2020 (Fig. 1a).

In other countries, implementations of NPI have been associated with fewer and delayed cases[7]; while a lack of coordination has been associated with disease spread and resurgence[8]. Although distancing measures were adopted across Brazil early in the pandemic, easing of these measures began as early as the end of March (Fig. 1c), often disregarding decisions by neighbouring municipalities, as illustrated in Fig. 1d for the state of Minas Gerais. We chose Minas Gerais to illustrate our dataset because of the high response rate to the survey. Of a total of 853 mayors in that state, only between 38–47 mayors failed to respond to specific NPI questions. We represented data absence with the colour white in Fig. 1d. We also chose to detail Minas Gerais because almost one-sixth of all Brazilian municipalities are located in that state. Municipal borders do not limit the flow for shopping trips or work commuting across towns[9,10]. Nevertheless, as Fig. 1c,d shows, decisions to ease NPIs were not coordinated between bordering cities.

Contributions and Recommendations

When the Brazilian Supreme Court ruled on 15 April 2020 that mayors and governors were autonomous in their decisions related to the pandemic[11], collecting local data on the management of the pandemic became urgent. With declining willingness of national and regional governments to impose national/regional lockdowns, the role of local measures to control the pandemic becomes increasingly important to understand transmission patterns at finer geographic scales. This high-resolution dataset on the NPIs in Brazil is an important contribution that also highlights challenges. Early and cohesive closure of non-essential activities was short-lived in Brazil, and municipalities are lifting distancing measures in an uncoordinated manner, starting as early as late March.

The easing of NPIs needs to be examined in relation to reductions in confirmed cases, hospital and testing capacities, mitigation policies such as compulsory use of face coverings, and the potential impact of local policies on neighbouring towns. City borders are porous and cities that have maintained strict social distancing policies may face a growing number of cases because of external decisions. Policy evaluation of Brazil's management of the pandemic will need to account for the uneven duration of control measures through NPIs – which include personal (physical distancing, isolation quarantine, hand hygiene, and face covering), environmental (surface

cleaning and ventilation) and social (travel restrictions, school and workplace closures, restriction on mass gatherings) – across the country[12]. This dataset allows for this assessment, aiding future research and policymaking.

Methods

Data collection. In order to collect these data, we started a specific collaboration with the Brazilian Confederation of Municipalities (CNM). The CNM is a non-partisan and non-profit organisation that works with mayors in Brazil, especially those that manage municipalities under 100,000 inhabitants, a focus that corresponds to 94% of Brazil's municipalities. As the largest municipal association in Brazil, CNM possesses contact details of Brazilian elected mayors. The capillarity of that organisation makes it an ideal partner for such large-scale data collection. When the COVID-19 outbreak started and the CNM conducted the survey, we formed a partnership to analyse and deposit the dataset and thus expand access to these data. There were no shared financial responsibilities between researchers and the surveying institution.

The details of this cooperation were established through a meeting followed by a written agreement signed by the first and last authors of this paper with CNM on 9 April 2020. The partnership was established because of the need to understand the impact of decentralized measures in Brazil and what decentralisation causes to the spread of infectious diseases. Upon establishing this collaboration, CNM added further questions to the questionnaire for their monitoring of municipalities, such as budgetary information possibly affected by the pandemic. CNM had already designed and conducted a previous survey independently[13], but upon our feedback, they added dates of implementation of NPIs to the survey questionnaire that we report on.

Mayors were contacted through a call centre. The CNM's call centre is independently run; they contact mayors regularly on different policy themes. The phone-based survey collected information on local NPI policies related to COVID-19, Mayors had the option to receive a protected password to respond to the questionnaire online at a later time. When mayors were unable to respond to the survey questions, they suggested an alternative respondent, such as the municipal health secretary.

Mayors and representatives that responded to the survey had the option of updating previous answers: they were contacted by phone multiple times and they could use a protected password to update online. This methodology acknowledged cases when municipalities, in the course of data collection, could have relaxed NPIs to later re-establish social distancing. The questionnaire aimed at the first date of NPI implementation and the current state of easing NPIs. Short-lived decisions on NPIs during the course of the survey were not captured in the questionnaire. However, call centre operators wrote an observation on the limited cases when municipalities reported erratic NPIs lifting. A total of 144 municipalities (2.58% of the total) described having re-opened non-essential services due to local businesses' and inhabitants' pressure, and/or because of reduced number of confirmed cases, and/or they followed the state governor. However, in those 144 cases, they soon decided to re-establish social distancing when the number of confirmed cases increased. We detailed such municipalities in Online-only Table 1.

In total, the questionnaire had 47 questions; our database has 5 columns related to the identification of the municipality and 13 of the 47 questions that were part of our collaboration to document NPI policy strategies: 6 thematic questions with respective 6 dates of implementation and 1 question pertaining to percentage.

Data classification. In summary, our dataset includes: (1) adoption of cordon sanitaire, (2) prohibition of agglomeration, (3) closure of all but essential services, (4) compulsory use of face covering, (5) reduction in public transportation services and if so, the percentage of the reduction, and (6) whether easing of the above measures were applied.

We had uniform data entry for all survey data. The dataset has information for the majority but not all municipalities in Brazil (4,027 of 5,568 municipalities and the Federal District). Below we offer a breakdown of the number of answers for each question: (Q1) adoption of cordon sanitaire: 3,976, (Q2) prohibition of agglomeration: 3,965, (Q3) closure of all but essential services: 3,958, (Q4) compulsory use of face covering: 3,952 (Q5) reduction in public transportation offer: 3,908, and (Q6) if there was already any easing of the above distancing measures, 3,947. When the answer to the above questions was yes, the column related to date of implementation was also populated.

We collected information on policies adapting a classification system on the 20 most frequent categories of NPIs to control the spread of COVID-19[4]. Considering the list of frequently adopted NPIs, we focused on the ones that would have a direct impact on the spatial mobility of residents. These measures were associated with a specific date of implementation. Adapting an international NPI classification to Brazil's municipal reality requires some explanation. We did not report on closure of education institutions because such measures were implemented at the state level[3]. Similarly, we did not report on airport restrictions because such policies are not a municipal duty; on the other hand, the implementation of cordon sanitaire was a local decision and we report on those.

We reported on mass gathering cancellations (such as discotheques and sport events) and small gatherings restrictions (the closure of all but essential services). Reduction in public transportation services was not among the 20 most frequent NPIs categories[4]. However, this measure was frequently implemented, and potentially causing unintended consequences. Reductions in public transport services combined with low levels of social isolation can result in overcrowding of transport stations and vehicles. We invite further examination of transportation reduction and mobility patterns in Brazilian municipalities. Finally, we included a question on the compulsory use of face covering, as the compulsory use of masks is usually related to the re-opening of non-essential services and decrees were locally passed. For all NPIs, we also ask the date of implementation, and that field was populated in the format DD/MM/YYYY (Table 1). We also adapted the naming of NPIs to the Brazilian context, with mass gathering cancellation being named as "prohibition of agglomeration", and small gathering cancellation as "closure of non-essential services".

During the 45 days of data collection (13 May to 31 July 2020), the CNM staff and the authors had access to a summary containing the total number of responses, which were classified as (1) complete, 3,174 interviews; (2) ongoing, 853 interviews; (3) pending, 1,536 interviews; or (4) without response, 6 interviews. We also received a partial dataset on the 13 of July.

Column names (in Portuguese)	Column names (in English)	Number of records
IBGE	Unique Id	5569
Município	Municipality	5569
UF	State Acronym	27
Capitais	Capitals	Sim (27) / Não (5542)
Região	Region	5
Q1. Barreiras sanitárias (posto de monitoramento de entrada e saída de pessoas no Município)	Q1. Cordon Sanitaire (monitoring of entrance and exit of people in the municipality)	3976
Q1. Data Início (se sim)	Q1. Start date (if yes)	2021
Q2. Medidas restritivas para diminuição da circulação/aglomeração de pessoas.	Q2. Restrictions to avoid circulation/ agglomeration of people	3965
Q2. Data Início (se sim)	Q2. Start date (if yes)	3707
Q3. Medidas de isolamento social, permitindo APENAS serviços essenciais.	Q3. Measures of social isolation, allowing ONLY essential services	3958
Q3. Data Início (se sim)	Q3. Start date (if yes)	2901
Q4. Uso obrigatório de máscaras faciais.	Q4. Compulsory use of face covering	3952
Q4. Data Início (se sim)	Q4. Start date (if yes)	3588
Q5. Foram adotadas medidas de redução na oferta de transporte público?	Q5. Were any measures implemented to reduce the offer of public transportation?	3908
Q5. Qual foi a porcentagem de redução	Q5. What was the percentage of reduction?	1647
Q5. Data Início (se sim)	Q5. Start date (if yes)	1590
Q6. Houve flexibilização das medidas restritivas e de isolamento social.	Q6. Were measures of restriction and social isolation eased?	3947
Q6. Data Início (se sim)	Q6. Start date (if yes)	2319

Table 1. Column names and data summary.

Finally, for mapping NPIs (Fig. 1), we used the official spatial datasets with the administrative boundaries of states and municipalities organized by the Brazilian Institute of Geography and Statistics (IBGE).

Data Records

The latest version of the data was updated on the 31 July 2020. We explain the time lapse between finishing the survey and making it available online below, which included budgeting time for validation. The dataset is available online (https://doi.org/10.5061/dryad.vdncjsxs2)[14]. An additional data report and a description of the project (containing all 47 survey questions) are available online[15]. We describe the dataset fields that pertain to our cooperation in detail below. We have kept the dataset in its original language, Portuguese, to increase the usage by health professionals and scholars in Brazil. We offer a translation to English to make these data of international use:

Main dataset. IBGE (unique_id): unique id for each municipality defined according to IBGE (Brazilian Institute for Geography and Statistics).

Município (Name of the municipality)

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

Capital, dropdown option: sim (yes); não (no);

Região (Region), dropdown options as follows: Centro-Oeste (Centre-West); Norte (North), Sul (South), Nordeste (Northeast), Sudeste (Southeast);

Q1. Barreiras sanitárias - posto de monitoramento de entrada e saída de pessoas no Município - (Cordon Sanitaire - monitoring of entrance and exit of people in the municipality);
Q1. Data Início (Q1. Start date)
Q2. Medidas restritivas para diminuição da circulação/aglomeração de pessoas (Restrictions to avoid circulation/ agglomeration of people);
Q2. Data Início (Q2. Start date);
Q3. Medidas de isolamento social, permitindo APENAS serviços essenciais (Measures of social isolation, allowing ONLY essential services);
Q3. Data Início (Q3. Start date);
Q4. Uso obrigatório de máscaras faciais (Compulsory use of face covers);
Q4. Data Início (Q4. Start date);
Q5. Foram adotadas medidas de redução na oferta de transporte público? (Were any measures implemented to reduce the offer of public transportation?);

Q5. Qual foi a porcentagem de redução? (What was the percentage of reduction?);
Q5. Data Início (Q5. Start date)
Q6. Houve flexibilização das medidas restritivas e de isolamento social? (Were measures of restriction and social isolation eased?);
Q6. Data Início (Q6. Start date)

Technical Validation

Interviewing mayors and health secretaries offers an important aspect of interpretation of laws and validation of dates related to non-pharmaceutical interventions. Even though decrees restricting physical contact or relaxing social distancing measures are available online, there were multiple laws on similar issues (e.g., a decree closing non-essential services followed by another one defining non-essential services, with a third one deciding on the duration of such activity). For that reason, consolidated information coming from those at the policy-making side increases precision.

To verify the correctness of received answers, we compared our dataset with existing information on adoption of NPIs on a city-level, which is mostly available for state-capitals. The eight capital cities for which we had data (Curitiba, Florianopólis, Fortaleza, Goiânia, João Pessoa, Manaus, Teresina, and Vitória), home to approximately 5% of the Brazilian population and distributed across all five Brazilian regions, offered answers and dates that are compatible with those collected by independent scholars looking at decrees[5].

In addition to comparison with other datasets, we checked our data for possible erroneous entries, such as municipalities that eased NPIs before having implemented them. Five municipalities (Ajuricaba/RS; Estiva Gerbi/SP; Paiçandu/PR; São Lourenço/MG; Senador Cortes/MG) added 01/03/2020 as dates when they relaxed NPIs, which is equal or earlier than the date they implemented social distancing measures. These dates of NPI relaxation are likely to be erroneous.

During data collection, call centre operators detailed a limited number of municipalities that adopted NPIs, eased those, and later re-implemented distancing measures (as detailed in Online-only Table 1). The limited number of municipalities describing such behaviour corresponds with findings from other independently collected data on NPIs[5]. Researchers looking at capitals and second large cities found out that between NPIs implementation and easing, variations usually refer to capacity of specific services (e.g., restaurants or public transport) with cases of re-opening followed by a new set of distancing policies being still relatively rare. The reasons for such stable behaviour can vary and require further examination. This may be connected with the political costs of changing policies frequently, especially in a year of municipal elections. In addition to that, Brazil only significantly reduced its infection rate in October and at the time of writing, the country goes through an increase in number of infections. A new round of NPIs could potentially happen in the near future.

Finally, our multi-faceted data validation process also included broadly reporting on the data. On 9 September 2020, the CNM made public its summary report[15] and we waited for two-weeks after that launch before submitting a first version of this dataset description. In this time period, no mayor requested revision of the data entered. Despite a lack of contestations from mayors to date, and the low number of possible erroneous data, omissions exist, and we invite users to complement this dataset using published decrees and media sources. The possibility of new lockdowns potentially followed by a new easing of NPIs would require a future survey.

Below we include examples of how these data could be used to allow for the continuity of policy surveys and guarantee a high level of cooperation between mayors and research units.

Usage Notes

This database offers an opportunity for researchers and policymakers to examine the potential impacts of NPIs on COVID-19 transmission and control in Brazil. This is a unique dataset as it collates responses for the majority of all Brazilian cities, while previous datasets have mainly focused on capital cities, states, or have looked at a national level[4,5]. Because these data were collected through a survey, answers given by municipal authorities may be inaccurate. Unfortunately, even the scrutiny of laws also allows for a level of inaccuracy in interpretation and discrepancies on the exact date of policy implementation may be disputed in some cases. We have mitigated this problem with early and well-broadcasted release of a report on this dataset and technical validation included comparing our dataset with existing others on capital cities, we also looked for erroneous entries in our database. We invite researchers to use the dataset as it is. If significant inaccuracies are found, we will update our dataset and will describe such an update when it takes place.

This dataset represents a baseline for further research as it describes how COVID-19 response took place in a continental country such as Brazil. Because not all municipal authorities answered to all questions, particularly in the North region of Brazil, we suggest users to consider additional sources of information to document missing policy implementation, preferably using official sources such as local decrees. However, as decrees are not always available online, secondary sources such as media reports may need to be consulted.

A significant contribution of this dataset comes from the 'release of NPIs' column. At the time of writing, agglomerations (e.g., football stadiums) remain prohibited in most cities in Brazil. The easing of NPIs therefore mainly relates to the re-opening of non-essential services. Given that re-opening was asynchronous, researchers trying to retrieve exact dates on the easing of NPIs do not have a target period and searching for laws can be tiresome, especially when considering the total number of municipalities in Brazil. Finally, decrees that established the re-opening of shops and restaurants were often modified a few days later, such modifications especially related to capacity of small gatherings. The amendment of decrees makes NPI dates particularly susceptible to errors, potentially over or underestimating the date of easing of social distancing. A survey with mayors and health authorities is thus a fundamental tool as it allows us to listen to those on the frontline to describe when was the pivotal date for the reopening of services in town, information otherwise blurred in different decrees across time.

Scholars using our dataset could investigate whether easing of NPIs preceded increases in population mobility levels, or if adherence to NPIs was already low when NPIs were still in place. The connection between easing of NPIs and the compulsory use of face covering also invites further examination on the potential mitigation effects of masks. As the pandemic progresses and as Brazil is a highly affected country, we invite researchers to use the data to understand the pandemic and support health policymakers in their efforts.

Received: 5 October 2020; Accepted: 2 February 2021;
Published online: 04 March 2021

References

1. Candido, D. S. *et al*. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260 (2020).
2. Mellan, T. *et al*. Report 21: Estimating COVID-19 cases and reproduction number in Brazil. *Imperial College London* https://doi.org/10.25561/78872 (2020).
3. de Souza, W. M. *et al*. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. *Nat. Hum. Behav.* **4**, 856–865 (2020).
4. Desvars-Larrive, A. *et al*. A structured open dataset of government interventions in response to COVID-19. *Sci. Data* **7**, 285 (2020).
5. Petherick, A. *et al*. Brazilian Sub-National Covid-19 Policy Responses. *GitHub* https://github.com/OxCGRT/Brazil-covid-policy (2020).
6. Candido, D. S. *et al*. Routes for COVID-19 importation in Brazil. *J. Travel Med.* **27**, 3 (2020).
7. Tian, H. *et al*. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638–642 (2020).
8. Rukṭanonchai, N. W. *et al*. Assessing the impact of coordinated COVID-19 exit strategies across Europe. *Science* **369**, 1465–1470 (2020).
9. Holtz, D. *et al*. Interdependence and the cost of uncoordinated responses to COVID-19. *Proc. Natl. Acad. Sci. USA* **117**, 19837–19843 (2020).
10. Peixoto, P. S., Marcondes, D., Peixoto, C. & Oliva, S. M. Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil. *PLoS One* **15**, 7 (2020).
11. Supremo Tribunal Federal. Ação direta de inconstitucionalidade 6341. *Portal.stf.jus.br* https://portal.stf.jus.br/processos/detalhe.asp?incidente=5880765 (2020).
12. Dye, C., Cheng, R. C. H., Dagpunar, J. S. & Williams, B. G. The scale and dynamics of COVID-19 epidemics across Europe. *R. Soc. Open Sci.* **7**, 201726 (2020).
13. Confederação Nacional de Municípios. Pesquisa sobre o novo coronavírus (Covid-19). *Cnm.org.br* https://www.cnm.org.br/biblioteca/exibe/14582 (2020).
14. de Souza Santos, A. A. *et al*. SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities. *Dryad* https://doi.org/10.5061/dryad.vdncjsxs2 (2020).
15. Confederação Nacional de Municípios. Pesquisa CNM sobre COVID-19: foco na gestão municipal e apoio dos entes federados. *Cnm.org.br* https://www.cnm.org.br/biblioteca/exibe/14729 (2020).

Acknowledgements

The authors thank the Brazilian Confederation of Municipalities (CNM) for the partnership as well as all mayors and health secretaries that responded to the questionnaire. We thank Alexander Mielke for valuable input. We also thank our funding agencies: WMS: FAPESP #2017/13981-0 and 2019/24251-9NRF: CADDE/FAPESP (MR/S0195/1 and FAPESP 18/14389-0) (http://caddecentre.org/) and Wellcome Trust and Royal Society Sir Henry Dale Fellowship 204311/Z/16/Z.SLL: Canadian Social Sciences and Humanities Doctoral Fellowship and the Oxford Martin School Programme on Pandemic Genomics. DSC: Clarendon Fund and Department of Zoology.

Author contributions

A.A.D.S.S., D.C.S., W.M.S. contributed equally: data interpretation, writing, figure, study design, revision. L.B.: figures, data interpretation and text revision. S.L.L., R.H.M.P., C.H.W., E.S. and N.R.F. contributed with data interpretation and text revision. A.A.D.S.S. and N.R.F. set up the collaboration with the Brazilian Confederation of Municipalities (CNM).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.A.d.S.S.

Chapter 3

Evolution and epidemic spread of SARS-CoV-2 in Brazil

This chapter is, to date, the largest peer-reviewed genomic epidemiology study on the early spread of SARS-CoV-2 in Brazil. It results from an outstanding logistical and genome sequencing effort to rapidly build and analyse a genomic dataset representative of the epidemiological situation in Brazil during the first two months of SARS-CoV-2 spread. It provides the first evidence for several importation events, genetic diversity and routes of national spread of the virus. This work was first made available on MedRxiv as a preprint on the 23rd June 2020 and published in *Science* in July 2020 and it is presented here in full.

Candido DS*, *Claro IM*, de Jesus JG*, Souza WM*, Moreira FRR*, Dellicour S*, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020 Sep 4;369(6508):1255–60.*

"Be fast, have no regrets... If you need to be right before you move, you will never win"

Dr Michael J Ryan

"Everything that happens twice will surely happen a third time"

Paulo Coelho — The Alchemist

CORONAVIRUS

Evolution and epidemic spread of SARS-CoV-2 in Brazil

Darlan S. Candido[1,2]*, Ingra M. Claro[2,3]*, Jaqueline G. de Jesus[2,3]*, William M. Souza[4]*, Filipe R. R. Moreira[5]*, Simon Dellicour[6,7]*, Thomas A. Mellan[8]*, Louis du Plessis[1], Rafael H. M. Pereira[9], Flavia C. S. Sales[2,3], Erika R. Manuli[2,3], Julien Thézé[10], Luiz Almeida[11], Mariane T. Menezes[5], Carolina M. Voloch[5], Marcilio J. Fumagalli[4], Thaís M. Coletti[2,3], Camila A. M. da Silva[2,3], Mariana S. Ramundo[2,3], Mariene R. Amorim[12], Henrique H. Hoeltgebaum[13], Swapnil Mishra[8], Mandev S. Gill[7], Luiz M. Carvalho[14], Lewis F. Buss[2], Carlos A. Prete Jr.[15], Jordan Ashworth[16], Helder I. Nakaya[17], Pedro S. Peixoto[18], Oliver J. Brady[19,20], Samuel M. Nicholls[21], Amilcar Tanuri[5], Átila D. Rossi[5], Carlos K. V. Braga[9], Alexandra L. Gerber[11], Ana Paula de C. Guimarães[11], Nelson Gaburo Jr.[22], Cecila Salete Alencar[23], Alessandro C. S. Ferreira[24], Cristiano X. Lima[25,26], José Eduardo Levi[27], Celso Granato[28], Giulia M. Ferreira[29], Ronaldo S. Francisco Jr.[11], Fabiana Granja[12,30], Marcia T. Garcia[31], Maria Luiza Moretti[31], Mauricio W. Perroud Jr.[32], Terezinha M. P. P. Castiñeiras[33], Carolina S. Lazari[34], Sarah C. Hill[1,35], Andreza Aruska de Souza Santos[36], Camila L. Simeoni[12], Julia Forato[12], Andrei C. Sposito[37], Angelica Z. Schreiber[38], Magnun N. N. Santos[38], Camila Zolini de Sá[39], Renan P. Souza[39], Luciana C. Resende-Moreira[40], Mauro M. Teixeira[41], Josy Hubner[42], Patricia A. F. Leme[43], Rennan G. Moreira[44], Maurício L. Nogueira[45], Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic Network, Neil M. Ferguson[8], Silvia F. Costa[2,3], José Luiz Proenca-Modena[12], Ana Tereza R. Vasconcelos[11], Samir Bhatt[8], Philippe Lemey[7], Chieh-Hsi Wu[46], Andrew Rambaut[47], Nick J. Loman[21], Renato S. Aguiar[39], Oliver G. Pybus[1], Ester C. Sabino[2,3]†, Nuno Rodrigues Faria[1,2,8]†

Brazil currently has one of the fastest-growing severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemics in the world. Because of limited available data, assessments of the impact of nonpharmaceutical interventions (NPIs) on this virus spread remain challenging. Using a mobility-driven transmission model, we show that NPIs reduced the reproduction number from >3 to 1 to 1.6 in São Paulo and Rio de Janeiro. Sequencing of 427 new genomes and analysis of a geographically representative genomic dataset identified >100 international virus introductions in Brazil. We estimate that most (76%) of the Brazilian strains fell in three clades that were introduced from Europe between 22 February and 11 March 2020. During the early epidemic phase, we found that SARS-CoV-2 spread mostly locally and within state borders. After this period, despite sharp decreases in air travel, we estimated multiple exportations from large urban centers that coincided with a 25% increase in average traveled distances in national flights. This study sheds new light on the epidemic transmission and evolutionary trajectories of SARS-CoV-2 lineages in Brazil and provides evidence that current interventions remain insufficient to keep virus transmission under control in this country.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel betacoronavirus with a 30-kb genome that was first reported in December 2019 in Wuhan, China (*1*, *2*). SARS-CoV-2 was declared a public health emergency of international concern on 30 January 2020. As of 12 July 2020, there were >12.5 million cases of coronavirus disease 2019 (COVID-19) and 561,000 deaths globally (*3*). The virus can be classified into two main phylogenetic lineages, A and B, which spread from Wuhan before strict travel restrictions were enacted (*4*, *5*) and now cocirculate around the world (*6*). The case fatality ratio of SARS-CoV-2 infection has been estimated at between 1.2 and 1.6% (*7*–*9*), with substantially higher ratios in those >60 years of age (*8*). Some estimates suggest that 18 to 56% of SARS-CoV-2 transmission is from asymptomatic or presymptomatic individuals (*10*–*13*), complicating epidemiological assessments and public health efforts to curb the pandemic.

Challenges of real-time assessment of transmission

Although the SARS-CoV-2 epidemics in several countries, including China, Italy, and Spain, have been brought under control through nonpharmaceutical interventions (NPIs) (*3*), the number of SARS-CoV-2 cases and deaths in Brazil continues to increase (*14*) (Fig. 1A). As of 12 July 2020, Brazil had reported 1,800,827 SARS-CoV-2 cases, the second-largest number in the world, and 70,398 deaths. More than one-third of the cases (34%) in Brazil are concentrated in the southeast region, which includes São Paulo city (Fig. 1B), the world's fourth-largest conurbation, where the first case in Latin America was reported on 25 February 2020 (*15*). Diagnostic assays for SARS-CoV-2 molecular detection were widely distributed across the regional reference centers of the national public health laboratory network from 21 February 2020 on (*16*, *17*). However, several factors, including delays in reporting, changes in notification, and heterogeneous access to testing across populations, obfuscate the real-time assessment of virus transmission using SARS-CoV-2 case counts (*15*). Consequently, a more accurate measure of SARS-CoV-2 transmission in Brazil is the number of reported deaths caused by severe acute respiratory infections (SARIs), which is provided by the Sistema Único de Saúde (SUS) (*18*). Changes in the opportunity for SARS-CoV-2 transmission are strongly associated with changes in average mobility (*18*–*20*) and can typically be measured by calculating the effective reproduction number, R, defined as the average number of secondary infections caused by an infected person. $R > 1$ indicates a growing epidemic, whereas $R < 1$ is needed to achieve a decrease in transmission.

We used a Bayesian semimechanistic model (*21*, *22*) to analyze SARI mortality statistics and human mobility data to estimate daily changes in R in São Paulo city (12.2 million inhabitants) and Rio de Janeiro city (6.7 million inhabitants), the largest urban metropoles in Brazil (Fig. 1, C and D). NPIs in Brazil consisted of school closures implemented between 12 and 23 March 2020 across the country's 27 federal units/states and store closures implemented between 13 and 23 March 2020. In São Paulo city, schools started closing on 16 March 2020 and stores closed 4 days later. At the start of the epidemics, we found $R > 3$ in São Paulo and Rio de Janeiro and, concurrent with the timing of state-mandated NPIs, R values fell close to 1.

Mobility-driven changes in *R*

Analysis of R values after NPI implementation highlights several notable mobility-driven features. There was a period immediately after NPIs, between 21 and 31 March 2020, when R was consistently <1 in São Paulo city (Fig. 1C). However, after this initial decrease, the R value for São Paulo rose to >1 and increased through time, a trend associated with increased population mobility. This can be seen in the Google transit stations index, which rose from −60 to −52%, and by a decrease in the social isolation index from 54 to 47%. By 4 May 2020, we estimate R = 1.3 [95% Bayesian credible interval (BCI): 1.0 to 1.6] in both São Paulo and Rio de Janeiro cities (table S1). However, we note that there were instances in the previous 7 days when the 95% credible intervals for R included values <1, drawing attention to the fluctuations and uncertainty in the estimated R for both cities.

Early sharing of genomic sequences, including the first SARS-CoV-2 genome, Wuhan-Hu-1, released on 10 January (*23*), has enabled unprecedented global levels of molecular testing for an emerging virus (*24*, *25*). However, despite the thousands of virus genomes deposited on public access databases, there is a lack of consistent sampling structure and there are limited data from Brazil (*26*–*28*), which

hampers accurate reconstructions of virus movement and transmission using phylogenetic analyses. To investigate how SARS-CoV-2 became established in the country, and to quantify the impact of NPIs on virus spatiotemporal spread, we tested a total of 26,732 samples from public and private laboratories using real-time quantitative polymerase chain reaction (RT-qPCR) assays and found 7944 (29%) to be positive for SARS-CoV-2. We then focused our sequencing efforts on generating a large and spatially representative genomic dataset with curated metadata to maximize the association between the number of sequences and the

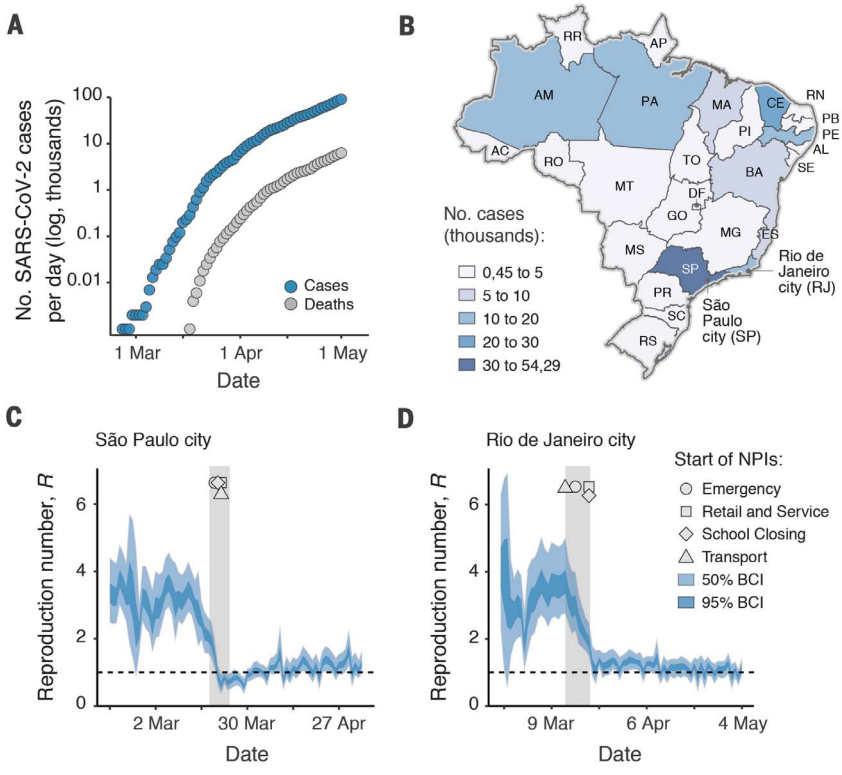


Fig. 1. SARS-CoV-2 epidemiology and epidemic spread in Brazil. (**A**) Cumulative number of SARS-CoV-2 reported cases (blue) and deaths (gray) in Brazil. (**B**) States are colored according to the number of cumulative confirmed cases by 30 April 2020. (**C** and **D**) *R* over time for the cities of São Paulo (C) and Rio de Janeiro (D). *R* values were estimated using a Bayesian approach incorporating the daily number of deaths and four variables related to mobility data (a social isolation index from Brazilian geolocation company *InLoco* and Google mobility indices for time spent in transit stations, parks, and the average between groceries and pharmacies, retail and recreational, and workspaces). Dashed horizontal line indicates $R = 1$. Gray area and geometric symbols show the times at which NPIs were implemented. BCIs of 50 and 95% are shown as shaded areas. The two-letter ISO 3166-1 codes for the 27 federal units in Brazil are provided in the supplementary materials.

[1]Department of Zoology, University of Oxford, Oxford, UK. [2]Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [3]Departamento de Moléstias Infecciosas e Parasitárias, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [4]Centro de Pesquisa em Virologia, Faculdade de Medicina de Ribeirão Preto, Ribeirão Preto, Brazil. [5]Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. [6]Spatial Epidemiology Lab, Université Libre de Bruxelles, Brussels, Belgium. [7]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium. [8]MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, London, UK. [9]Institute for Applied Economic Research, Brasília, Brazil. [10]Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, Saint-Genès-Champanelle, France. [11]Laboratório de Bioinformática, Laboratório Nacional de Computação Científica, Petrópolis, Brazil. [12]Departamento de Genética, Evolução, Microbiologia e Imunologia, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, Brazil. [13]Department of Mathematics, Imperial College London, London, UK. [14]Escola de Matemática Aplicada (EMAp), Fundação Getúlio Vargas, Rio de Janeiro, Brazil. [15]Department of Electronic Systems Engineering, University of São Paulo, São Paulo, Brazil. [16]Usher Institute, University of Edinburgh, Edinburgh, UK. [17]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil. [18]Departamento de Matemática Aplicada, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil. [19]Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK. [20]Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, UK. [21]Institute for Microbiology and Infection, University of Birmingham, Birmingham, UK. [22]DB Diagnósticos do Brasil, São Paulo, Brazil. [23]LIM 03 Laboratório de Medicina Laboratorial, Hospital das Clínicas Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [24]Instituto Hermes Pardini, Belo Horizonte, Brazil. [25]Departamento de Cirurgia, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [26]Simile Instituto de Imunologia Aplicada Ltda, Belo Horizonte, Brazil. [27]Laboratório DASA, São Paulo, Brazil. [28]Laboratório Fleury, São Paulo, Brazil. [29]Laboratório de Virologia, Instituto de Ciências Biomédicas, Universidade Federal de Uberlândia, Uberlândia, Brazil. [30]Centro de Estudos da Biodiversidade, Universidade Federal de Roraima, Boa Vista, Brazil. [31]Divisão de Doenças Infecciosas, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, Brazil. [32]Hospital Estadual Sumaré, Universidade Estadual de Campinas, Campinas, Brazil. [33]Departamento de Doenças Infecciosas e Parasitárias, Faculdade de Medicina, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. [34]Divisão de Laboratório Central do Hospital das Clínicas, da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [35]Department of Pathobiology and Population Sciences, Royal Veterinary College, Hatfield, UK. [36]University of Oxford, Latin American Centre, Oxford School of Global and Area Studies, Oxford, UK. [37]Departamento de Clínica Médica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, Brazil. [38]Departamento de Patologia Clínica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, Brazil. [39]Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [40]Departamento de Botânica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [41]Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [42]Departamento de Biologia Celular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [43]Centro de Saúde da Comunidade, Universidade Estadual de Campinas, Campinas, Brazil. [44]Centro de Laboratórios Multiusuários, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [45]Laboratório de Pesquisas em Virologia, Faculdade de Medicina de São José do Rio Preto, São José do Rio Preto, São Paulo, Brazil. [46]Mathematical Sciences, University of Southampton, Southampton, UK. [47]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

*These authors contributed equally to this work.

†Corresponding author. Email: sabinoec@usp.br (E.C.S.); nfaria@ic.ac.uk (N.R.F.)

Corrected 10 September 2020. See full text.

number of SARS-CoV-2 confirmed cases per state.

Spatially representative sequencing efforts

We generated 427 new SARS-CoV-2 genomes with >75% genome coverage from Brazilian samples collected between 5 March and 30 April 2020 (figs. S1 to S3 and data S1). For each state, the time between the date of the first reported case and the collection date of the first sequence analyzed in that state was only 4.5 days on average (Fig. 2A). For eight federal states, genomes were obtained from samples collected up to 6 days before the first case notifications. The genomes generated here were collected in 85 municipalities across 18 of 27 federal units spanning all regions in Brazil (Fig. 2A and fig. S2). Sequenced genomes were obtained from samples collected 4 days on average (median, range: 0 to 29 days) after the onset of symptoms and were generated in three laboratories using harmonized sequencing and bioinformatic protocols (table S2). When we include 63 additional available sequences from Brazil deposited in GISAID (29) (see data S1 and S2), we found the dataset to be representative of the spatial heterogeneity of the Brazilian epidemic. Specifically, the number of genomes per state strongly correlated with SARI SARS-CoV-2 confirmed cases and SARI cases with unknown etiology per state (n = 490 sequences from 21 states, Spearman's correlation, ρ = 0.83; Fig. 2A). This correlation varied from 0.70 to 0.83 when considering SARI cases and deaths caused by SARS-CoV-2 and SARI cases and deaths from unknown etiology (fig. S4). Most (n = 485/490) Brazilian sequences belong to SARS-CoV-2 lineage B, with only five strains belonging to lineage A (two from Amazonas, one from Rio Grande do Sul, one from Minas Gerais, and one from Rio de Janeiro; data S1 and fig. S5 show detailed lineage information for each sequence). Moreover, we used an in silico assessment of diagnostic assay specificity for Brazilian strains (n = 490) to identify potential mismatches in some assays targeting these strains. We found that the forward primers of the Chinese CDC and Hong Kong University nucleoprotein-targeting RT-qPCR may be less appropriate for use in Brazil than other diagnostic assays, for which few or no mismatches were identified (fig. S6 and table S3). The impact of these mismatches on the sensitivity of these assays should be confirmed experimentally. If sensitivity is affected, then the use of duplex RT-qPCR assays that concurrently target different genomic regions may help in the detection of viruses with variants in primer- or probe-binding regions.

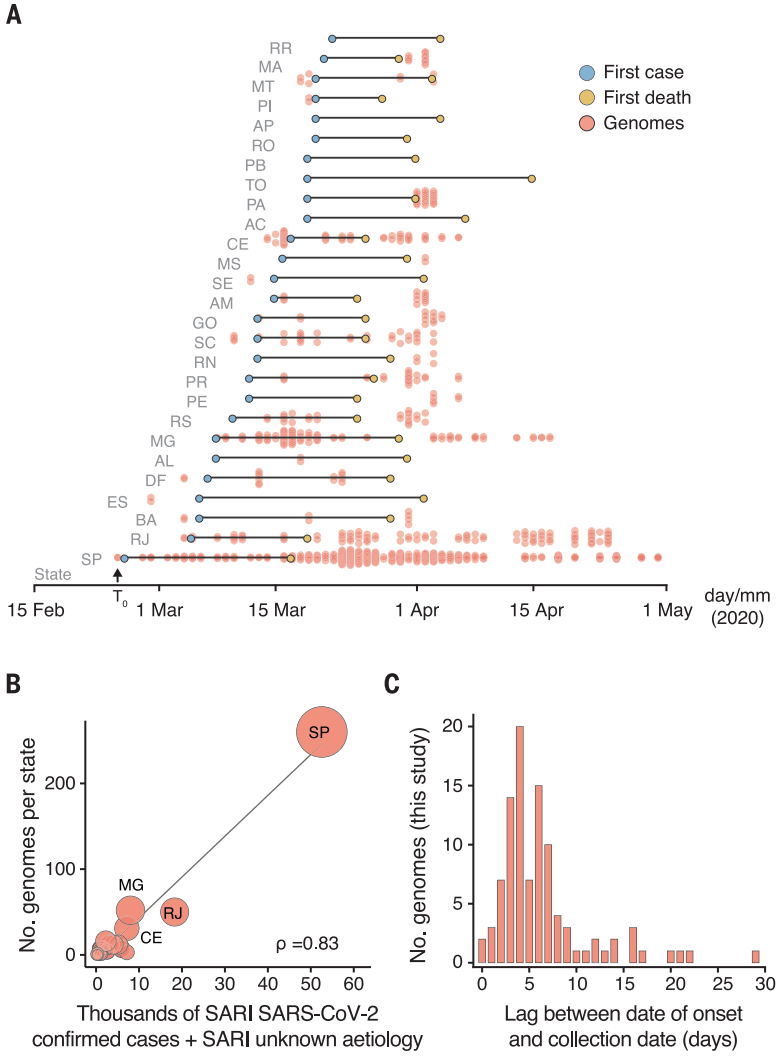


Fig. 2. Spatially representative genomic sampling. (**A**) Dumbbell plot showing the time intervals between date of collection of sampled genomes, notification of first cases, and first deaths in each state. Red lines indicate the lag between the date of collection of first genome sequence and first reported case. The key for the two-letter ISO 3166-1 codes for Brazilian federal units (or states) are provided in the supplementary materials. (**B**) Spearman's rank correlation between the number of SARI SARS-CoV-2 confirmed and SARI cases with unknown etiology against the number of sequences for each of the 21 Brazilian states included in this study (see also fig. S4). Circle sizes are proportional to the number of sequences for each federal unit. (**C**) Interval between the date of symptom onset and the date of sample collection for the sequences generated in this study.

Phylogenetic analyses and international introductions

We estimated maximum likelihood and molecular clock phylogenies for a global dataset with a total of 1182 genomes sampled from 24 December 2019 to 30 April 2020 (root-to-tip genetic distance correlation with sampling dates, r^2 = 0.53; Fig. 3A and fig. S7). We inferred a median evolutionary rate of 1.13×10^{-3} (95% BCI: 1.03 to 1.23×10^{-3}) substitutions per site per year using an exponential growth coalescent model, equating to 33 changes per year on average across the virus genome. This is within the range of evolutionary rates estimated for other human coronaviruses (*30*–*33*). We estimate the date of the common ancestor (TMRCA) of the SARS-CoV-2 pandemic to around mid-November 2019 (median = 19 November 2019, 95% BCI: 26 October 2019 to 6 December 2019), which is consistent with recent findings (*34*, *35*).

Phylogenetic analysis revealed that the majority of the Brazilian genomes (76%, n = 370/490) fell into three clades, hereafter referred to as Clade 1 (n = 186/490, 38% of Brazilian strains), Clade 2 (n = 166, 34%), and Clade 3 (n = 18/490, 4%) (Fig. 3A and figs. S8 and S9), which were largely in agreement with those identified in a phylogenetic analysis using 13,833 global genomes. The most recent common ancestors of the three main Brazilian clades (Clades 1 to 3) were dated from 28 February (21 February to 4 March 2020) (Clade 1),

Corrected 10 September 2020. See full text.

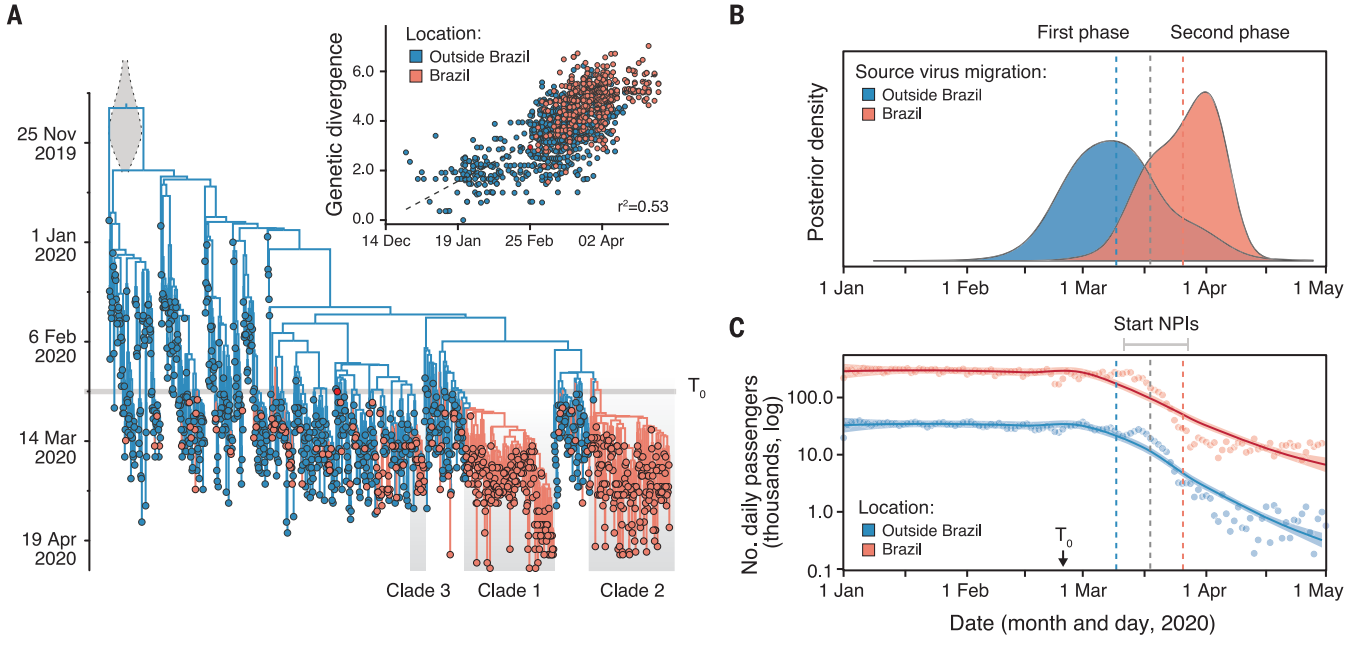


Fig. 3. Evolution and spread of SARS-CoV-2 in Brazil. (**A**) Time-resolved maximum clade credibility phylogeny of 1182 SARS-CoV-2 sequences, 490 of which are from Brazil (salmon) and 692 from outside of Brazil (blue). The largest Brazilian clades are highlighted by gray boxes (Clade 1, Clade 2, and Clade 3). Inset shows a root-to-tip regression of genetic divergence against dates of sample collection. Red tip corresponds to the first reported case in Brazil. (**B**) Dynamics of SARS-CoV-2 import events in Brazil. Dates of international and national (between federal states) migration events were estimated from virus genomes using a phylogeographic approach. The first phase was dominated by virus migrations from outside of Brazil, whereas the second phase was marked by virus spread within Brazil. Dashed vertical lines correspond to the mean posterior estimate for migration events from outside of Brazil (blue) and within Brazil (red). (**C**) Locally estimated scatterplot smoothing of the daily number of international (blue) and national (red) air passengers in Brazil in 2020. T_0, date of first reported case in Brazil (25 February 2020).

22 February (17 to 24 February 2020) (Clade 2), to 11 March (9 to 12 March 2020) (Clade 3) (Fig. 3A and fig. S10). This indicates that community-driven transmission was already established in Brazil by early March, suggesting that international travel restrictions initiated after this period would have had limited impact. Brazilian Clade 1 is characterized by a nucleotide substitution in the spike protein (G25088T, numbering relative to GenBank reference NC_045512.2) and circulates predominantly in São Paulo state (n = 159, 85.4%; figs. S9 and S11). Clade 2 is defined by two nucleotide substitutions in ORF6 (T27299C) and nucleoprotein (T29148C); this is the most spatially widespread lineage, with sequences from a total of 16 states in Brazil. Clade 3 is concentrated in Ceará state (n = 16, 89%) and falls in a global cluster with sequences mainly from Europe. In the Amazon region, where the epidemic is expanding rapidly (*14, 22*), we found evidence for multiple national and international introductions, with 37% (n = 7/19) of sequences from Pará and Amazonas states clustering in Clade 1 and 32% (n = 6/19) in Clade 2.

Time-measured phylogeographic analyses revealed at least 102 (95% BCI: 95 to 109) international introductions of SARS-CoV-2 in Brazil (Fig. 3A and figs. S8 and S12). This represents an underestimate of the real number of introductions because we sequenced, on average, only one out of 200 confirmed cases. Most of these estimated introductions were directed to internationally well-connected states (*36*) such as São Paulo (36% of all imports), Minas Gerais (24%), Ceará (10%), and Rio de Janeiro (8%) (fig. S12). We further assessed the contribution of international versus national virus lineage movement events through time (Fig. 3B). In the first phase of the epidemic, we found an increasing number of international introductions until 10 March 2020 (Fig. 2B). Limited available travel history data (*15*) suggested that these early cases were predominantly acquired from Italy (26%, n = 70 of 266 unambiguously identified country of infection) and the United States (28%, n = 76 of 266). After this initial phase, we found that the estimated number of international imports decreased concomitantly with the decline in the number of international passengers traveling to Brazil (Fig. 3, B and C, and S13). By contrast, despite the declines in the number of passengers traveling on national flights (Fig. 3C), we detected an increase in virus lineage movement events between Brazilian regions at least until early April 2020.

Modeling spatiotemporal spread within Brazil

To better understand virus spread across spatiotemporal scales within Brazil, we used a continuous phylogeographic model that maps phylogenetic nodes to their inferred origin locations (*37*) (Fig. 4). We distinguished branches that remain within a state versus those that cross a state to infer the proportion of within-state versus between-state observed virus movement.

We estimate that during the first epidemic phase, SARS-CoV-2 spread mostly locally and within state borders. By contrast, the second phase was characterized by long-distance movement events and the ignition of the epidemic outside of the southeast region of Brazil (Fig. 4A). Throughout the epidemic, we found that within-state virus lineage movement was, on average, 5.1-fold more frequent than between-state movement. Moreover, our data suggest that within-state virus spread and, to a lesser extent, between-state virus spread decreased after the implementation of NPIs (Fig. 4B). However, the more limited sampling after 6 April 2020 (see fig. S2) decreased inferred virus lineage movement to the present (Figs. 3B and 4B).

We found that the average route length traveled by passenger increased by 25% during the second phase of the epidemic (Fig. 4C) despite a concomitant reduction in the number of passengers flying within Brazil (Fig. 3C). The increase in the average route length after NPI implementation resulted from a larger reduction in the number of air passengers flying on shorter-distance journeys compared with those flying on longer-distance journeys. For example, we found an 8.8-fold reduction in

Corrected 10 September 2020. See full text.

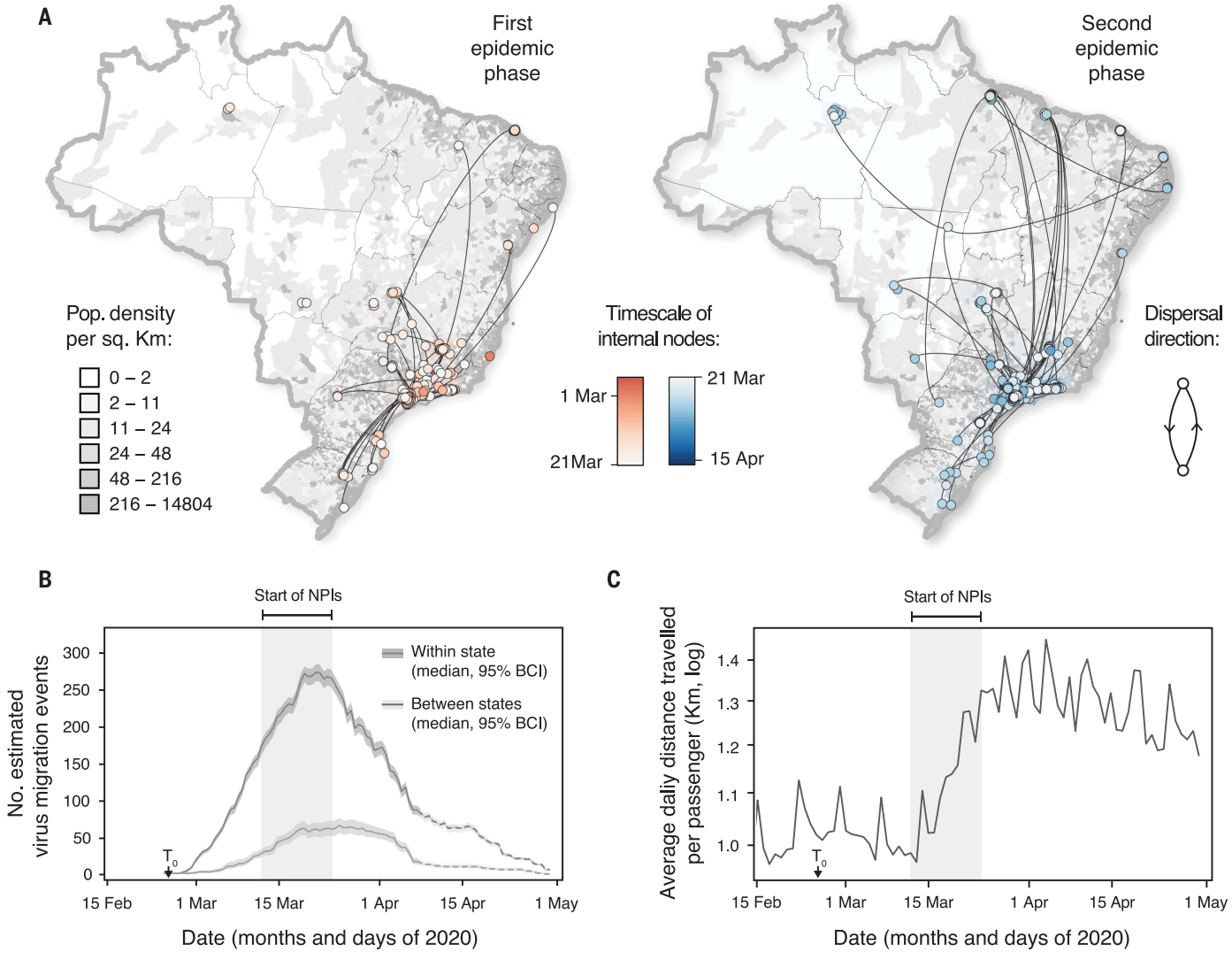


Fig. 4. Spread of SARS-CoV-2 in Brazil. (**A**) Spatiotemporal reconstruction of the spread of Brazilian SARS-CoV-2 clusters containing more than two sequences during the first (left) and the second (right) epidemic phase (Fig. 3B). Circles represent nodes of the maximum clade credibility phylogeny and are colored according to their inferred time of occurrence. Shaded areas represent the 80% highest posterior density interval and depict the uncertainty of the phylogeographic estimates for each node. Solid curved lines denote the links between nodes and the directionality of movement. Sequences belonging to clusters with fewer than three sequences were also plotted on the map with no lines connecting them. Background population density for each municipality was obtained from the Brazilian Institute of Geography (https://www.ibge.gov.br/). See fig. S14 for details of virus spread in the southeast region. (**B**) Estimated number of within-state (or within a given federal unit) and between-state (or between federal units) virus migrations over time. Dashed lines indicate estimates obtained during the period of limited sampling (fig. S2). (**C**) Average distance in kilometers traveled by an air passenger per day in Brazil. The number of daily air passengers is shown in Fig. 3B. Light gray boxes indicate the starting dates of NPIs across Brazil.

the number of passengers flying in flight legs <1000 km, compared with a 4.4-fold reduction in those flying >2000 km (fig. S15). These findings emphasize the roles of within- and between-state mobility as a key driver of both local and interregional virus spread, with highly populated and well-connected urban conurbations in the southeast region acting as the main sources of virus exports within the country (fig. S12).

Discussion

We provide a comprehensive analysis of SARS-CoV-2 spread in Brazil showing the importance of community- and nation-wide measures to control the COVID-19 epidemic in Brazil. Although NPIs initially reduced virus transmission and spread, the continued increase in the number of cases and deaths in Brazil highlights the urgent need to prevent future virus transmission by implementing rapid and accessible diagnostic screening, contact tracing, quarantining of new cases, and coordinated social and physical distancing measures across the country (*38*). With the recent relaxation of NPIs in Brazil and elsewhere, continued molecular, immunological, and genomic surveillance are required for real-time data-driven decisions. Our analysis shows how changes in mobility may affect global and local transmission of SARS-CoV-2 and demonstrates how combining genomic and mobility data can complement traditional surveillance approaches.

REFERENCES AND NOTES

1. F. Wu *et al.*, *Nature* **579**, 265–269 (2020).
2. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, *Nat. Med.* **26**, 450–452 (2020).
3. World Health Organization, *Coronavirus Disease (COVID-2019) Situation Reports* (2020); www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

4. H. Tian *et al*., *Science* **368**, 638–642 (2020).
5. M. U. G. Kraemer *et al*., *Science* **368**, 493–497 (2020).
6. A. Rambaut *et al*., *Nat. Microbiol.* (2020).
7. T. W. Russell *et al*., *Euro Surveill.* **25**, 2000256 (2020).
8. R. Verity *et al*., *Lancet Infect. Dis.* **20**, 669–677 (2020).
9. J. T. Wu *et al*., *Nat. Med.* **26**, 506–510 (2020).
10. M. M. Arons *et al*., *N. Engl. J. Med.* **382**, 2081–2090 (2020).
11. L. Ferretti *et al*., *Science* **368**, eabb6936 (2020).
12. E. Lavezzo *et al*., *Nature* (2020).
13. K. Mizumoto, K. Kagaya, A. Zarebski, G. Chowell, *Euro Surveill.* **25**, 2000180 (2020).
14. Brazilian Ministry of Health, *Painel de Casos de Doença Pelo Coronavírus 2019 (COVID-19) No Brasil Pelo Ministério da Saúde* (2020); http://covid.saude.gov.br.
15. W. M. de Souza *et al*., *Nat. Hum. Behav.* **4**, 856–865 (2020).
16. J. Croda *et al*., *Rev. Soc. Bras. Med. Trop.* **53**, e20200167 (2020).
17. J. Croda, L. Garcia, *Epidemiol. Ser. Saúde* **29**, e2020002 (2020).
18. S. B. Oliveira *et al*., Monitoring social distancing and SARS-CoV-2 transmission in Brazil using cell phone mobility data. medRxiv 2020.04.30.20082172 [Preprint] (5 May 2020); https://doi.org/10.1101/2020.04.30.20082172.
19. S. M. Kissler, Reductions in commuting mobility predict geographic differences in SARS-CoV-2 prevalence in New York City (Harvard DASH Repository, 2020); https://dash.harvard.edu/bitstream/handle/1/42665370/Kissler_etal_NYC_mobility.pdf?sequence = 1&isAllowed = y.
20. H. J. T. Unwin *et al*., *Report 23: State-Level Tracking of COVID-19 in the United States (21-05-2020)* (Imperial College London, 2020); https://doi.org/10.25561/79231.
21. S. Flaxman *et al*., *Nature* **584**, 257–261 (2020).
22. T. A. Mellan *et al*., *Report 21: Estimating COVID-19 Cases and Reproduction Number in Brazil* (2020); https://doi.org/10.25561/78872.
23. Y.-Z. Zhang, E. C. Holmes, Novel 2019 coronavirus genome, Virological (2020); https://virological.org/t/novel-2019-coronavirus-genome/319.
24. V. M. Corman *et al*., *Euro Surveill.* **25**, 2000045 (2020).
25. T. Thi Nhu Thao *et al*., *Nature* **582**, 561–565 (2020).
26. P. C. Resende *et al*., Genomic surveillance of SARS-CoV-2 reveals community transmission of a major lineage during the early pandemic phase in Brazil. bioRxiv 020.06.17.158006 [Preprint] (2020); https://doi.org/10.1101/2020.06.17.158006.
27. J. Xavier *et al*., *Emerg. Microbes Infect.* **9**, 1824–1834 (2020).
28. V. A. Nascimento *et al*., *Memoirs of the Oswaldo Cruz Institute* 10.1590/0074-02760200310 (2020).
29. Y. Shu, J. McCauley, *Euro. Surveill.* **22**, 30494 (2017).
30. M. Cotten *et al*., *Lancet* **382**, 1993–2002 (2013).
31. M. Cotten *et al*., *mBio* **5**, e01062-13 (2014).
32. G. Dudas, L. M. Carvalho, A. Rambaut, T. Bedford, *eLife* **7**, e31257 (2018).
33. Z. Zhao *et al*., *BMC Evol. Biol.* **4**, 21 (2004).
34. S. Duchene *et al*., Temporal signal and the phylodynamic threshold of SARS-CoV-2. bioRxiv 2020.05.04.077735 [Preprint] (2020); https://doi.org/10.1101/2020.05.04.077735.
35. J. Lu *et al*., *Cell* **181**, 997–1003.e9 (2020).
36. D. D. S. Candido *et al*., *J. Travel Med.* **27**, taaa042 (2020).
37. S. Dellicour *et al*., A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. bioRxiv 2020.05.05.078758 [Preprint] (2020); https://doi.org/10.1101/2020.05.05.078758.
38. World Health Organization, Coronavirus disease 2019 (COVID-19): Situation report –72 (WHO, 2020); https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200401-sitrep-72-covid-19.pdf?sfvrsn = 3dd8971b_2.
39. Centre for Genomic Pathogen Surveillance, Imperial College London, Report of 427 novel genomes from Brazil and the associated metadata, Microreact (2020); https://microreact.org/project/rKjKLMrjdPVHKR1erUzKyi.
40. Data and code for: D. S. Candido *et al*., Evolution and epidemic spread of SARS-CoV-2 in Brazil, Dryad (2020); https://doi.org/10.5061/dryad.rxwdbrv5z.

ACKNOWLEDGMENTS

A full list acknowledging those involved in the diagnostics and generation of new sequences as part of the CADDE-Genomic-Network can be found in the supplementary materials. We thank the administrators of the GISAID database for supporting rapid and transparent sharing of genomic data during the COVID-19 pandemic. A full list acknowledging the authors submitting data used in this study can be found in data S2. We thank P. Resende (FIOCRUZ), T. Adelino (FUNED), C. Sacchi (IAL), V. Nascimento (FIOCRUZ Amazonia), and their colleagues for submitting Brazilian data to GISAID; A. Pinter (SUCEN), N. Gouveia (USP), and I. Marcílio de Souza (HCFM-USP) for fruitful discussions; L. Matkin and J. Quick for logistic support; and the UNICAMP Task Force against Covid-19 for support in generating genomes from Campinas. The analysis of openly available epidemiological data from https://covid.saude.gov.br/ has benefited from the COVID-19 surveillance efforts by the Secretaria de Vigilância em Saúde, Ministry of Health, Brazil. **Funding:** This project was supported by a Medical Research Council-São Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0) (http://caddecentre.org/). FAPESP further supports I.M.C. (2018/17176-8 and 2019/12000-1), J.G.J. (2018/17176-8 and 2019/12000-1, 18/14389-0), F.C.S.S. (2018/25468-9), W.M.S. (2017/13981-0, 2019/24251-9), M.F. (2018/09383-3), T.M.C. (2019/07544-2), C.A.M.S. (2019/21301-5), H.I.N. (2018/14933-2), P.S.P. (16/18445-7), M.L.N. (20/04836-0), and J.L.M. (2020/04558-0 and 2016/00194-8). N.R.F. is supported by a Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z). D.S.C. is supported by the Clarendon Fund and by the Department of Zoology, University of Oxford. S.D. is supported by the Fonds National de la Recherche Scientifique (FNRS, Belgium). J.T. and P.L. are supported by European Union's Horizon 2020 project MOOD (874850). This project was supported by CNPq (M.T.M., M.L.N., and A.T.R.V.: 303170/2017-4; R.S.A.: 312688/2017-2 and 439119/2018-9; R.P.S.: 310627/2018-4; and W.M.S.: 408338/2018-0), FAPERJ (A.T.R.V.: E-26/202.826/2018 and R.S.A.: 202.922/2018). M.S.R. is supported by FMUSP. C.A.P., G.M.F., J.H., and M.R.A. are supported by CAPES. O.J.B. is supported by a Sir Henry Wellcome Fellowship funded by the Wellcome Trust (206471/Z/17/Z). R.P.S. is supported by FAPEMIG (APQ-00475-20). M.M.T. is supported by Instituto Nacional de Ciância e Tecnologia em Dengue (INCT Dengue 465425/2014-3). A.T.R.V. is supported by FINEP (01.16.0078.00). P.L. and N.J.L. are supported by the Wellcome Trust ARTIC network (collaborators award no. 206298/Z/17/Z). P.L. and A.R. are supported by the European Research Council (grant no. 725422 -ReservoirDOCS). O.G.P., N.R.F., and L.d.P. are supported by the Oxford Martin School. This work received funding from the U.K. Medical Research Council under a concordat with the U.K. Department for International Development. We additionally acknowledge support from Community Jameel and the NIHR Health Protection Research Unit in Modelling Methodology. **Author contributions:** Conceptualization: D.S.C., I.M.C., J.G.J., E.C.S., N.R.F.; Formal analysis: D.S.C., I.M.C., J.G.J., W.M.S., F.R.R.M., S.D., T.A.M., LP, R.H.M.P., J.T., L.A., C.M.V., H.H., S.M., M.S.G., L.M.C., L.F.B., C.A.P., O.J.B., S.M.N., S.C.H., J.L.P.M., A.T.R.V., S.B., O.G.P., P.L., C.H.W., R.S.A., N.R.F.; Investigation: D.S.C., I.M.C., J.G.J., W.M.S., F.R.R.M., R.H.M.P., F.C.S.S., E.R.M., M.T.M., C.M.V., M.J.F., T.M.C., C.A.M.S., M.S.R., M.R.A., J.A., H.N., P.S.P., A.T., A.D.R., C.K.V.B., A.L.G., A.P.G., N.G., C.S.A., A.C.S.F., C.X.L., J.E.L., C.G., G.M.F., R.S.F., F.G., M.T.G., M.L.M., M.W.P., T.M.P.P.C., C.S.L., A.A.S.S., C.L.S., J.F., A.C.S., A.Z.S., M.N.N.S., C.Z.S., R.P.S., L.C.R.M., M.M.T., J.H., P.A.F.L., R.G.M., M.L.N., S.F.C., J.L.P.M., A.T.R.V., R.S.A., E.C.S., N.R.F.; Interpretation: D.S.C., I.M.C., J.G.J., W.M.S., F.R.R.M., S.D., T.A.M., L.P., R.H.M.P., S.C.H., A.A.S.S., N.M.F., A.T.R.V., S.B., P.L., C.H.W., A.R., R.S.A., O.G.P., E.C.S., N.R.F.; Writing – original draft: D.S.C., I.M.C., J.G.J., W.M.S., F.R.R.M., S.D., T.A.M., R.S.A., O.G.P., E.C.S., N.R.F.; Writing – review & editing: All authors have read and approved the final version of the manuscript. Funding acquisition: W.M.S., M.L.N., N.M.F., J.L.P.M., A.T.R.V., N.J.L., R.S.A., O.G.P., E.C.S., N.R.F. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The 427 SARS-CoV newly generated genomes from this study can be found on GISAID under the accession IDs: EPI_ISL_470568-470655 and EPI_ISL_476152-476490. An interactive visualization of the temporal, geographic and mutational patterns in our data can be found at https://microreact.org/project/rKjKLMrjdPVHKR1erUzKyi (*39*). Reads have been deposited to accession numbers PRJEB39487 (IMT-USP and UNICAMP) and PRJNA640656 (UFRJ-LNCC). All data, code, and materials used in the analysis are available on DRYAD (*40*). The IRB protocol number is CAAE 30127020.0.0000.0068 as described in the materials and methods. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1255/suppl/DC1
Materials and Methods
Figs. S1 to S15
Tables S1 to S3
List of Members of the CADDE Genomic Network
References (*41*–*77*)
Data S1 and S2
MDAR Reproducibility Checklist

10 June 2020; accepted 16 July 2020
Published online 23 July 2020
10.1126/science.abd2161

Science

Evolution and epidemic spread of SARS-CoV-2 in Brazil

Darlan S. CandidoIngra M. ClaroJaqueline G. de JesusWilliam M. SouzaFilipe R. R. MoreiraSimon DellicourThomas A. MellanLouis du PlessisRafael H. M. PereiraFlavia C. S. SalesErika R. ManuliJulien ThézéLuiz AlmeidaMariane T. MenezesCarolina M. VolochMarcilio J. FumagalliThaís M. ColettiCamila A. M. da SilvaMariana S. RamundoMariene R. AmorimHenrique H. HoeltgebaumSwapnil MishraMandev S. GillLuiz M. CarvalhoLewis F. BussCarlos A. Prete Jr.Jordan AshworthHelder I. NakayaPedro S. PeixotoOliver J. BradySamuel M. NicholsAmilcar TanuriÁtila D. RossiCarlos K. V. BragaAlexandra L. GerberAna Paula de C. GuimarãesNelson Gaburo Jr.Cecila Salete AlencarAlessandro C. S. FerreiraCristiano X. LimaJosé Eduardo LeviCelso GranatoGiulia M. FerreiraRonaldo S. Francisco Jr.Fabiana GranjaMarcia T. GarciaMaria Luiza MorettiMauricio W. Perroud Jr.Terezinha M. P. P. CastiñeirasCarolina S. LazariSarah C. HillAndreza Aruska de Souza SantosCamila L. SimeoniJulia ForatoAndrei C. SpositoAngelica Z. SchreiberMagnun N. N. SantosCamila Zolini de SáRenan P. SouzaLuciana C. Resende-MoreiraMauro M. TeixeiraJosy HubnerPatricia A. F. LemeRennan G. MoreiraMaurício L. NogueiraNeil M. FergusonSilvia F. CostaJosé Luiz Proenca-ModenaAna Tereza R. VasconcelosSamir BhattPhilippe LemeyChieh-Hsi WuAndrew RambautNick J. LomanRenato S. AguiarOliver G. PybusEster C. SabinoNuno Rodrigues Faria

Science, 369 (6508), • DOI: 10.1126/science.abd2161

The spread of SARS-CoV-2 in Brazil

Brazil has been hard-hit by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic. Candido *et al.* combined genomic and epidemiological analyses to investigate the impact of nonpharmaceutical interventions (NPIs) in the country. By setting up a network of genomic laboratories using harmonized protocols, the researchers found a 29% positive rate for SARS-CoV-2 among collected samples. More than 100 international introductions of SARS-CoV-2 into Brazil were identified, including three clades introduced from Europe that were already well established before the implementation of NPIs and travel bans. The virus spread from urban centers to the rest of the country, along with a 25% increase in the average distance traveled by air passengers before travel bans, despite an overall drop in short-haul travel. Unfortunately, the evidence confirms that current interventions remain insufficient to keep virus transmission under control in Brazil.

Science, this issue p. 1255

Chapter 4

Genomic epidemiology of variants of concern (VOCs) in Brazil

December and January 2020 were marked by the identification of the first VOCs (Alpha in the UK, Beta in South Africa and Gamma in Brazil) followed by a global increase the incidence of SARS-CoV-2 cases and deaths. This chapter covers two publications on two of these new variants. **Chapter 4.1** describes genomic and epidemiological aspects of the first confirmed cases of the Alpha VOC in Brazil. It was first made available as a virological.org post on the 31st December 2020 and published in the Journal of Emerging Infectious Diseases. It is presented here fully. **Chapter 4.2** resulted from an incredible collaborative effort to describe and understand the emergence of the Gamma VOC in Manaus, Brazil, amidst an upsurge of cases in the region. I was responsible/involved in all aspects of the genomic epidemiology of this paper, and, as such, it is only partly presented in this thesis. It was first made available as a preprint on MedRxiv on the 3rd of March 2021 and subsequently published in *Science*.

Claro IM, da Silva Sales FC, Ramundo MS, ***Candido DS***, Silva CAM, de Jesus JG, et al. Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020. Emerging Infect Dis. 2021 Mar;27(3):970–2.

Faria NR*, Mellan TA*, Whittaker C*, Claro IM*, ***Candido D da S****, Mishra S*, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 2021 May 21;372(6544):815–21.

Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020

Ingra Morales Claro,[1] Flavia Cristina da Silva Sales,[1] Mariana Severo Ramundo, Darlan S. Candido, Camila A.M. Silva, Jaqueline Goes de Jesus, Erika R. Manuli, Cristina Mendes de Oliveira, Luciano Scarpelli, Gustavo Campana, Oliver G. Pybus, Ester Cerdeira Sabino,[2] Nuno Rodrigues Faria,[2] José Eduardo Levi[2]

Author affiliations: University of São Paulo, São Paulo, Brazil (I.M. Claro, F.C.S. Sales, M.S. Ramundo, D.S. Candido, C.A.M. Silva, J.G. de Jesus, E.R. Manuli, E.C. Sabino, N.R. Faria, J.E. Levi); University of Oxford, Oxford, UK (D.S. Candido, O.G. Pybus, N.R. Faria); Diagnósticos da América SA (DASA), Baueri, Brazil (C.M. de Oliveira, L. Scarpelli, G. Campana, J.E. Levi); Imperial College London, London, UK (N.R. Faria)

DOI: https://doi.org/10.3201/eid2703.210038

In December 2020, research surveillance detected the B.1.1.7 lineage of severe acute respiratory syndrome coronavirus 2 in São Paulo, Brazil. Rapid genomic sequencing and phylogenetic analysis revealed 2 distinct introductions of the lineage. One patient reported no international travel. There may be more infections with this lineage in Brazil than reported.

Genomic sequencing and analysis during the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic have led to identification of ≈800 distinct SARS-CoV-2 lineages worldwide. A new phylogenetic cluster, B.1.1.7 lineage or variant of concern 202012/01, is characterized by 17 unique mutations and was first detected in southeastern England in late September 2020 (A. Rambaut et al., unpub. data, https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563). As of January 17, 2021, this lineage had been confirmed in 38 countries (https://cov-lineages.org/global_report_B.1.1.7.html). Epidemiologic and phylogenetic studies suggest that the rapid epidemic growth of B.1.1.7 in the United Kingdom is caused by its higher transmissibility (E. Volz et al., unpub. data, https://www.medrxiv.org/content/10.1101/2020.12.30.20249034v2; N. Davies, unpub. data, https://cmmid.github.io/topics/covid19/uk-novel-variant.html), which could lead to increased incidence and higher peaks in hospitalizations and deaths (N. Davies, unpub. data, https://cmmid.github.io/topics/covid19/uk-novel-variant.html).

We confirm 2 cases of infection with SARS-CoV-2 B.1.1.7 lineage in Latin America. On December 30, 2020, we received saliva samples from 2 patients for genomic sequencing as part of research surveillance activities. Patient 1 was a woman 20–30 years of age residing in São Paulo, Brazil, who reported no travel outside of Brazil. Her symptoms began on December 21, and testing was conducted the next day. Patient 2 was a man 30–40 years of age who was tested in São Paulo on December 22 after having traveled from London on December 19. Ethics approval for this study was confirmed by the national ethics review board (Comissão Nacional de Ética em Pesquisa, protocol no. CAAE 30127020.0.0000.0068).

PCR testing (TaqPath COVID-19 PCR; ThermoFisher Scientific, https://www.thermofisher.com) performed as previously described (*1*) indicated that patient 1 was positive for the open reading frame 1ab (cycle threshold [C_t] 25.8) and nucleoprotein (C_t 24.5) gene targets and patient 2 was positive for open reading frame 1ab (C_t 28.1) and nucleoprotein (C_t 27.29), but both were negative for the spike gene target. The 2 spike-gene dropout samples were identified among 400 samples collected during November 4–December 25, 2020.

For each sample, we conducted nanopore sequencing in duplicate by using the ARTIC protocol (https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w). Concentrations of double-stranded DNA for the library-negative controls were below detection levels, indicating no contamination. We conducted whole-genome sequencing of SARS-CoV-2 by using the MinION platform (Oxford Nanopore Technologies, https://nanoporetech.com). By December 31, sequencing statistics revealed 56,565 mapped reads for patient 1 and 51,761 for patient 2. Consequently, 28,023 bases for patient 1 and 26,339 for patient 2 were covered at >25× depth. Consensus sequences covered 92.4% of the Wuhan Hu-1 reference genome (GenBank accession no. MN908947.3) for patient 1 and 87.1% for patient 2. For the 2 newly generated genome sequences, we identified the B.1.1.7 lineage (assignment probability = 1.0) by using the pangolin COVID-19 Lineage Assigner version 2.1.6 (*2*) (https://pangolin.cog-uk.io). The 2 genomes were made publicly available on GISAID (http://www.gisaid.org) on December 31, 2020 (identification nos. EPI_ISL_754236 for patient 1 and EPI_ISL_754237 for patient 2).

We next estimated a rapid maximum-likelihood phylogenetic tree (*3*,*4*) for a multiple sequence align-

[1]These first authors contributed equally to this article.

[2]These senior authors contributed equally to this article.

ment (*5*) with the new sequences and 127 publicly available B.1.1.7 genomes from around the world available on GISAID (*6*) as of December 31, 2020 (https://github.com/CADDE-CENTRE/VOC-Lineage-Brazil). The virus genome recovered from patient 1 grouped within a well-supported cluster (bootstrap 85%) of 10 sequences (60% from the United Kingdom) (Figure). This finding is consistent with the travel history of an asymptomatic family member who was positive for SARS-CoV-2 (according to a rapid test performed on December 23, 2020), who arrived in São Paulo on December 17 after traveling from Italy to the United Kingdom and, after a short stay, from London to São Paulo, and who was in close contact with patient 1. The sequence from patient 2 clustered with good statistical support (bootstrap 79.4%) with a sequence collected in the United Kingdom on November 27. Patient 2 had traveled from London to São Paulo on December 19 and was symptomatic when saliva was collected on December 22. Phylogenetic analysis suggests that this infection represents a second, independent introduction of the B.1.1.7 lineage from the United Kingdom to Brazil; patient 2 was not epidemiologically linked to patient 1.

Because information about this lineage from locations outside the United Kingdom is limited, our interpretations based on phylogenetic data might be biased by the different numbers of available genome sequences shared around the globe. Moreover, the samples that we analyzed were selected from only 2 cases confirmed by reverse transcription PCR in São Paulo; thus, our genomes were obtained from a small fraction of targeted spike-gene failure, and frequency of detection in our nonrandom sample does not represent prevalence of this lineage at the population level.

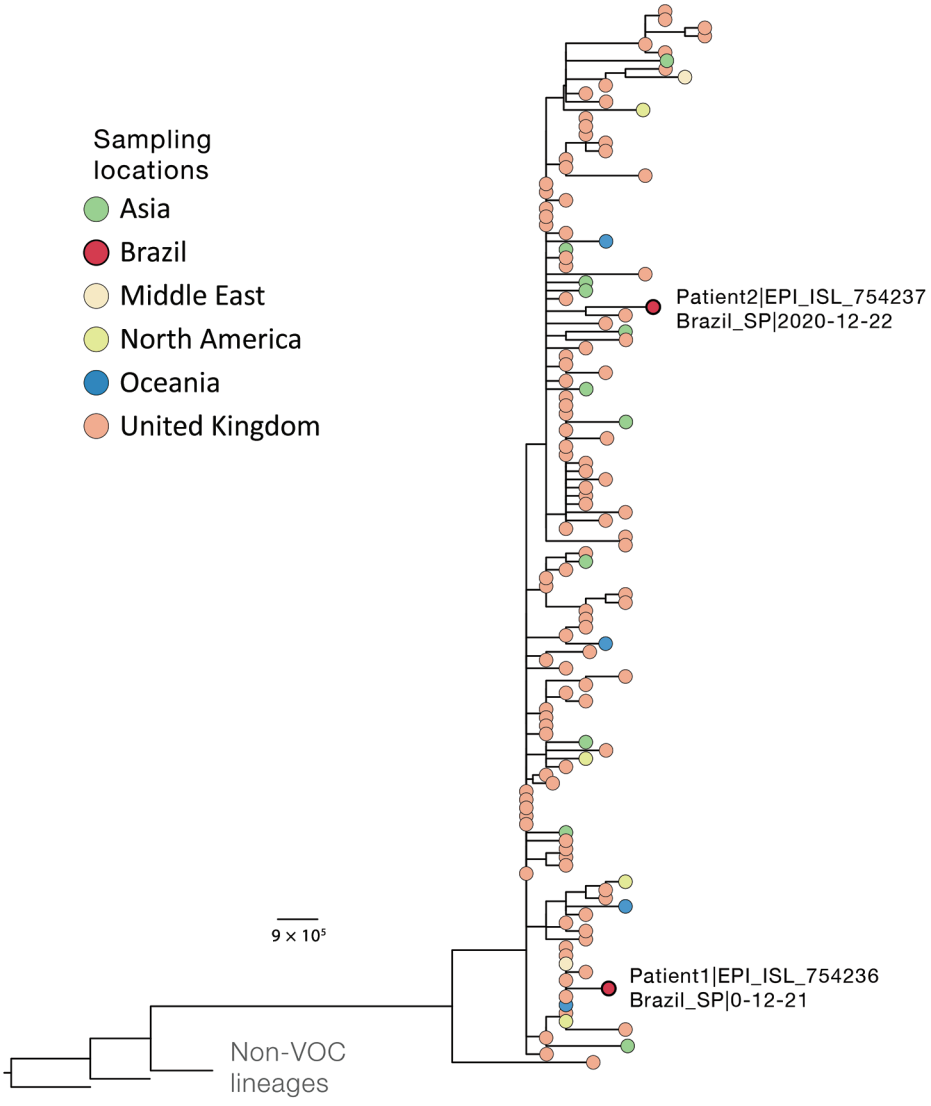


Figure. Phylogenetic context of novel severe acute respiratory syndrome coronavirus 2 B.1.1.7 genomes isolated from 2 patients in Brazil (labeled on figure), December 2020. Downsampling for the phylogenetic analysis of the B.1.1.7 SARS-CoV-2 variant (n = 4,693, December 31, 2020) was performed by selecting 1 sequence per country per day. As outgroups, we included 2 B.1.1 sequences from the United Kingdom that were closely related to the lineage of interest and sequence WH04 from Wuhan, China (GISAID identification no. EPI_ISL_406801; http://www.gisaid.org). Details on multiple alignment and phylogenetic tree reconstruction are described elsewhere (*4*). Tree file, aligned sequences, and GISAID acknowledgment tables are available at https://github.com/CADDE-CENTRE/VOC-Lineage-Brazil. Scale bar indicates nucleotide substitutions per site. VOC, variant of concern.

Despite temporary suspension of all flights to or from Brazil from or through the United Kingdom as of December 25, 2020 (http://www.gov.uk/foreign-travel-advice/brazil), it is likely that the number of SARS-CoV-2 lineage B.1.1.7 infections in Brazil is higher than that reported. Increasing genomic surveillance of B.1.1.7 and other variants of concern that carry mutations of potential biological significance (e.g., E484K in the spike protein; C.M. Voloch, unpub data, https://www.medrxiv.org/content/10.1101/2020.12.23.20248598v1) is imperative for monitoring vaccination effectiveness and contextualizing the epidemiology and evolution of SARS-CoV-2 in Latin America.

Acknowledgments

We thank all researchers who are working around the clock to generate and share genome data worldwide on GISAID (http://www.gisaid.org). GISAID acknowledgment tables are available at https://github.com/CADDE-CENTRE/VOC-Lineage-Brazil.

This project was supported by a Medical Research Council-São Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0) (http://caddecentre.org/). N.R.F. is supported by a Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z).

About the Author

Mrs. Claro is a PhD student at the Department of Infectious Disease, School of Medicine & Institute of Tropical Medicine, University of São Paulo, São, Paulo, Brazil. She is a part of the Centre for Arbovirus Discovery, Diagnostics, Genomics and Epidemiology team and has primary research interests in real-time epidemiologic surveillance of viruses of public importance for Brazil.

References

1. Vogels CBF, Watkins AE, Harden CA, Brackney DE, Shafer J, Wang J, et al. SalivaDirect: a simplified and flexible platform to enhance SARS-CoV-2 testing capacity. Med. 2020 Dec 26 [Epub ahead of print]. https://doi.org/10.1016/j.medj.2020.12.010
2. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5:1403–7. https://doi.org/10.1038/s41564-020-0770-5
3. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4. https://doi.org/10.1093/molbev/msaa015
4. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020;369:1255–60. https://doi.org/10.1126/science.abd2161
5. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. https://doi.org/10.1093/molbev/mst010
6. Shu Y, McCauley J. GISAID: Global Initiative on Sharing All Influenza Data—from vision to reality. Euro Surveill. 2017;30;22:30494.

Address for correspondence: Nuno R. Faria, St Mary's Hospital, Praed St, Paddington, London W2 1NY, UK; email: nfaria@ic.ac.uk; and José Eduardo Levi, Avenida Juruá 548, Alphaville, Barueri, SP 06455-010, Brazil; email: jose.levi.ext@dasa.com.br

Mycobacterium bovis Pulmonary Tuberculosis, Algeria

Fatah Tazerart, Jamal Saad, Abdellatif Niar, Naima Sahraoui,[1] Michel Drancourt[1]

Author affiliations: Université Ibn Khaldoun de Tiaret, Tiaret, Algeria (F. Tazerart); Université de Blida, Blida, Algeria (F. Tazerart, N. Sahraoui); Institut Hospitalo-Universitaire Méditerranée Infection, Marseille, France (F. Tazerart, J. Saad, M. Drancourt); Aix-Marseille-University, Marseille (J. Saad, M. Drancourt); Laboratoire de Reproduction des Animaux de la Ferme, Université Ibn Khaldoun de Tiaret, Tiaret (A. Niar)

DOI: https://doi.org/10.3201/eid2703.191823

We analyzed 98 *Mycobacterium tuberculosis* complex isolates collected in 2 regions of Algeria in 2015–2018 from 93 cases of pulmonary tuberculosis. We identified 93/98 isolates as *M. tuberculosis* lineage 4 and 1 isolate as *M. tuberculosis* lineage 2 (Beijing). We confirmed 4 isolates as *M. bovis* by whole-genome sequencing.

In Algeria, interpreting tuberculosis (TB) incidence, estimated at 53–88 cases/100,000 population in 2017 (*1*), is limited by the fact that the diagnosis relies on microscopic examination of clinical samples. Iso-

[1]These authors equally contributed to this work.

CORONAVIRUS

Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil

Nuno R. Faria[1,2,3,4]*†, Thomas A. Mellan[1,2]†, Charles Whittaker[1,2]†, Ingra M. Claro[3,5]†, Darlan da S. Candido[3,4]†, Swapnil Mishra[1,2]†, Myuki A. E. Crispim[6,7], Flavia C. S. Sales[3,5], Iwona Hawryluk[1,2], John T. McCrone[8], Ruben J. G. Hulswit[9], Lucas A. M. Franco[3,5], Mariana S. Ramundo[3,5], Jaqueline G. de Jesus[3,5], Pamela S. Andrade[10], Thais M. Coletti[3,5], Giulia M. Ferreira[11], Camila A. M. Silva[3,5], Erika R. Manuli[3,5], Rafael H. M. Pereira[12], Pedro S. Peixoto[13], Moritz U. G. Kraemer[4], Nelson Gaburo Jr.[14], Cecilia da C. Camilo[14], Henrique Hoeltgebaum[15], William M. Souza[16], Esmenia C. Rocha[3,5], Leandro M. de Souza[3,5], Mariana C. de Pinho[3,5], Leonardo J. T. Araujo[17], Frederico S. V. Malta[18], Aline B. de Lima[18], Joice do P. Silva[18], Danielle A. G. Zauli[18], Alessandro C. de S. Ferreira[18], Ricardo P. Schnekenberg[19], Daniel J. Laydon[1,2], Patrick G. T. Walker[1,2], Hannah M. Schlüter[15], Ana L. P. dos Santos[20], Maria S. Vidal[20], Valentina S. Del Caro[20], Rosinaldo M. F. Filho[20], Helem M. dos Santos[20], Renato S. Aguiar[21], José L. Proença-Modena[22], Bruce Nelson[23], James A. Hay[24,25], Mélodie Monod[15], Xenia Miscouridou[15], Helen Coupland[1,2], Raphael Sonabend[1,2], Michaela Vollmer[1,2], Axel Gandy[15], Carlos A. Prete Jr.[26], Vitor H. Nascimento[26], Marc A. Suchard[27], Thomas A. Bowden[9], Sergei L. K. Pond[28], Chieh-Hsi Wu[29], Oliver Ratmann[15], Neil M. Ferguson[1,2], Christopher Dye[4], Nick J. Loman[30], Philippe Lemey[31], Andrew Rambaut[8], Nelson A. Fraiji[6,32], Maria do P. S. S. Carvalho[6,33], Oliver G. Pybus[4,34]‡, Seth Flaxman[15]‡, Samir Bhatt[1,2,35]*‡, Ester C. Sabino[3,5]*‡

Cases of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection in Manaus, Brazil, resurged in late 2020 despite previously high levels of infection. Genome sequencing of viruses sampled in Manaus between November 2020 and January 2021 revealed the emergence and circulation of a novel SARS-CoV-2 variant of concern. Lineage P.1 acquired 17 mutations, including a trio in the spike protein (K417T, E484K, and N501Y) associated with increased binding to the human ACE2 (angiotensin-converting enzyme 2) receptor. Molecular clock analysis shows that P.1 emergence occurred around mid-November 2020 and was preceded by a period of faster molecular evolution. Using a two-category dynamical model that integrates genomic and mortality data, we estimate that P.1 may be 1.7- to 2.4-fold more transmissible and that previous (non-P.1) infection provides 54 to 79% of the protection against infection with P.1 that it provides against non-P.1 lineages. Enhanced global genomic surveillance of variants of concern, which may exhibit increased transmissibility and/or immune evasion, is critical to accelerate pandemic responsiveness.

Brazil has experienced high mortality during the COVID-19 pandemic, recording >300,000 deaths and >13 million reported cases, as of March 2021. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection and disease burden have been highly variable across the country, with the state of Amazonas in north Brazil being the worst-affected region (*1*). Serological surveillance of blood donors in Manaus, the capital city of Amazonas and the largest city in the Amazon region, has suggested >67% cumulative attack rates by October 2020 (*2*). Similar but slightly lower seroprevalences have also been reported for cities in neighboring regions (*3*, *4*). However, the level of previous infection in Manaus was clearly not sufficient to prevent a rapid resurgence in SARS-CoV-2 transmission and mortality there during late 2020 and early 2021 (*5*), which has placed substantial pressure on the city's health care system.

Here, we show that the second wave of infection in Manaus was associated with the emergence and rapid spread of a new SARS-CoV-2 lineage of concern, named lineage P.1. The lineage carries a distinctive constellation of mutations (table S1), including several that have been previously determined to be of virological importance (*6*–*10*) and that are located in the spike protein receptor binding domain (RBD), the region of the virus involved in recognition of the angiotensin-converting enzyme-2 (ACE2) cell surface receptor (*11*). Using genomic data, structure-based mapping of mutations of interest onto the spike protein, and dynamical epidemiology modeling of genomic and mortality data, we investigated the emergence of the P.1 lineage and explored epidemiological explanations for the resurgence of COVID-19 in Manaus.

Identification and nomenclature of the P.1 lineage in Manaus

In late 2020, two SARS-CoV-2 lineages of concern were discovered through genomic surveillance, both characterized by sets of notable mutations: lineage B.1.351, first reported in South Africa (*12*), and lineage B.1.1.7, detected in the UK (*13*). Both variants have transmitted rapidly in the countries where they were discovered and spread to other regions (*14*, *15*). Analyses indicate that B.1.1.7 has higher transmissibility and causes more severe illness as compared with those of previously circulating lineages in the UK (*1*, *16*, *17*).

After a rapid increase in hospitalizations in Manaus caused by severe acute respiratory infection (SARI) in December 2020 (Fig. 1A), we focused ongoing SARS-CoV-2 genomic surveillance (*2*, *18*–*22*) on recently collected samples from the city (supplementary materials, materials and methods, and table S2). Before this, only seven SARS-CoV-2 genome sequences from Amazonas were publicly available (SARS-CoV-2 was first detected in Manaus

[1]MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK. [2]The Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), School of Public Health, Imperial College London, London, UK. [3]Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [4]Department of Zoology, University of Oxford, Oxford, UK. [5]Departamento de Moléstias Infecciosas e Parasitárias, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [6]Fundação Hospitalar de Hematologia e Hemoterapia, Manaus, Brazil. [7]Diretoria de Ensino e Pesquisa, Fundação Hospitalar de Hematologia e Hemoterapia, Manaus, Brazil. [8]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. [9]Division of Structural Biology, Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [10]Departamento de Epidemiologia, Faculdade de Saúde Pública da Universidade de São Paulo, Sao Paulo, Brazil. [11]Laboratório de Virologia, Instituto de Ciências Biomédicas, Universidade Federal de Uberlândia, Uberlândia, Brazil. [12]Institute for Applied Economic Research–Ipea, Brasília, Brazil. [13]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil. [14]DB Diagnósticos do Brasil, São Paulo, Brazil. [15]Department of Mathematics, Imperial College London, London, UK. [16]Virology Research Centre, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil. [17]Laboratory of Quantitative Pathology, Center of Pathology, Adolfo Lutz Institute, São Paulo, Brazil. [18]Instituto Hermes Pardini, Belo Horizonte, Brazil. [19]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. [20]CDL Laboratório Santos e Vidal, Manaus, Brazil. [21]Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [22]Laboratory of Emerging Viruses, Department of Genetics, Evolution, Microbiology, and Immunology, Institute of Biology, University of Campinas (UNICAMP), São Paulo, Brazil. [23]Instituto Nacional de Pesquisas da Amazônia, Manaus, Brazil. [24]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [25]Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [26]Departamento de Engenharia de Sistemas Eletrônicos, Escola Politécnica da Universidade de São Paulo, São Paulo, Brazil. [27]Department of Biomathematics, Department of Biostatistics, and Department of Human Genetics, University of California, Los Angeles, CA, USA. [28]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA. [29]Mathematical Sciences, University of Southampton, Southampton, UK. [30]Institute for Microbiology and Infection, University of Birmingham, Birmingham, UK. [31]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium. [32]Diretoria Clínica, Fundação Hospitalar de Hematologia e Hemoterapia do Amazonas, Manaus, Brazil. [33]Diretoria da Presidência, Fundação Hospitalar de Hematologia e Hemoterapia do Amazonas, Manaus, Brazil. [34]Department of Pathobiology and Population Sciences, The Royal Veterinary College, London, UK. [35]Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark.

†These authors contributed equally to this work. ‡These authors contributed equally to this work.

***Corresponding author. Email: n.faria@imperial.ac.uk (N.R.F.); samir.bhatt@sund.ku.dk (S.B.); sabinoec@usp.br (E.S.C.)**

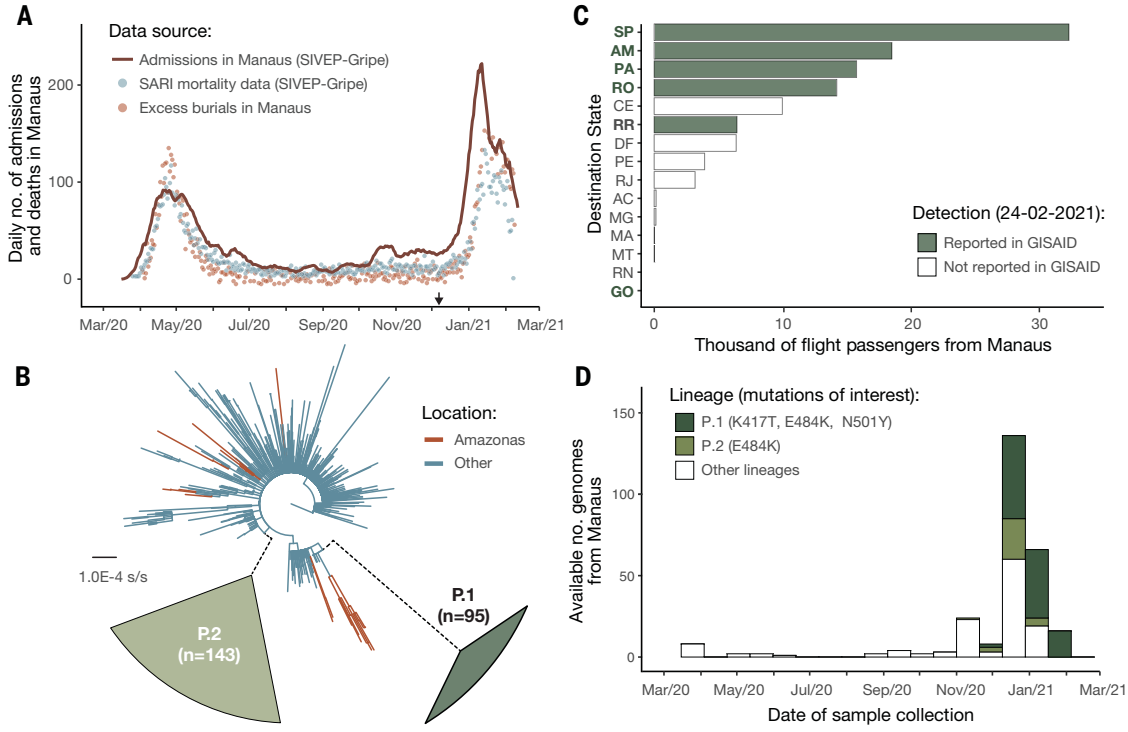


Fig. 1. SARS-CoV-2 epidemiological, diagnostic, genomic, and mobility data from Manaus. (**A**) Dark solid line shows the 7-day rolling average of the COVID-19 confirmed and suspected daily time series of hospitalizations in Manaus. Admissions in Manaus are from Fundação de Vigilância em Saúde do Amazonas (*66*). Green dots indicate daily severe acute respiratory mortality records from the SIVEP-Gripe (Sistema de Informação de Vigilância Epidemiológica da Gripe) database (*67*). Red dots indicate excess burial records based on data from Manaus Mayor's office for comparison (supplementary materials, materials and methods). The arrow indicates 6 December 2020, the date of the first P.1 case identified in Manaus by our study. (**B**) Maximum likelihood tree (n = 962 viral genomes) with B.1.1.28, P.1, and P.2 sequences, with collapsed views of P.1 and P.2 clusters and highlighting other sequences from Amazonas state, Brazil. Ancestral branches leading to P.1 and P.2 are shown as dashed lines. A more detailed phylogeny is available in fig. S3. Scale bar is shown in units of nucleotide substitutions per site (s/s). (**C**) Number of air travel passengers from Manaus to all states in Brazil was obtained from National Civil Aviation Agency of Brazil (www.gov.br/anac). The ISO 3166-2:BR codes of the states with genomic reports of P.1 [GISAID (*68*), as of 24 February 2021], are shown in bold. An updated list of GISAID genomes and reports of P.1 worldwide is available at https://cov-lineages.org/global_report_P.1.html. (**D**) Number of genome sequences from Manaus belonging to lineages of interest (supplementary materials, materials and methods). Spike mutations of interest are denoted.

on 13 March 2020) (*19*, *23*). We sequenced SARS-CoV-2 genomes from 184 samples from patients seeking COVID-19 testing in two diagnostic laboratories in Manaus between November and December 2020, using the ARTIC V3 multiplexed amplicon scheme (*24*) and the MinION sequencing platform. Because partial genome sequences can provide useful epidemiological information, particularly regarding virus genetic diversity and lineage composition (*25*), we harnessed information from partial (n = 41 viral sequences, 25 to 75% genome coverage), as well as near-complete (n = 95 viral sequences, 75 to 95%) and complete (n = 48 viral sequences, ≥95%) sequences from Manaus (figs. S1 to S4), together with other available and published genomes from Brazil for context. Viral lineages were classified by using the Pangolin (*26*) software tool (http://pangolin.cog-uk.io), nextclade (https://clades.nextstrain.org), and standard phylogenetic analysis using complete reference genomes.

Our early data indicated the presence of a novel SARS-CoV-2 lineage in Manaus that contained 17 amino acid changes (including 10 in the spike protein), three deletions, four synonymous mutations, and a four–base-pair nucleotide insertion compared with the most closely related available sequence (GISAID ID: EPI_ISL_722052) (Fig. 1B; lineage-defining mutations can be found in table S1) (*27*). This lineage was given a new designation, P.1, on the basis that (i) it is phylogenetically and genetically distinct from ancestral viruses, (ii) associated with rapid spread in a new area, and (iii) carries a constellation of mutations that may have phenotypic relevance (*26*). Phylogenetic analysis indicated that P.1—and another lineage, P.2 (*19*)—were descendants of lineage B.1.1.28 that was first detected in Brazil in early March 2020 (Fig. 1B). Our preliminary results were shared with local teams on 10 January 2021 and published online on 12 January 2021 (*27*). Concurrently, cases of SARS-CoV-2 P.1 infection were reported in Japan in travelers from Amazonas (*28*). As of 24 February 2021, P.1 had been confirmed in six Brazilian states, which in total received >92,000 air passengers from Manaus in November 2020 (Fig. 1C). Genomic surveillance first detected lineage P.1 on 6 December 2020 (Fig. 1A), after which the frequency of P.1 relative to other lineages increased rapidly in the tested samples from Manaus (Fig. 1D; lineage frequency information can be found in fig. S5). Retrospective genome sequencing might be able to recover earlier P.1 genomes. Between 2 November 2020 and 9 January 2021, we observed 7137 SARI cases and 3144 SARI deaths in Manaus (Fig. 1A). We generated a total of 182 SARS-CoV-2 sequences from Manaus during this period. This corresponds to one genome for each 39 SARI cases in Manaus, and this ratio is >100-fold higher as compared with the average number of shared genomes per reported case during the same period in Brazil.

Dating the emergence of the P.1 lineage

We used molecular-clock phylogenetics to understand the emergence and evolution of lineage P.1 (*25*). We first regressed root-to-tip genetic distances against sequence sampling dates (*29*) for the P.1, P.2, and B.1.1.28 lineages separately (figs. S6 to S8). This exploratory analysis revealed similar evolutionary rates within each lineage but greater root-to-tip distances for P.1 compared with B.1.1.28 (fig. S8), suggesting that the emergence of P.1 was preceded by a period of faster molecular evolution. The B.1.1.7 lineage exhibits similar evolutionary characteristics (*13*), which was hypothesized to have occurred in a chronically infected or immunocompromised patient (*30*, *31*).

To date the emergence of P.1, while accounting for a faster evolutionary rate along

its ancestral branch, we used a local molecular clock model (*32*) with a flexible nonparametric demographic tree prior (*33*). Using this approach, we estimated the date of the common ancestor of the P.1 lineage to be around 15 November 2020 [median, 95% Bayesian credible interval (BCI), 6 October to 24 November 2020; mean, 9 November 2020] (fig. S9). This is only 3 to 4 weeks before the resurgence in SARS-CoV-2 confirmed cases in Manaus (Figs. 1A and 2 and fig. S9). The P.1 sequences formed a single well-supported group (posterior probability = 1.00) that clustered most closely with B.1.1.28 sequences from Manaus (Fig. 2, "AM"), suggesting that P.1 emerged there. The earliest P.1 samples were detected in Manaus (*34*). The first known travel-related cases were detected in Japan (*28*) and São Paulo (table S3) and were both linked to travel from Manaus. Furthermore, the local clock model statistically confirmed a higher evolutionary rate for the branch immediately ancestral to lineage P.1 compared with lineage B.1.1.28 as a whole [Bayes factor (BF) = 6.04].

Our data indicate multiple introductions of the P.1 lineage from Amazonas to Brazil's southeastern states (Fig. 2). We also detected seven small well-supported clusters of P.2 sequences from Amazonas (two to six sequences, posterior probability = 1.00). Virus exchange between Amazonas state and the urban metropolises in southeast Brazil largely follows patterns of national air travel mobility (Fig. 1D and fig. S10).

Infection with P.1 and sample viral loads

We analyzed all quantitative reverse transcription polymerase chain reaction (RT-PCR) SARS-CoV-2–positive results from a laboratory that has provided testing in Manaus since May 2020 (Fig. 1A and data file S1), with the aim of exploring trends in sample quantitative RT-PCR cycle threshold (Ct) values, which are inversely related to sample virus loads and transmissibility (*35*). By focusing on data from a single laboratory, we reduced instrument and process variation that can affect Ct measurements.

We analyzed a set of quantitative RT-PCR positive cases for which virus genome sequencing and lineage classification had been undertaken (n = 147 samples). Using a logistic function (Fig. 3A), we found that the fraction of samples classified as P.1 increased from 0 to 87% in around 7 weeks (table S4), quantifying the trend shown in Fig. 1C. We found a small but statistically significant association between P.1 infection and lower Ct values, for both the *E* gene (lognormal regression, P = 0.029, n = 128 samples, 65 of which were P.1) and *N* gene (P = 0.01, n = 129 samples, 65 of which were P.1), with Ct values lowered by 1.43 [0.17 to 2.60, 95% confidence interval (CI)] and 1.91 (0.49 to 3.23) cycles in the P.1 lineage on average, respectively (Fig. 3B).

Using a larger sample of 942 Ct values (including an additional 795 samples for which no lineage information was available), we investigated Ct values across three time periods characterized by increasing P.1 relative abundance. Average Ct values for both the *E* and *N* genes declined through time, as both case numbers and the fraction of P.1 infections increased, with Ct values significantly lower in period 3 as compared with period 1 (*E* gene, P = 0.12 and P < 0.001 for comparison of time periods 2 and 3 to period 1; *N* gene, P = 0.14 and P < 0.001, respectively) (Fig. 3C). Analyses of Ct values for samples from a different laboratory, also based in Manaus, showed similarly significant declines between the first and third time periods defined here (P < 0.0001 for both *E* and *N* genes) (fig. S11 and data file S3).

However, population-level Ct distributions are sensitive to changes in the average time since infection when samples are taken, so that median Ct values can decrease during epidemic growth periods and increase during epidemic decline (*36*). To account for this effect, we assessed the association between P.1 infection and Ct levels while controlling for the delay between symptom onset and sample collection. Statistical significance was lost for both data sets (E gene, P = 0.15, n = 42 samples, 22 of which were P.1; N gene, P = 0.12, n = 42 samples, 22 of which were P.1). Owing to this

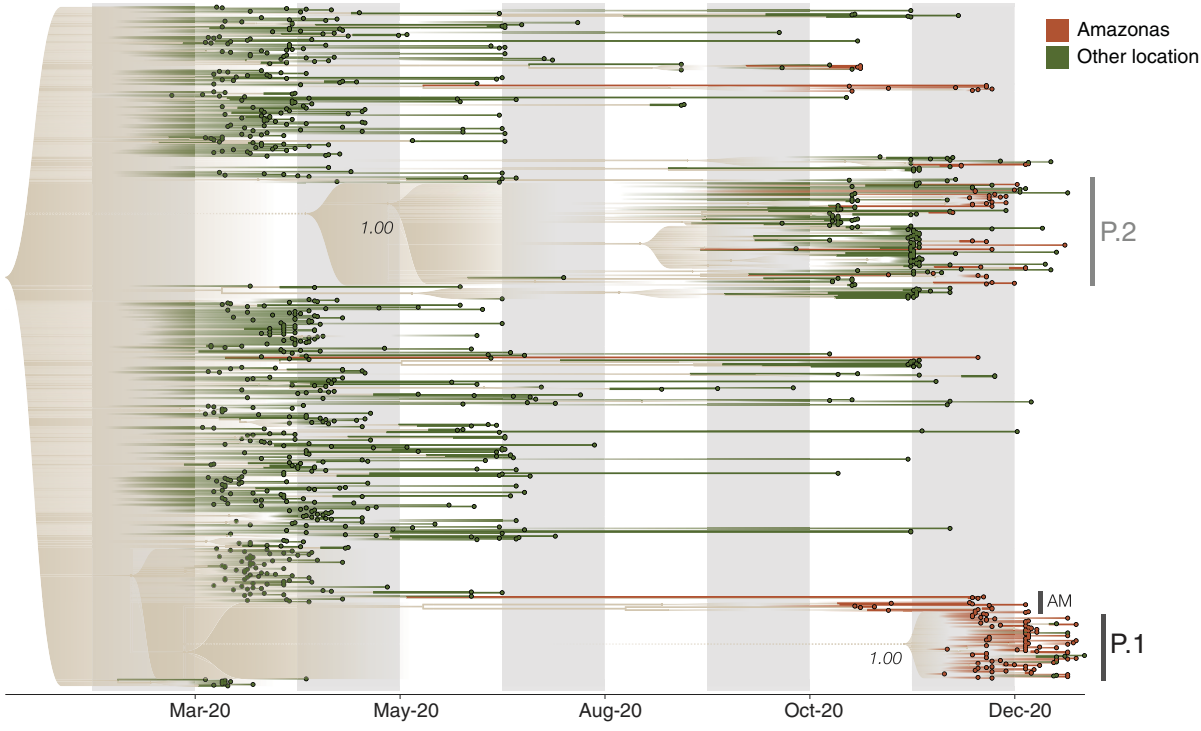


Fig. 2. Visualization of the time-calibrated maximum clade credibility tree reconstruction for B.1.1.28, P.1, and P.2 lineages in Brazil. Terminal branches and tips of Amazonas state are colored in brown, and those from other locations are colored in green (n = 962 viral genomes). Nodes with posterior probabilities of <0.5 have been collapsed into polytomies, and their range of divergence dates are illustrated as shaded expanses.

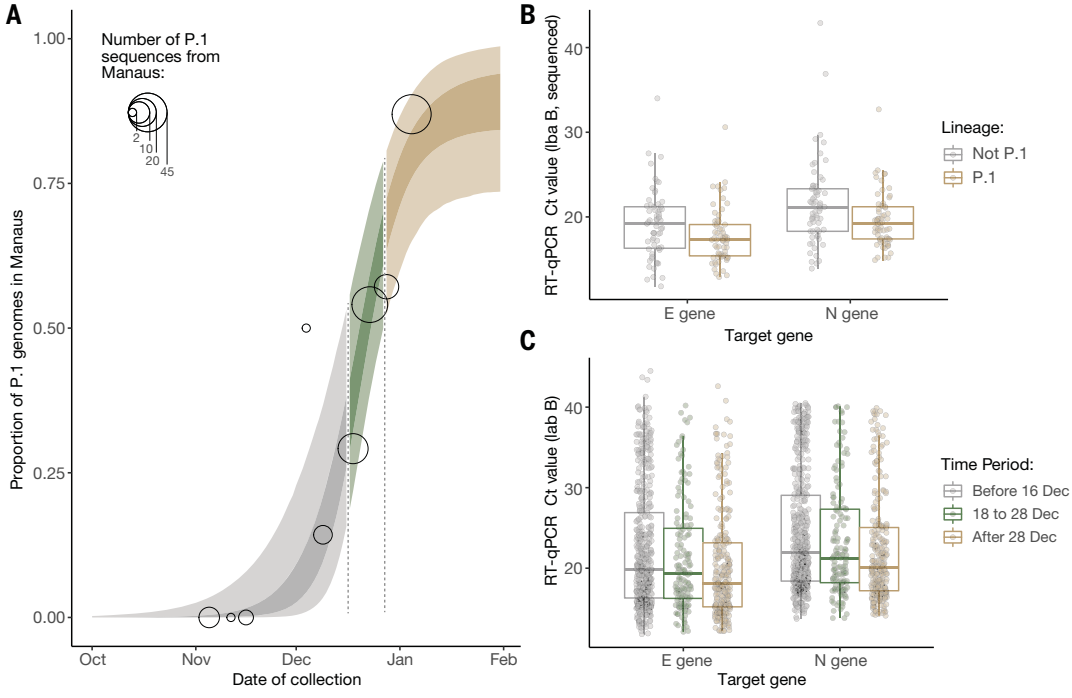


Fig. 3. Temporal variation in the proportion of sequenced genomes belonging to P.1, and trends in quantitative RT-PCR Ct values for COVID-19 infections in Manaus. (**A**) Logistic function fitting to the proportion of genomes in sequenced infections that have been classified as P.1 (black circles, size indicating number of infections sequenced), divided up into time periods when the predicted proportion of infections that are due to P.1 is <1/3 (light brown), between 1/3 and 2/3 (green), and greater than 2/3 (gray). For the model fit, the darker ribbon indicates the 50% credible interval, and the lighter ribbon indicates the 95% credible interval. For the data points, the gray thick line is the 50% exact binomial CI, and the thinner line is the 95% exact binomial CI. (**B**) Ct values for genes E and N in a sample of symptomatic cases presenting for testing at a health care facility in Manaus (laboratory A), stratified according to the period defined in (A) in which the oropharyngeal and nasal swab collections occurred. (**C**) Ct values for genes E and N in a subsample of 184 infections included in (B) that had their genomes sequenced (dataset A).

confounding factor, we cannot distinguish whether P.1 infection is associated with increased viral loads (*37*) or a longer duration of infection (*38*).

Mathematical modeling of lineage P.1 epidemiological characteristics

We next explored epidemiological scenarios that might explain the recent resurgence of transmission in Manaus (*39*). To do this, we extended a semimechanistic Bayesian model of SARS-CoV-2 transmissibility and mortality (*40*–*42*) to include two categories of virus ("P.1" and "non-P.1") and to account for infection severity, transmissibility, and propensity for reinfection to vary between the categories. It also integrates information on the timing of P.1 emergence in Manaus using our molecular clock results (Fig. 2). The model explicitly incorporates waning of immune protection after infection, parameterized on the basis of dynamics observed in recent studies (*16*, *43*), to explore the competing hypothesis that waning of prior immunity might explain the observed resurgence (*42*). We used the model to evaluate the statistical support that P.1 possesses altered epidemiological characteristics compared with local non-P.1 lineages. Epidemiological model details and sensitivity analyses (tables S5 to S10) can be found in the supplementary materials. The model is fitted to both COVID-19 mortality data [with a correction for systematic reporting delays (*44*, *45*)] and the estimated increase through time in the proportion of infections due to P.1 derived from genomic data (table S4). We assumed that within-category immunity wanes over time (50% wane within a year, although sensitivity analyses varying the rapidity of waning are presented in table S7) and that cross-immunity (the degree to which previous infection with a virus belonging to one category protects against subsequent infection with the other) is symmetric between categories.

Our results suggest that the epidemiological characteristics of P.1 are different from those of previously circulating local SARS-CoV-2 lineages, but the results also highlight substantial uncertainty in the extent and nature of this difference. Plausible values of transmissibility and cross-immunity exist in a limited area but are correlated (Fig. 4A, with the extent of immune evasion defined as 1 minus the inferred cross-immunity). This is expected because in the model, a higher degree of cross-immunity means that greater transmissibility of P.1 is required to generate a second epidemic. Within this plausible region of parameter space, P.1 can be between 1.7 and 2.4 times more transmissible (50% BCI, 2.0 median, with a 99% posterior probability of being >1) than local non-P1 lineages and can evade 21 to 46% (50% BCI, 32% median, with a 95% posterior probability of being able to evade at least 10%) of protective immunity elicited by previous infection with non-P.1 lineages, corresponding to 54 to 79% (50% BCI, 68% median) cross-immunity (Fig. 4A). The joint-posterior distribution is inconsistent with a combination of highly increased transmissibility and low cross-immunity and, conversely, also with near-complete cross-immunity but only a small increase in transmissibility (Fig. 4A). Moreover, our results further show that natural immunity waning alone is unlikely to explain the observed dynamics in Manaus, with support for P.1 possessing altered epidemiological characteristics robust to a range of values assumed for the date of the lineage's emergence and the rate of natural immunity waning (tables S5 and S7). We caution that these results are not generalizable to other settings; more detailed and direct data are needed to identify the exact degree and nature of the changes to the epidemiological characteristics of P.1 compared with previously circulating lineages.

We estimate that infections are 1.2 to 1.9 times more likely (50% BCI, median 1.5, 90% posterior probability of being >1) to result in mortality in the period after the emergence of P.1, compared with before, although posterior estimates of this relative risk are also correlated with inferred cross-immunity (Fig. 4B). More broadly, the recent epidemic in Manaus has strained the city's health care system, leading to inadequate access to medical care (*46*). We therefore cannot determine whether the estimated increase in relative mortality risk is due to P.1 infection, stresses on the Manaus

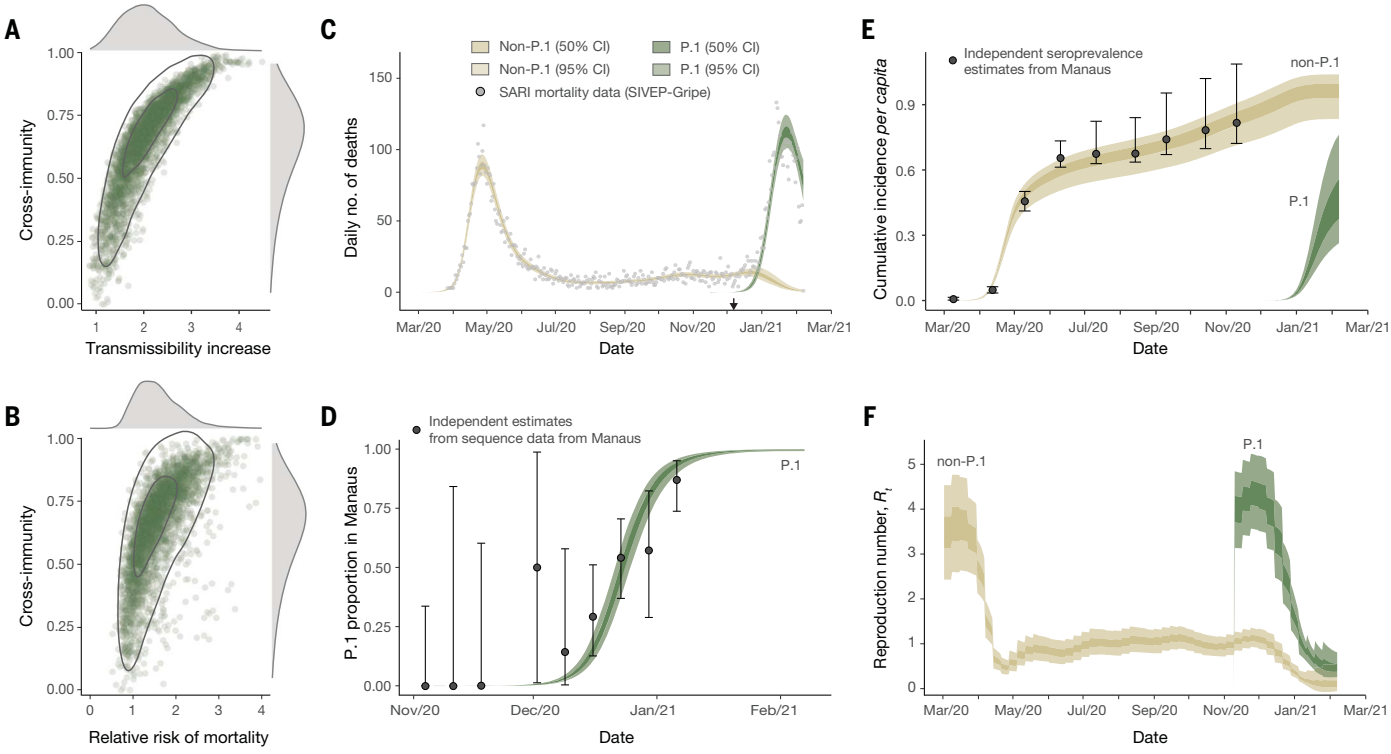


Fig. 4. Estimates of the epidemiological characteristics of P.1 inferred from a multicategory Bayesian transmission model fitted to data from Manaus, Brazil. (**A**) Joint posterior distribution of the cross-immunity and transmissibility increase inferred through fitting the model to mortality and genomic data. Gray contours indicate posterior density intervals ranging from the 95 and 50% isoclines. Marginal posterior distributions for each parameter shown along each axis. (**B**) As for (A), but showing the joint-posterior distribution of cross-immunity and the inferred relative risk of mortality in the period after emergence of P.1 compared with the period prior. (**C**) Daily incidence of COVID-19 mortality. Points indicate severe acute respiratory mortality records from the SIVEP-Gripe database (*67*, *69*). Brown and green ribbons indicate model fit for COVID-19 mortality incidence, disaggregated by mortality attributable to non-P.1 lineages (brown) and the P.1 lineage (green). (**D**) Estimate of the proportion of P.1 infections through time in Manaus. Black data points with error bars are the empirical proportion observed in genomically sequenced cases (Fig. 3A), and green ribbons (dark = 50% BCI, light = 95% BCI) are the model fit to the data. (**E**) Estimated cumulative infection incidence for the P.1 and non-P.1 categories. Black data points with error bars are reversion-corrected estimates of seroprevalence from blood donors in Manaus (*2*). Colored ribbons are the model predictions of cumulative infection incidence for non-P.1 lineages (brown) and P.1 lineages (green). These points are shown for reference only and were not used to fit the model. (**F**) Bayesian posterior estimates of trends in reproduction number R_t for the P.1 and non-P.1 categories.

health care system, or both. Detailed clinical investigations of P.1 infections are needed. Our model makes the assumption of a homogeneously mixed population and therefore ignores heterogeneities in contact patterns (differences in private versus public hospitals are provided in fig. S13). This is an important area for future research. The model fits observed time series data from Manaus on COVID-19 mortality (Fig. 4C) and the relative frequency of P.1 infections (Fig. 4D) and also captures previously estimated trends in cumulative seropositivity in the city (Fig. 4E). We estimate the reproduction number (R_t) on 7 February 2021 to be 0.1 (median, 50% BCI, 0.04 to 0.2) for non-P.1 and 0.5 (median, 50% BCI, 0.4 to 0.6) for P.1 (Fig. 4F).

Characterization and adaptation of a constellation of spike protein mutations

Lineage P.1 contains 10 lineage-defining amino acid mutations in the virus spike protein (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, and T1027I) compared with its immediate ancestor (B.1.1.28). In addition to the possible increase in the rate of molecular evolution during the emergence of P.1, we found by use of molecular selection analyses (*47*) evidence that eight of these 10 mutations are under diversifying positive selection (table S1 and fig. S14). (Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. In the mutants, other amino acids were substituted at certain locations; for example, K417T indicates that lysine at position 417 was replaced by threonine.)

Three key mutations present in P.1—N501Y, K417T, and E484K—are in the spike protein RBD. The former two interact with human ACE2 (hACE2) (*11*), whereas E484K is located in a loop region outside the direct hACE2 interface (fig. S14). The same three residues are mutated with the B.1.351 variant of concern, and N501Y is also present in the B.1.1.7 lineage. The independent emergence of the same constellation of mutations in geographically distinct lineages indicates a process of convergent molecular adaptation. Similar to SARS-CoV-1 (*48*–*50*), mutations in the RBD may increase affinity of the virus for host ACE2 and consequently influence host cell entry and virus transmission. Recent molecular analysis of B.1.351 (*51*) indicates that the three P.1 RBD mutations may similarly enhance hACE2 engagement, providing a plausible hypothesis for an increase in transmissibility of the P.1 lineage. Moreover, E484K is associated with reduced antibody neutralization (*6*, *9*, *52*, *53*). RBD-presented epitopes account for ~90% of the neutralizing activity of sera from individuals previously infected with SARS-CoV-2 (*54*); thus, tighter binding of P.1 viruses to hACE2 may further reduce the effectiveness of neutralizing antibodies.

Conclusion

We show that P.1 most likely emerged in Manaus in mid-November, where high attack rates have been previously reported. High rates of mutation accumulation over short time periods have been reported in chronically infected or immunocompromised patients (*13*). Given a sustained generalized epidemic in Manaus, we believe that this is a potential scenario for P.1 emergence. Genomic surveillance and early data sharing by teams worldwide have led to the rapid detection and characterization of SARS-CoV-2 and new variants of concern (VOCs) (*25*), yet such surveillance is still limited in many settings. The P.1 lineage is spreading rapidly across Brazil (*55*), and this lineage has now been detected in >*36* countries (*56*). But existing virus genome sampling strategies are often inadequate for determining the true extent of VOCs in Brazil, and more detailed data are needed to address the impact of different epidemiological and evolutionary processes in their emergence. Sustainable genomic surveillance efforts to track variant frequency [for example, (*57*–*59*)] coupled with analytical tools to quantify lineage dynamics [for example, (*60*, *61*)] and anonymized epidemiological surveillance data (*62*, *63*) could enable enhanced real-time surveillance of VOCs worldwide. Studies to evaluate real-world vaccine efficacy in response to P.1 are urgently needed. Neutralization titers represent only one component of the elicited response to vaccines, and minimal reduction of neutralization titers relative to earlier circulating strains is not uncommon. Until an equitable allocation and access to effective vaccines is available to all, nonpharmaceutical interventions should continue to play an important role in reducing the emergence of new variants.

REFERENCES AND NOTES

1. P. C. Hallal *et al*., *Lancet Glob. Health* **8**, e1390–e1398 (2020).
2. L. F. Buss *et al*., *Science* **371**, 288–292 (2021).
3. C. Álvarez-Antonio *et al*., medRxiv 21249913 [Preprint] 20 January 2021. doi:10.1101/2021.01.17.21249913.
4. M. Mercado, M. Ospina, Institutoo Nacional de Salud, "Seroprevalencia de SARS-CoV-2 durante la epidemia en Colombia: estudio país" (2020); www.ins.gov.co/BibliotecaDigital/Seroprevalencia-estudio-colombia.pdf.
5. Fundação de Vigilância em Saúde do Amazonas, "Perfil clínico e demográfico dos casos de Covid-19 no estado do Amazonas: uma análise comparativa entre 2020 e 2021," No. 17 (2021); www.fvs.am.gov.br/media/publicacao/boletim_covid_17.pdf.
6. A. J. Greaney *et al*., *Cell Host Microbe* **29**, 463–476.e6 (2021).
7. T. N. Starr *et al*., *Cell* **182**, 1295–1310.e20 (2020).
8. M. A. Suchard, R. E. Weiss, J. S. Sinsheimer, *Syst. Biol.* **52**, 48–54 (2003).
9. Z. Wang *et al*., *Nature* (2021).
10. Y. Weisblum *et al*., *eLife* **9**, e61312 (2020).
11. J. Lan *et al*., *Nature* **581**, 215–220 (2020).
12. H. Tegally *et al*., *Nature* (2021).
13. A. Rambaut *et al*., on behalf of COVID-19 Genomics Consortium UK (CCoG-UK), "Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations" (2020); https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
14. SAMRC Report on Weekly Deaths in South Africa (2021); www.samrc.ac.za/reports/report-weekly-deaths-south-africa?bc=254.
15. N. L. Washington *et al*., Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. medRxiv 21251159 [Preprint] 7 February 2021). doi:10.1101/2021.02.06.21251159.
16. C. H. Hansen, D. Michlmayr, S. M. Gubbels, K. Mølbak, S. Ethelberg, *Lancet* **397**, 1204–1212 (2021).
17. E. Volz *et al*., *Nature* (2021).
18. D. D. S. Candido *et al*., *J. Travel Med.* **27**, taaa042 (2020).
19. D. S. Candido *et al*., *Science* **369**, 1255–1260 (2020).
20. W. M. de Souza *et al*., *Nat. Hum. Behav.* **4**, 856–865 (2020).
21. J. G. Jesus *et al*., *Rev. Inst. Med. Trop. São Paulo* **62**, e30 (2020).
22. I. M. Claro *et al*., *Emerg. Infect. Dis.* **27**, 970–972 (2021).
23. V. A. D. Nascimento *et al*., *Mem. Inst. Oswaldo Cruz* **115**, e200310 (2020).
24. J. R. Tyson *et al*., *bioRxiv* 2020.09.04.283077 (2020).
25. World Health Organization, (2021). Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health, 8 January 2021. World Health Organization (2021); https://apps.who.int/iris/handle/10665/338480.
26. A. Rambaut *et al*., *Nat. Microbiol.* **5**, 1403–1407 (2020).
27. N. R. Faria *et al*., on behalf of CADDE Genomic Network, "Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings" (2021); https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586.
28. T. Fujino *et al*., *Emerg. Infect. Dis.* **27**, (2021).
29. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, *Virus Evol.* **2**, vew007 (2016).
30. B. Choi *et al*., *N. Engl. J. Med.* **383**, 2291–2293 (2020).
31. V. A. Avanzato *et al*., *Cell* **183**, 1901–1912.e9 (2020).
32. A. J. Drummond, M. A. Suchard, *BMC Biol.* **8**, 114 (2010).
33. M. S. Gill *et al*., *Mol. Biol. Evol.* **30**, 713–724 (2013).
34. F. N. Naveca *et al*, *Nat. Portfolio* 10.21203/rs.3.rs-275494/v1 (2021).
35. M. Marks *et al*., *Lancet Infect. Dis.* S1473-3099(20)30985-3 (2021).
36. J. A. Hay, L. Kennedy-Shaffer, S. Kanjilal, M. Lipsitch, M. J. Mina, Estimating epidemiologic dynamics from single cross-sectional viral load distributions. medRxiv 20204222 [Preprint] 13 February 2021). doi:10.1101/2020.10.08.20204222.
37. M. Kidd *et al*., *J. Infect. Dis.* jiab082 (2021).
38. K. Stephen *et al*., Densely sampled viral trajectories suggest longer duration of acute infection with B.1.1.7 variant relative to non-B.1.1.7 SARS-CoV-2 (2021); https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37366884.
39. E. C. Sabino *et al*., *Lancet* **397**, 452–455 (2021).
40. H. J. T. Unwin *et al*., *Nat. Commun.* **11**, 6189 (2020).
41. S. Flaxman *et al*., *Nature* **584**, 257–261 (2020).
42. James Scott, Axel Gandy, Swapnil Mishra, Juliette Unwin, Seth Flaxman, Samir Bhatt, Epidemia - Modeling of epidemics using hierarchical Bayesian models; https://imperialcollegelondon.github.io/epidemia/index.html.
43. V. Hall *et al*., *bioRxiv* 21249642 [Preprint] 15 January 2021). doi:10.1101/2021.01.13.21249642.
44. S. F. McGough, M. A. Johansson, M. Lipsitch, N. A. Menzies, *PLOS Comput. Biol.* **16**, e1007735 (2020).
45. I. Hawryluk *et al*., Gaussian Process Nowcasting: Application to COVID-19 Mortality Reporting. *arXiv* **arXiv:2102.11249**, (2021).
46. Agencia Brasil, Covid-19: Amazonas já transferiu 424 pacientes para outros estados (2021); https://agenciabrasil.ebc.com.br/saude/noticia/2021-02/covid-19-amazonas-ja-transferiu-424-pacientes-para-outros-estados.
47. S. L. Pond, S. D. Frost, S. V. Muse, *Bioinformatics* **21**, 676–679 (2005).
48. X. X. Qu *et al*., *J. Biol. Chem.* **280**, 29588–29595 (2005).
49. H. D. Song *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2430–2435 (2005).
50. W. Li *et al*., *EMBO J.* **24**, 1634–1643 (2005).
51. D. Zhou *et al*., *Cell* S0092-8674(21)00226-9 (2021).
52. C. K. Wibmer *et al*., *Nat. Med.* (2021).
53. S. Cele *et al*., *Nature* (2021).
54. L. Piccoli *et al*., *Cell* **183**, 1024–1042.e21 (2020).
55. A. F. Martins *et al*., *Euro Surveill.* **26**, 2100276 (2021).
56. A. O'Toole *et al*., The COVID-19 Genomics UK (COG-UK) consortium, et al., "Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2" (2021); https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592.
57. TESSy. The European Surveillance System (TESSy) (2015); www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy.
58. COG-UK, COVID-19 Genomics UK Consortium (2020); www.cogconsortium.uk.
59. SPHERES, SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance; (2020); www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html.
60. L. du Plessis *et al*., *Science* **371**, 708–712 (2021).
61. E. Volz *et al*., *Cell* **184**, 64–75.e11 (2021).
62. B. Xu, M. U. G. Kraemer, Open COVID-19 Data Curation Group, *Lancet Infect. Dis.* **20**, 534 (2020).
63. Global.health (2021); https://global.health.
64. GitHub Repository; https://github.com/CADDE-CENTRE.
65. S., CADDE CENTRE, C. Whittaker, CADDE-CENTRE/Novel-SARS-CoV-2-P1-Lineage-in-Brazil: Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil (peer-review version out soon). Zenodo (2021); doi:10.5281/zenodo.4676853.
66. Fundação de Vigilância em Saúde do Amazonas, Amazonas, "Dados epidemiológicos e financeiros das ações de combate à COVID-19. Publicações (2021); www.fvs.am.gov.br/publicacoes.
67. SRAG 2020 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19 - Open Data; https://opendatasus.saude.gov.br/dataset/bd-srag-2020.
68. Y. Shu, J. McCauley, *Euro Surveill.* **22**, 30494 (2017).
69. SRAG 2021 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19 - Open Data; https://opendatasus.saude.gov.br/dataset/bd-srag-2021.

ACKNOWLEDGMENTS

We thank L. Matkin (University of Oxford), M. Oikawa (Universidade Federal do ABC), and A. Acosta (University of Sao Paulo) for logistic support and C. Sachi (Instituto Adolfo Lutz) for agreeing with the use of unpublished sequence data available in GISAID before publication. We thank the anonymous reviewers for their considerations and suggestions. We thank the administrators of the GISAID database for supporting rapid and transparent sharing of genomic data during the COVID-19 pandemic. A full list acknowledging the authors publishing data used in this study can be found in data file S4. **Funding:** This work was supported by a Medical Research Council-São Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0) (https://caddecentre.org); FAPESP (E.C.S.: 18/14389-0; I.M.C: 2018/17176-8 and 2019/12000-1, F.C.S.S.: 2018/25468-9; J.G.d.J.: 2018/17176-8, 2019/12000-1, 18/14389-0; T.M.C.: 2019/07544-2; C.A.M.S.: 2019/21301-5; W.M.S.: 2017/13981-0, 2019/24251-9; L.M.d.S.: 2020/04272-9; M.C.d.P.: 2019/21568-1; V.H.N.: 2018/12579-7; C.A.P.: 2019/21858-0; and P.S.P.: 16/18445-7; J.L.P.-M.: 2020/04558-0); Wellcome Trust and Royal Society (N.R.F.: Sir Henry Dale Fellowship: 204311/Z/16/Z); Wellcome Trust (Wellcome Centre for Human Genetics: 203141/Z/16/Z); Clarendon Fund and Department of Zoology, University of Oxford (D.d.S.C.); Medical Research Council (T.A.B and R.J.G.H: MR/S007555/1); European Molecular Biology Organisation (R.J.G.H.: ALTF 869-2019); CNPq (R.S.A.: 312688/2017-2, 439119/2018-9; W.M.S.: 408338/2018-0, 304714/2018-6; V.H.N.: 304714/2018-6); FAPERJ (R.S.A.: 202.922/2018); FFMUSP (M.S.R.: 206.706; C.A.P.); Imperial College COVID-19 Research Fund (H.M.S. and S.F.); CAPES (G.M.F. and C.A.P., Code 001); Wellcome Trust Collaborator Award (P.L., A.R., and N.J.L.: 206298/Z/17/Z); European Research Council (P.L and A.R.: 725422-ReservoirDOCS); European Union's Horizon 2020 project MOOD (P.L. and M.U.G.K.: 874850); U.S. National Institutes of Health (M.A.S.: U19 AI135995); Oxford Martin School (O.G.P.); Branco Weiss Fellowship (M.U.G.K); Covid-19 Research Fund (S.F.); EPSRC (S.F.: EP/V002910/1; M.M. through the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning); BMGF (S.B.); UKRI (S.B.); Novo Nordisk Foundation (S.B.); Academy of Medical Sciences (S.B.); BRC (S.B.); MRC (S.B.); and Bill & Melinda Gates Foundation (O.R.: OPP1175094). We acknowledge support from the Rede Corona-ômica BR MCTI/FINEP affiliated to RedeVírus/MCTI (FINEP 01.20.0029.000462/20, CNPq 404096/2020-4), FAPESP project 2018/12579-7 CNPq project 304714/2018-6 (V.H.N.), EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning at Imperial and Oxford (M.M.), and the Bill & Melinda Gates Foundation (OPP1175094) (O.R.). This work received funding from the UK Medical Research Council under a concordat with the UK Department for International Development. We additionally acknowledge support from Community Jameel and the NIHR Health Protection Research Unit in Modelling Methodology. Last,

we also gratefully acknowledge support from Oxford Nanopore Technologies for a donation of sequencing reagents and NVIDIA Corporation and Advanced Micro Devices for a donation of parallel computing resources. **Author contributions:** Conceptualization: N.R.F., T.A.M., C.W., I.M.C., D.d.S.C., A.R., C.D., O.G.P., S.F., S.B., and E.C.S. Methodology: N.R.F., T.A.M., C.W., I.M.C., D.d.S.C., S.M., F.C.S.S., I.H., M.S.R., J.G.d.J., L.A.M.F., P.S.A., T.M.C., C.A.M.S., E.R.M., J.T.M., R.H.M.P., P.S.P., M.U.G.K., R.J.G.H., T.A.B., O.G.P., M.A.S., S.L.K.P., O.R., N.M.F., N.J.L., P.L., A.R., C.D., S.F., S.M., and E.C.S. Investigation: N.R.F., T.A.M., C.W., I.M.C., D.d.S.C., S.M., M.A.E.C., F.C.S.S., I.H., M.S.R., J.G.d.J., L.A.M.F., P.S.A., T.M.C., C.A.M.S., E.R.M., J.T.M., R.H.M.P., P.S.P., M.U.G.K., R.J.H.H, N.G., W.M.S., L.J.T.A., C.d.C.C., H.H., G.M.F., E.C.R., L.M.d.S., M.C.d.P., F.S.V.M., A.B.d.L., J.d.P.S., D.A.G.Z., A.C.d.S.F., R.P.S., D.J.L., P.G.T.W., H.M.S., A.L.P.d.S., M.S.V., , V.S.D.C., R.M.F.F., H.M.d.S., R.S.A., B.N., J.A.H., M.M., X.M., H.C., R.S., A.G., M.A.S., T.A.B., S.L.K.P., C.-H.W., O.R., N.M.F., C.A.P., V.H.N., N.J.L., P.L., A.R., N.A.F., M.d.P.S.S.C., C.D., O.G.P., S.F., S.B., and E.C.S. Visualization: N.R.F., T.A.M., C.W., D.d.S.C., I.M.C., J.T.M., A.R., S.L.K.P., T.A.B., C.W., and S.B. Funding acquisition: N.R.F., N.J.L., A.R., O.G.P., N.A.F., S.F., S.B., and E.C.S. Project administration: N.R.F. and E.C.S. Supervision: N.R.F., O.G.P., A.R., C.D., N.J.L., S.B., and E.C.S. Writing, original draft: N.R.F., T.A.M., C.W., I.M.C., D.d.S.C., S.F., S.B., O.G.P., C.D., and E.C.S. Writing, review and editing: All authors. **Competing interests:** S.B. declares that he advises on The Scientific Pandemic Influenza Group on Modelling (SPI-M) and advises the FCA on a legal matter regarding COVID-19 infections in England in March 2020. He is not paid for either of these advisory roles, and neither are related to the work in this paper. All other authors declare that they have no competing interests. **Data and materials availability:** All data, code, and materials used in the analysis are available in a dedicated GitHub Repository (*64*, *65*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/372/6544/815/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S16
Tables S1 to S10
References (*70*–*102*)
Data Files S1 to S6
MDAR Reproducibility Checklist

25 February 2021; accepted 11 April 2021
Published online 14 April 2021
10.1126/science.abh2644

Science

Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil

Nuno R. FariaThomas A. MellanCharles WhittakerIngra M. ClaroDarlan da S. CandidoSwapnil MishraMyuki A. E. CrispimFlavia C. S. SalesIwona HawrylukJohn T. McCroneRuben J. G. HulswitLucas A. M. FrancoMariana S. RamundoJaqueline G. de JesusPamela S. AndradeThais M. ColettiGiulia M. FerreiraCamila A. M. SilvaErika R. ManuliRafael H. M. PereiraPedro S. PeixotoMoritz U. G. KraemerNelson Gaburo Jr.Cecilia da C. CamiloHenrique HoeltgebaumWilliam M. SouzaEsmenia C. RochaLeandro M. de SouzaMariana C. de PinhoLeonardo J. T. AraujoFrederico S. V. MaltaAline B. de LimaJoice do P. SilvaDanielle A. G. ZauliAlessandro C. de S. FerreiraRicardo P. SchnekenbergDaniel J. LaydonPatrick G. T. WalkerHannah M. SchlüterAna L. P. dos SantosMaria S. VidalValentina S. Del CaroRosinaldo M. F. FilhoHelem M. dos SantosRenato S. AguiarJosé L. Proença-ModenaBruce NelsonJames A. HayMélodie MonodXenia MiscouridouHelen CouplandRaphael SonabendMichaela VollmerAxel GandyCarlos A. Prete Jr.Vitor H. NascimentoMarc A. SuchardThomas A. BowdenSergei L. K. PondChieh-Hsi WuOliver RatmannNeil M. FergusonChristopher DyeNick J. LomanPhilippe LemeyAndrew RambautNelson A. FraijiMaria do P. S. S. CarvalhoOliver G. PybusSeth FlaxmanSamir BhattEster C. Sabino

Science, 372 (6544), • DOI: 10.1126/science.abh2644

Unmitigated spread in Brazil

Despite an extensive network of primary care availability, Brazil has suffered profoundly during the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic. Using daily data from state health offices, Castro *et al.* analyzed the pattern of spread of COVID-19 cases and deaths in the country from February to October 2020. Clusters of deaths before cases became apparent indicated unmitigated spread. SARS-CoV-2 circulated undetected in Brazil for more than a month as it spread north from Sã o Paulo. In Manaus, transmission reached unprecedented levels after a momentary respite in mid-2020. Faria *et al.* tracked the evolution of a new, more aggressive lineage called P.1, which has 17 mutations, including three (K417T, E484K, and N501Y) in the spike protein. After a period of accelerated evolution, this variant emerged in Brazil during November 2020. Coupled with the emergence of P.1, disease spread was accelerated by stark local inequalities and political upheaval, which compromised a prompt federal response.

Science, abh1558 and abh2644, this issue p. 821 and p. 815

Science (ISSN 1095-9203) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Chapter 5

SARS-CoV-2 hospital-associated transmission dynamics in São Paulo, Brazil: a retrospective genomic surveillance study

This chapter presents the first genomic epidemiology study assessing the within- and between- hospital SARS-CoV-2 transmission dynamics in Brazil. It was developed retrospectively with epidemiological and sequence data from the largest hospital complex in Latin America, the Hospital das Clínicas of the São Paulo University Medical School, in which an interesting set up was in place during the first wave of SARS-CoV-2 spread in Brazil. Only one medical institute of the complex was dedicated to treating COVID-19 patients, while all other institutes were considered "COVID-free". This shows evidence for SARS-CoV-2 within-hospital transmission to be higher in non-COVID-19 hospitals. This work has recently been submitted for publication in The Lancet Infectious Diseases and it is presented here in full.

"Be safe, be smart, be kind"

Dr. Tedros Adhanom Ghebreyesus

SARS-CoV-2 hospital-associated transmission dynamics in São Paulo, Brazil: a retrospective genomic surveillance study

Darlan S. Candido MSc[1,2]*, Ingra M. Claro PhD[2,3]*, Mariana S. Ramundo PhD[2,3]*, Alessandra Luna-Muschi MD[2,3], Bernardo Gutierrez MSc[1,4], Flavia C. S. Sales BSc[2,3], Jaqueline G. de Jesus PhD[2,3], Erika R. Manuli MSc[2,3], Thaís M. Coletti BSc[2,3], Camila A. M. da Silva BSc[2,3], Pamela S. Andrade MSc[2,3,5], Giulia M. Ferreira MSc[2,3,6], Esmenia C. Rocha BSc[2,3], Leandro M. de Souza[2,3], Mariana C. de Pinho[2,3], Leonardo J. T. Araujo PhD[2,3,7], Cecilia S. Alencar PhD[8], Rinaldo F. Siciliano MD PhD[9], Rogério Zeigler MD[9], Cristhieni Rodrigues MD PhD[9], Maria Emília B. de Souza[10], Michely F. Vieira[10], Raquel Keiko L. Ito MD[3,10], Thais Guimaraes MD PhD[3,11], Tania M. V. Strabelli MD PhD[9], Edson Abdala MD PhD[3,10], Izabel Oliva Marcilio de Souza MD PhD[12], Elizabeth de Faria MD[13], Maura Salaroli Oliveira MD PhD[3,14], Anna S. Levin MD PhD[3,14], Ester C. Sabino MD PhD[2,3], Silvia F. Costa MD PhD[2,3†], Nuno R. Faria, PhD[1,2,15 †]

1 Department of Zoology, University of Oxford, Oxford, OX1 3SZ, United Kingdom.

2 Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, 05403-000, Brazil.

3 Departamento de Moléstias Infecciosas e Parasitárias, Faculdade de Medicina da Universidade de São Paulo, Sao Paulo, 05403-000, Brazil.

4 Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito USFQ

5 Faculdade de Saúde Pública da Universidade de São Paulo, São Paulo, 01246-904, Brazil.

6 Laboratório de Virologia, Instituto de Ciências Biomédicas, Universidade Federal de Uberlândia, Uberlândia, 38400-902, Brazil.

7 Centro de Patologia, Instituto Adolfo Lutz, São Paulo, 01246-000, Brazil.

8 Laboratório de Medicina Laboratorial, Hospital das Clínicas Faculdade de Medicina da Universidade de São Paulo, São Paulo, 05403-000, Brazil.

9 Unidade de Controle de Infecção Hospitalar, Instituto do Coração, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Sao Paulo, 05403-000, Brazil.

10 Instituto do Cancer do Estado de São Paulo, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Sao Paulo, 05403-000, Brazil.

11 Subcomissão de Controle de Infecção Hospitalar, Instituto Central, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Sao Paulo, 05403-000, Brazil.

12 Núcleo de Vigilância Epidemiológica, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Sao Paulo, 05403-000, Brazil.

13 Centro de Atendimento ao Colaborador, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Sao Paulo, 05403-000, Brazil

14 Grupo de Controle de Infecção Hospitalar, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Sao Paulo, 05403-000, Brazil.

15 MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, London, SW7 2AZ, United Kingdom.

Correspondence to:

Nuno R. Faria

Email: n.faria@imperial.ac.uk (N.R.F.);

Silvia F. Costa

Email: silviacosta@usp.br

*Authors contributed equally.

†Authors contributed equally.

Abstract

Background: Brazil reported its first SARS-CoV-2 case on 26 February 2020 in an international traveler returning to São Paulo, Brazil. By 10 June 2020, 3,898 healthcare workers (HCW) and patients at the Hospital das Clínicas (HC) in São Paulo, the largest hospital complex in Latin America, had tested positive for SARS-CoV-2 RNA. We aimed to provide insight into the transmission of SARS-CoV-2 in healthcare workers and patients, and within and between HC institutes during the early phase of the epidemic in Brazil.

Methods: We analyzed epidemiological data from SARS-CoV-2 RT-PCR confirmed cases between 13 March to 10 June 2020. A total of 340 SARS-CoV-2 genomes were generated from healthcare workers and patients from two HC institutes not receiving COVID-19 patients (institutes A and C) and one institute receiving exclusively COVID-19 patients (institute B). Within- and between-institute transmission clusters were identified and within-cluster transmission dynamics was assessed using logistic regression analysis and a suite of phylogenetic genetic analyses.

Results: SARS-CoV-2 weekly incidence in healthcare workers was highest in institutes not receiving COVID-19 patients, and decreased by 75%, 54%, 48% for institutes C, A, B, respectively, after universal masking was adopted. We found a total of 86 hospital-acquired patient infections in HC during the study period, 81.4% (n=70) in institute C. Of these, 74.3% of these reported after mandatory universal masking. The average cluster size and cluster duration were larger in non-COVID institutes. Sequences from non-COVID institutes were more likely to cluster together than sequences from institute B (odds ratio, OR=4.17 and OR=3.48, for C and A institutes). The proportion of estimated viral importation events from outside the HC complex to the different HC institutes was highest for institute B (83.64%) compared to institute A (67.85%), and C (57.70%). The statistical support for virus migration from non-COVID institutes A and C to institute B was strong (Bayes factor = 113.7 and 84.4, respectively).

Interpretation: The hospital-associated SARS-CoV-2 transmission was higher in non-COVID-19 healthcare institutes compared to a COVID-19 healthcare institute during the first epidemic wave in

São Paulo, with our data supporting virus infection non-COVID-19 institutes as a source of infection in a COVID-19 institute, suggesting that risk perception and compliance was lower in non-COVID institutes.

Research in Context

On 5 October 2020, we searched PUBMED for studies including the following search terms: ("SARS-CoV-2" OR "COVID-19" OR "coronavirus disease 2019") AND ("genom*" OR "sequenc*" OR "WGS") AND ("nosocomial transmission" OR "hospital outbreak" OR "hospital-acquired" OR "healthcare-associated" OR "health-care associated"). No search restrictions were applied. Our search retrieved 62 studies, out of which only 25 applied genomic epidemiology to uncover hospital-associated SARS-CoV-2 transmission, 23 original works, and 2 reviews. Most studies covered the early stages of the pandemic (14, 61%), focused in a single hospital (17, 74%), presented evidence for hospital-associated SARS-CoV-2 transmission (20, 87%), and focused on describing individual clusters rather than general transmission patterns (18, 78%, median sample size= 44, range 3–764). There is a general consensus over the importance of universal masking and genomic epidemiology for hospital-associated outbreak control. Studies looking at general transmission patterns show evidence for most cases being linked to super spreading events and highlight the role of healthcare workers in hospital-associated transmission, although patients might be more likely infected by other patients. No studies used genetic data to investigate transmission patterns and dynamics between and within non-COVID and COVID hospitals.

Added-value of this study

We aimed to understand the transmission patterns within and between non-COVID-19 and COVID-19 only institutes that are part of the largest hospital complex in Latin America. We show that hospital-associated SARS-CoV-2 transmission was higher in non-COVID institutes, driven by larger clusters and of longer duration, even after mandatory universal masking. We also uncover between-hospital transmission events and temporal patterns of hospital-associated transmission.

Implications of all the available evidence

While separating COVID-19 from non- COVID-19 patients in different wards/hospitals and mandatory universal masking reduce HCW cases and prevent larger within-hospital outbreaks, adequate HCW risk perception and adherence are extremely important for the effectiveness of these policies. In times of COVID-19 fatigue, hospitals should work closely with HCW to increase awareness and compliance within and outside of the hospital environment.

Introduction

Brazil reported the first confirmed case of SARS-CoV-2 on 26 February 2020 (1), and has since experienced two large continuous COVID-19 waves. As of 13 January 2021, 22,724,232 SARS-CoV-2 cases and 620,641 deaths attributed to SARS-CoV-2 have been reported in Brazil, the highest numbers in Latin America (2). During this period, the State of São Paulo reported 20% of all cases in Brazil (4,298,180), with 1,422,413 reported cases in the city of São Paulo alone.

Studies conducted during the early stages of the COVID-19 pandemic testing for viral RNA or antibodies have reported variable prevalence of SARS-CoV-2 (1.1% to 9.8%) amongst HCW across different countries (3–10). In a university hospital in the United Kingdom, HCW infection rates were higher amongst HCW from COVID-19-dedicated units (22.6%) compared to those working in non–COVID-19 units (8.6%) or working in multiple wards (11.2%) (11). A retrospective analysis of 435 cases amongst inpatients in a hospital in London found that 66 (15%) were definitely or probably acquired at the hospital (12).

Asymptomatic and pre-symptomatic HCW can become inadvertent vehicles of transmission to other HCW and non-COVID patients (13). However, identifying SARS-CoV-2 transmission clusters remains a challenging task in part because of unidentified asymptomatic cases that may be missed during contact tracing, and in part, because consensus virus genome sequences from a transmission cluster are often identical due to the virus' relatively slow evolutionary rate. These limitations can partly be overcome when epidemiological and genomic data are analyzed jointly (14–17). To date, most studies using genomic epidemiological approaches to identify SARS-CoV-2 within-hospital transmission have focused on the identification and description of specific transmission clusters, rather than understanding dynamics of epidemiologically-linked clusters of transmission (12,18–23). Moreover, analyses comparing the dynamics between and within different hospital units in large hospital complexes remain scarce, particularly in Latin America where the SARS-CoV-2 pandemic hit hardest.

Understanding SARS-CoV-2 hospital-associated transmission in the early stages of an epidemic is of great importance to improve healthcare response and preparedness for future outbreaks. Here we investigate the patterns of SARS-CoV-2 early transmission in HC São Paulo, the largest hospital complex in Latin America, where all COVID-19 patients were hospitalized in a dedicated building, by combining insights from epidemiological data, genome sequencing, and phylogenetic analysis.

Methods

Epidemiological context

We performed a retrospective study at a large reference teaching tertiary healthcare complex specialized in high complexity cases called Hospital das Clínicas (HC), affiliated with the University of São Paulo, Brazil. The HC complex has approximately 2,200-beds and 22,000 healthcare workers directly involved in patient care offered in nine specialized institutes. From 30 April 2020 to 02 September 2020, one institute was designated as an exclusive COVID-19 hospital (from here on referred to as Institute B). The other institutes were designated for non-COVID-19 patients. HCW were not allowed to move between buildings. Patients with suspected SARS-CoV-2 infection from non-COVID-19 institutes were maintained in individual rooms until RT-PCR confirmation when they would be transferred to the COVID-19 only institute. Universal masking, here defined as mandatory masking to all hospital staff, was adopted at different epidemiological weeks across the institutes: week 15 (institute A), week 17 (institute C), and week 19 (COVID Institute B and other institutes). Detailed information on COVID measures taken across the 3 institutes can be seen in Appendix p 1-3.

Study overview: clinical samples and metadata collection

This study was approved by the national research ethics commission (Comissão Nacional de Ética em Pesquisa) under protocol number CAAE 30127020.0.0000.0068. Patient and HCW SARS-CoV-2 testing were done at the Hospital Central Clinical Laboratory using real-time quantitative polymerase chain reaction (RT-qPCR) on naso-oropharyngeal swabs **(Corman et al., 2020; Waggoner et al., 2020)**. All individuals with RT-qPCR positive samples collected between 13 March to 10 June 2020 were included in this study and results from subsequent tests after the first RT-qPCR positive result were excluded. Clinical and epidemiological data collection included age, sex, home address, occupation, unit of work within the hospital, date of onset of symptoms, symptoms, need for hospitalization, and clinical outcome. Geocoding of residential Zip codes of patients and staff was done by using Google Maps via the geocode function implemented in the R package ggmap.

Additional information such as the complete medical history of patients while in the hospital was retrieved only for clustered patients and health workers.

Patient classification

To classify patients according to the time in days between symptom onset and hospitalization, we adapted the Public Health England (PHE) guidelines (24,25). The PHE classification system considers a 14-day period between a SARS-CoV-2 exposure and COVID-19 symptoms, with an average of five days. Patients were classified into one of four groups: Group 1 (community): symptom onset before hospital admission or up to 2 days after hospital admission; Group 2 (Indeterminate hospital-associated): symptom onset between 3-7 days after hospital admission; Group 3 (Suspected hospital-associated): symptom onset between 8-14 days after hospital admission; Group 4 (Hospital-associated): symptom onset >14 days after hospital admission.

Genome Sequencing

Of a total of 3,898 positive individuals, 454 (12%) samples from Institutes A, B, and C were selected for virus genomic sequencing which was performed at the Institute of Tropical Medicine, University of São Paulo, Brazil. Information on the number of samples sequenced per institute can be seen in appendix p 13. Genome amplification was performed using the n-CoV-2019 Artic protocol (https://artic.network/ncov-2019) (see supplementary material) and libraries were sequenced using the Oxford Nanopore Technologies portable genome sequencer, MinION. A genome reference-based assembly pipeline was used for consensus sequence generation with a minimum coverage depth of 20x. See appendix p 3-4.

Analysis

SARS-CoV-2 sequences from Brazil with collection date up to the 20th May 2020 (most recent date in our HCW dataset) (n=1860) were downloaded from GISAID (26–28) and appended to a previously described global dataset of 1,182 viral genomes. The resulting dataset was aligned to the reference NC_045512.2 using MAFFT v 7.450 (29) and manually edited using AliView. As previously described (30), we further filtered down our dataset by maintaining only sequences with at least 75% consensus sequence coverage. TempEst v.1.5.3 (31) analyses and visual inspection of the alignment in AliView were used to identify and remove sequences with unusual divergence. No recombination signal was found using RDP 4 (32). Three final datasets were generated: Dataset 1 consisted of 2,550 sequences, including 340 sequences from this study; Dataset 2, sequences with >90% coverage
(n=2259); and Dataset 3, a reduced version of Dataset 2, including all HCFMUSP sequences with coverage >90% (n=234) (see supplementary material). Pangolin version V3.1.11 (33) was used for lineage assignment.

Maximum likelihood phylogenies was inferred using IQ-TREE v.2.0 (34) under the best substitution model as determined by ModelFinder (35) implemented in the IQ-TREE pipeline. Bayesian time-rooted phylogenies for Dataset 3 were estimated using BEAST v1.10.4 (36) running with BEAGLE (37) and a discrete phylogeographic approach was used to understand the temporal dynamic of hospital-associated SARS-CoV-2 transmission. Hospital-associated clusters were defined according to the content of HC sequences (sequences from this study) and according to the statistical support obtained from the phylogenetic analysis. Compartmentalization analysis was perfomed using Simmond's Association Index implemented in the Hypothesis Testing Using Phylogenies (HYPHY) (38), and a binomial logistic regression analyses was performed to identify patient characteristics associated with clustering of genomic sequences. Results were reported as the odds ratio (OR) over the baseline variables and p values <0.05 were considered statistically significant. For the household geographical distances analysis, a Mann-Whitney U test was performed in R Studio 1.2.1335. See appendix p 4-9.

Results

Epidemiological context

From 25 February 2020 to 8 June 2020, the municipality of São Paulo, Brazil reported a total of 75,699 COVID-19 cases (Figure 1a). The first SARS-CoV-2 positive cases from HC-FMUSP were reported on epidemiological week 11 (from 9 March 2020), with HCW cases across all nine different institutes. In the following week, the first COVID-19 patient from HC, a community-acquired case, was hospitalized. On 30 March 2020, Institute B was converted into a COVID-19-only institute, while the other institutes were required to transfer COVID-19 patients to Institute B (Figures 1a and 1b).

A total of 12,134 SARS-CoV-2 tests were performed on samples from patients and symptomatic HCW working at the HC-FMUSP hospital complex, of which 3,933 (32%) were positive (appendix p 10). 3,898 SARS-CoV-2 positive individuals were included in this study, 2,008 (51.5%) patients and 1,890 (48.5%) HCW (Figures 1a). A flowchart with the complete study design can be seen in appendix p 10. Given that 91.8% (3,578) of cases from HC were reported by only three institutes - Institute B (2,159 cases, 55.4%), Institute C (716 cases, 18.4%), and Institute A (703 cases, 18%) - we explored the potential differences in the COVID-19 transmission dynamics between non-COVID-19 and COVID-19 institutes (Figure 1B and appendix p 11). A summary of epidemiological and sociodemographic characteristics from all cases in this study can be seen in appendix p 29.

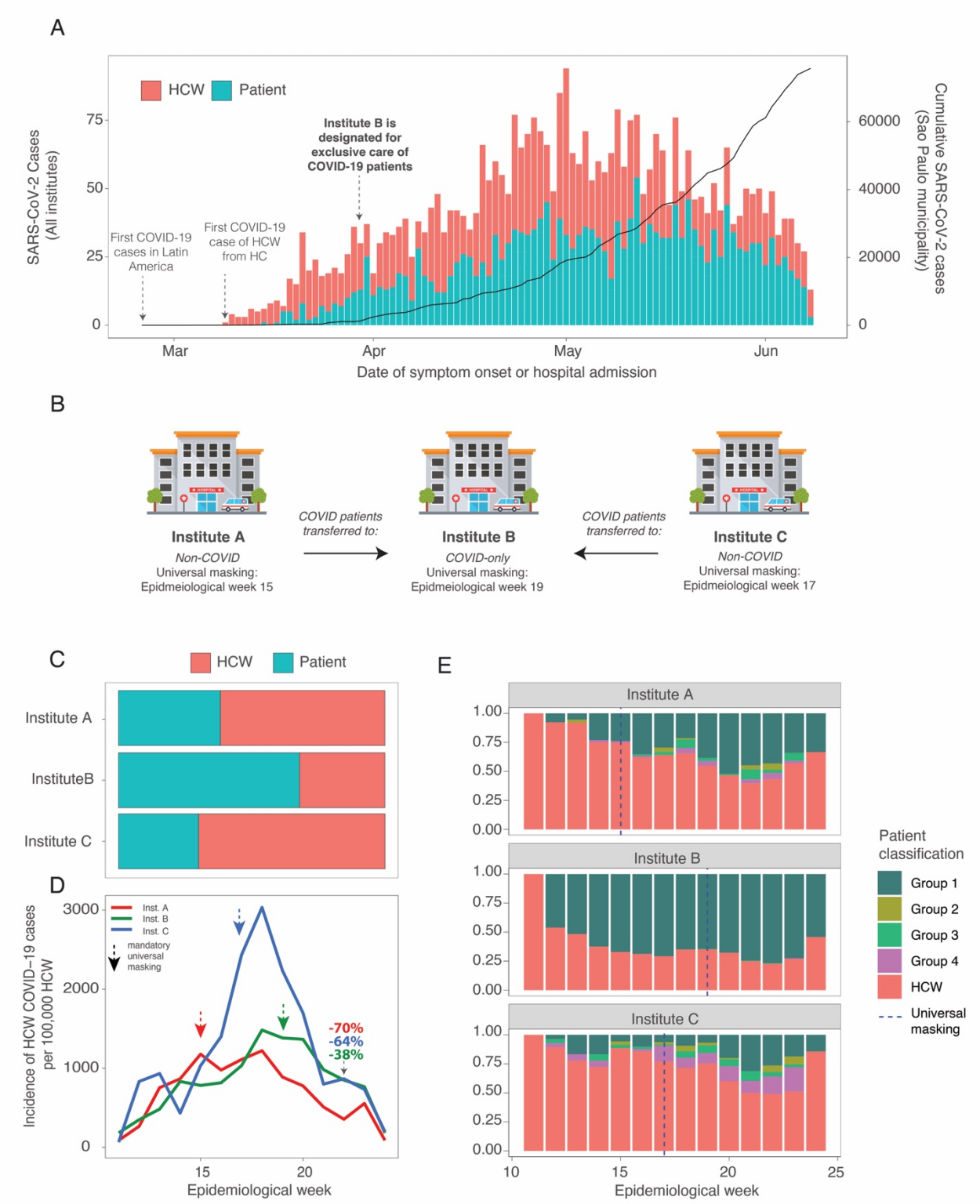


Figure 1. Epidemiological context of HC SARS-CoV-2 hospital-associated transmission. **(A)** Time series of COVID-19 positive cases across all institutes from Hospital das Clínicas (HC-FMUSP) and cumulative COVID-19 cases for the municipality of São Paulo. Colors depict whether cases occurred in HCW (red) or patients (blue). The dotted line marks the date of adoption of universal masking. **(B)**

Proportion of cases from A, B, and C institutes stratified by HCW/patient. **(C)** Incidence of HCW COVID-19 cases from Institute A (red), Institute B (green), and Institute C (blue) per epidemiological week. The dotted line marks the epidemiological week of adoption of universal masking in each institute. Percentages represent the reduction in the incidence of HCW COVID-19 cases for each institute at week 23, having week 18 as a reference (the week before universal masking was implemented). Percentage colors follow the pattern for line colors and represent the different institutes. **(D)** Proportion of patients according to time between the onset of symptoms and hospitalization. Patients were discretized into four groups: group 1 (community-acquired), group 2 (indeterminate acquisition), group 3 (suspected hospital transmission), group 4 (hospital-acquired) (see methods).

Epidemiological evidence for hospital-associated transmission

Most COVID-19 cases in non-COVID-19 institutes occurred in HCW, 61.7% in Institute A and 70% in Institute C; in contrast to 32% in Institute B (Figure 1b). Incidence of HCW cases was also higher amongst non-COVID-19 institutes in most epidemiological weeks before the enforcement of universal masking, especially for Institute C, reaching a peak incidence of 3,033/100,000 HCW (Figure 1D). Incidence of HCW cases decreased by 70% in Institute A, 64% in Institute C, and 38% in Institute B, after 7, 5, and 3 weeks of the implementation of universal masking, respectively - note that universal masking was adopted at different epidemiological weeks: 15 (A), 17 (C) and 19 (B) (Figure 1C).

We used the time gap between symptom onset and patient hospitalization as a proxy for SARS-CoV-2 hospital transmission and categorized patients into four groups (see methods). A total of 167 patients (8.55 %, total of 1,952 patients) from the three institutes had symptom onset >2 days after hospitalization (groups 2-4) and were possibly linked to hospital-associated transmission (Figure 1d and S3a): 27 (16.2%) belonged to group 2 (indeterminate acquisition), 54 (32.3%) to group 3 (suspected hospital transmission), and 86 (51.5%) to group 4 (hospital-acquired). The majority of group 4 COVID-19 cases occurred at Institute C, 70 (81.4%), while 15 (17.4%) occurred at Institute A, and one at Institute B (1.2%) (Figure 2a, appendix p 32). 52 (74.3%) group 4 cases from Institute C were reported after universal masking was already mandatory (Fig. 1D). The time between

hospitalization and symptom onset for group 4 patients ranged from 15 to 175 days (mean=42.5 days, median=27.5 days) (appendix p 12).

Hospital-associated SARS-CoV-2 genetic diversity and clustering

To further support the epidemiological evidence of hospital-associated transmission and characterize its dynamics, we randomly selected 454 SARS-CoV-2 positive samples (hospital workers and patients). From those, we obtained a total of 340 SARS-CoV-2 genomes with coverage >75%, approximately 10% of all reported cases (67 new GISAID submissions, appendix p 13 and 34-45). Most sequences from the three institutes belonged to lineage B.1.1.28, followed by lineages B.1.1.33, B.1 and B.1.1, the main lineages circulating in São Paulo at the time (appendix p 14).

Using datasets B and C (>90% coverage sequences), we were able to identify 16 clusters potentially associated with hospital transmission in the three institutes, comprised by a total of 73 sequences (Figure 2 and appendix p 46). Cluster size ranged from 2-12 sequences from this study, with an average cluster size of 4.6 sequences (median=3.5) (Figures 3a and b, appendix p 47). Within-cluster diversity was on average 1.22 SNPs (median=1, range 0-6 SNPs) (Figure 3D). We have also found maximum within-cluster pairwise diversity to be correlated to cluster duration (days), R2=0.6, and Spearman's rho = 0.56 (Figure 2B). Most clusters, 12 (75%), were defined by one single mutation (appendix p 49).

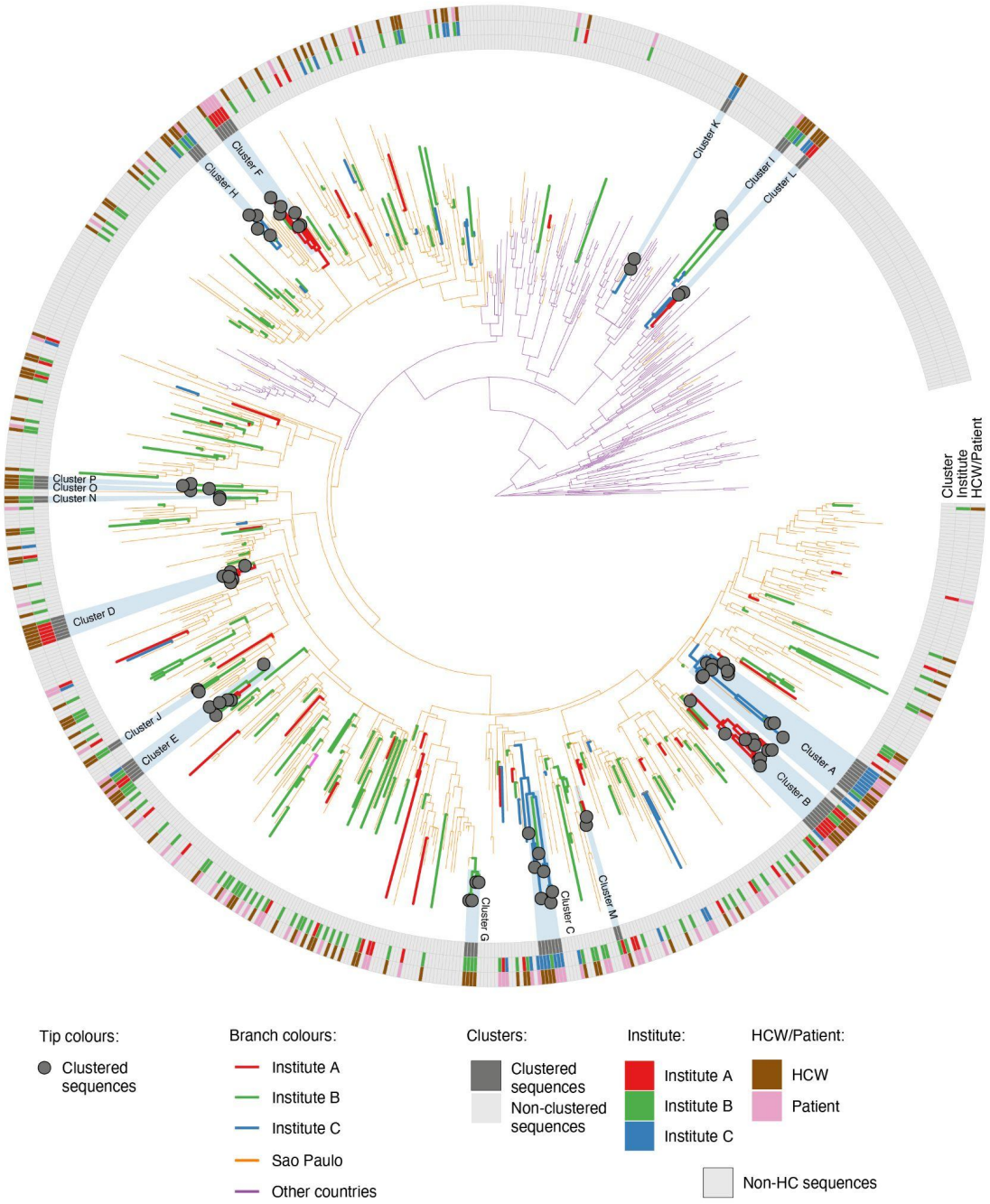


Figure 2. Hospital-associated SARS-CoV-2 transmission clusters. Time-stamped maximum clade credibility phylogeny inferred from dataset 3 (841 sequences), including 234 HC sequences (see methods). Regression of root-to-tip genetic divergence against sampling dates retrieved an R^2 of 0.57 (appendix p 15). Tips are colored according to hospital-associated transmission clusters and branches are colored according to inferred node location (Institutes A, B and C, São Paulo state, and Other).

Heatmaps depict the institute of collection for each HC sample and information on whether a sample belonged to an HCW or a patient. An expanded version of Figure 2 can be found in appendix 2.

Most clusters, 13 (81.25%), were mostly composed (≥70%) by sequences from a single institute (Figure 3C). Most clustered sequences belonged to HCW, 50 (68.5%) (figure 3D). Half of the clusters contained only sequences from HCW, while no clusters consisted of sequences from patients only. We also observe that only 19.7% of the sequences from Institute B are clustered; this proportion was 43.4% and 50% for Institute A and Institute C, respectively (Figure 4C). Moreover, despite presenting the largest number of clusters (n=6), Institute B clusters had the smallest average size, 2.5 sequences (median=2, range 2-4 sequences), and duration, 2.8 days (median=1.5 days, range 0-7 days), while Institute C had the smallest number of clusters (n=3), but the highest average cluster size (median=7, range 2-12 sequences), and the longest average duration, 23.7 days (median=27 days, range 4-40 days). Other cluster characteristics can be seen in appendix p 17 and 47.

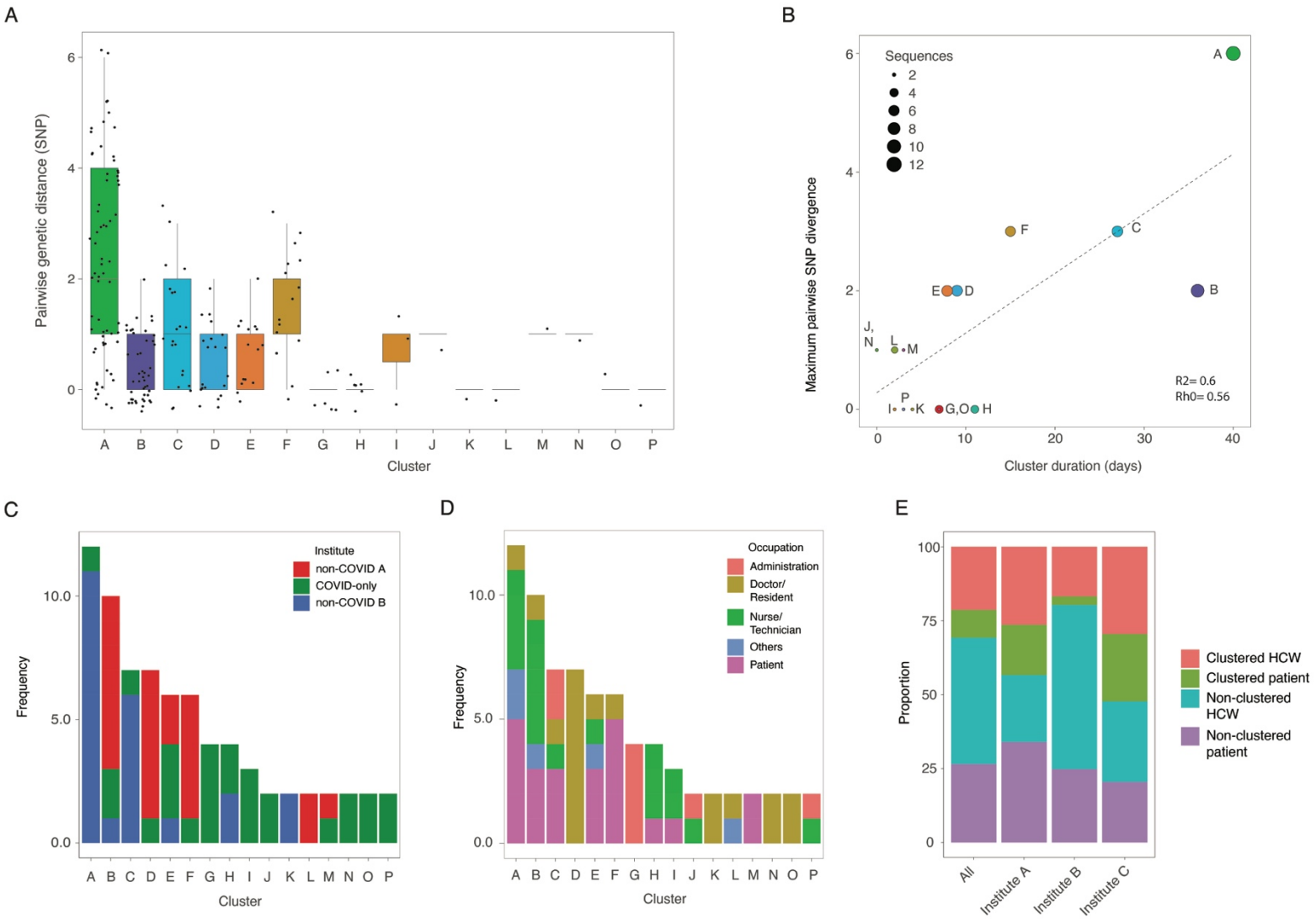

A
Pairwise genetic distance (SNP)
Cluster
B
Maximum pairwise SNP divergence
Cluster duration (days)
Sequences
R2= 0.6
Rh0= 0.56
C
Frequency
Cluster
Institute
non-COVID A
COVID-only
non-COVID B
D
Frequency
Cluster
Occupation
Administration
Doctor/ Resident
Nurse/ Technician
Others
Patient
E
Proportion
All
Institute A
Institute B
Institute C
Clustered HCW
Clustered patient
Non-clustered HCW
Non-clustered patient

Figure 3. Characteristics of the 16 phylogenetic clusters potentially associated with hospital transmission. (A) Pairwise genetic distances of sequences in each cluster. (B) Correlation between cluster maximum pairwise genetic divergence and cluster duration. (C) Frequency of sequences in each cluster according to the institute of origin. (D) Frequency of sequences in each cluster according to occupation. (E) Proportion of sequences per institute according to clustering status. Proportion was calculated considering a total of 234 sequences with coverage >90% used for cluster analysis.

Factors linked to hospital-associated clustering

To explore the differences between clustered (n=72) and non-clustered sequences (n=162) across the three institutes, we used logistic regression models to assess predictors that would best explain such patterns. Model 1 revealed that sequences from non-COVID-19 Institutes C and A were at greater odds for clustering than sequences from Institute B (COVID-only) (OR=4.46; and OR=3.59, respectively) (Table 1). HCW from both Institutes A and C were nine times more likely to cluster than patients from Institute B (OR=9.92; OR=9.21, model 2), while Institute B HCW have a non-significant tendency for higher odds of clustering (OR=2.72, p=0.08). Across all institutes, medical residents were the only occupation at greater odds for clustering compared to patients (OR=4.10, p=0.0007, appendix p 51-53, model 3). OR for different occupations in each institute can be seen in appendix p 51-53.

Table 1. Age and Sex-adjusted Odds ratio and p-values for clustering logistic models 1 and 2

Logistic Model Parameters	**Level**	**aOdds Ratio**	**p-value**
Model 1:			
Variables: Institute + HCW/patient +			
Age + Sex			
Base level: Institute B; Patient	(Intercept)	0.17	**0.002**
	Institute A	3.48	**0.00074**
	Institute C	4.17	**0.0002**
	HCW	1.63	0.21
Model 2:			
Variables: HCW/patient per Institute			
+ Age + Sex			
Base level: Patient.Institute B*	(Intercept)	0.11	**0.0034**
	HCW.Institute A	7.95	**0.0033**
	HCW.Institute B*	2.36	**0.017**
	HCW.Institute C	7.49	**0.0043**
	Patient.Institute A	4.45	**0.027**

	Patient.Institute C	7.43	**0.0050**

HCW: healthcare worker; * Institute B was COVID-19-only.

To assess the degree of compartmentalization of clustered sequences given different traits, we used the tree-based method Simmonds Association Index (AI). When the analysis was performed on all clustered sequences (n=72), a compartmentalization signal was identified for the trait institute (AI =0.45, BS=1000). Although all institutes are contributing to the signal, Institute A (AI= 0.37, BS=1000) and Institute C (AI=0.44, BS=999) have higher degrees of population structure when compared to Institute B (AI= 0.56, BS=998). However, analysis of sequences from each institute separately revealed a compartmentalization signal for traits HCW vs Patients (AI =0.36, BS=999) and occupation (AI =0.42, BS= 1000) for Institute A sequences only (appendix p 54).

Dynamics of hospital-associated SARS-CoV-2 transmission

We next used genomic data to infer the proportion of imported cases from each Institute, and in turn, infer the extent to which hospital-associated transmission happened within each institute. Time-measured phylogeographic analysis performed on dataset 3 revealed that Institute B had the highest proportion of import-associated cases (from São Paulo, international or other institutes), 83.64% (BCI: 76.64 to 92.70%), followed by Institute A, 67.85% (BCI: 60.38% to 71.70%), and Institute C, 57.70% (BCI: 52.23% to 63.63%) (Figure 4a). These data also imply that Institute A , and especially Institute C, would have had a higher degree of hospital-associated SARS-CoV-2 transmission than Institute B. To validate our results and take potential sampling bias into account, we also estimated the expected percent of imports for each institute by randomly reshuffling the institute assigned to each sequence. Averages for ten simulations and individual runs are shown in appendix p 18-19, and confirm that the expected average percent of imports for Institute C and Institute A should

be higher than the observed while that of Institute ? should be lower. Similar results are observed when sequences were discretized as from a HCW or patient (appendix p 20-22).

We also identified the location transition rates with strongest statistical support (Bayes Factor > 10) (39). Interestingly, strong statistical support was found for transitions from non-COVID institutes A and C to Institute B (Bayes factor = 113.7 and 84.4, respectively) (Figure 4b). Information on all location transition counts, rates, and Bayes factors can be found in appendix p 55. Temporally, transitions from institutes A and B into institute C peaked 3 weeks after the peak of São Paulo imports into these institutes (Figure 4C). Similar results can be observed for transitions between HCW and patients (Figure 4D).

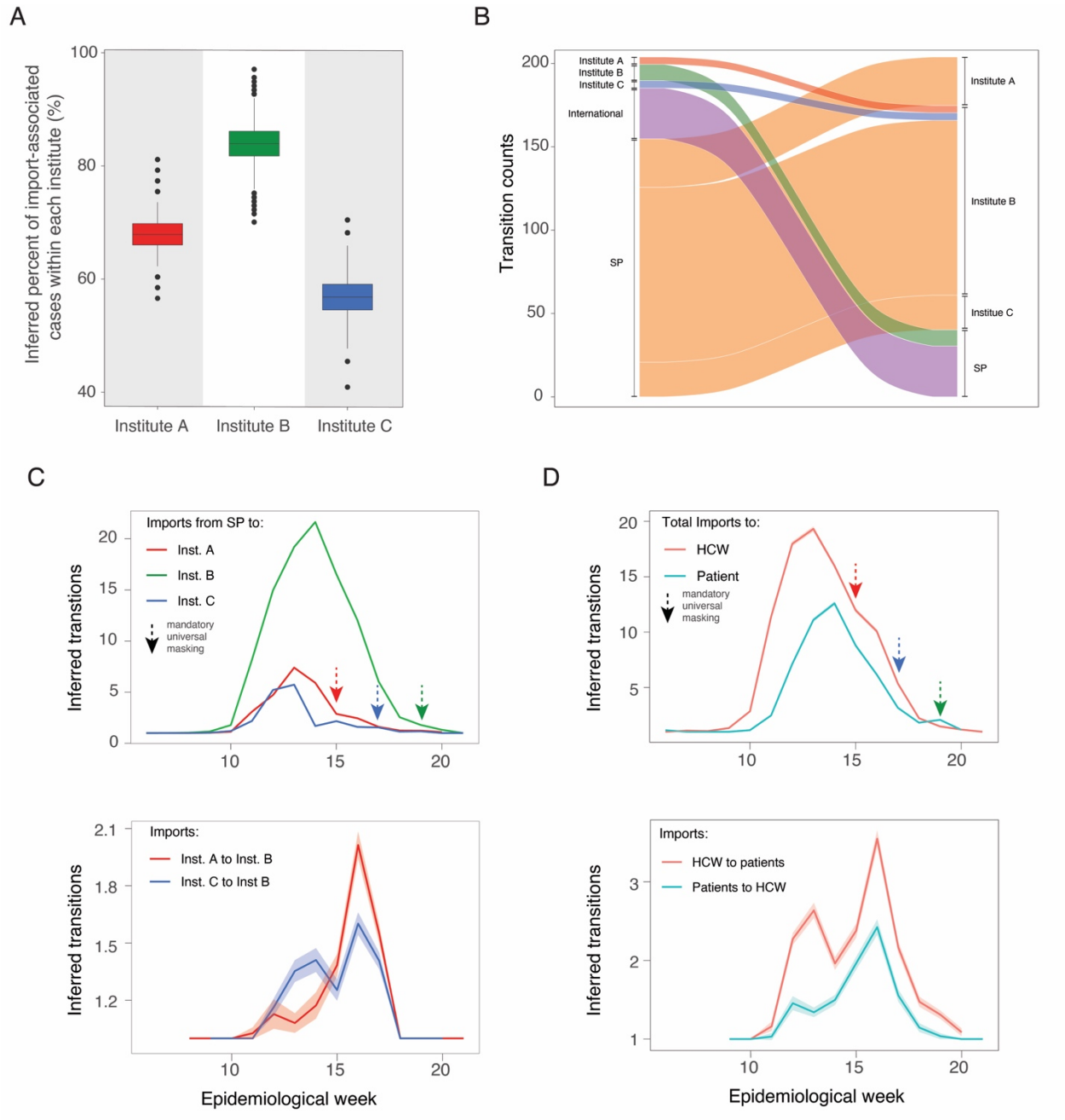


Figure 4. Proportion of SARS-CoV-2 imported cases in HC institutes and inferred transitions into and between institutes. (A) Proportion (%) of inferred total imports to Institutes A, B, and C. (B) Location transition counts (Markov jumps) for transition rates with strong statistical support (BF>10). Alluvial plots are proportional to the Markov jumps counts for each specific location transition. Colors identify the location of origin for each transition: Institute A (red), Institute B (green), non-COVID-B (blue), Other (purple), São Paulo (orange). (C) Inferred transitions (Markov Jumps) from São Paulo to each

institute over time (up) and between institutes (down). (D) Inferred total imported HCW and patient cases (up) and between HCW and Patients (down).

Epidemiological links of hospital-associated transmission clusters

We categorized the epidemiological link for each individual as strong, possible, and unclear (see methods). We find that 31 (42.5%) of the individuals had a strong epidemiological link with at least one other individual from the cluster, 31 (42.5%) had a possible link, and 11 (15%) had an unclear link (appendix p 23 and 47). Full information on epidemiological links for each cluster can be seen in appendix p 24-27. To exclude household interaction between clustered individuals, we used the pairwise distance between sequences in the same cluster and compared it to non-clustered sequences. Clustered sequences tended to come from cases residing in slightly more distant households (median=18.26 km) compared to non-clustered sequences (median= 16.2 km) (p-value=0.07, Mann-Whitney test) - (appendix p 28 and 47-50).

Discussion

While most COVID-19 hospital-associated transmission studies have focused on single hospitals or on mixed hospital data, we provide the first comparison of SARS-CoV-2 transmission across COVID-19 and non-COVID-19 hospitals during the early stages of the pandemic. We provide evidence of higher hospital-associated SARS-CoV-2 transmission in non-COVID-19 hospitals compared to a COVID-19-only hospital using different types of data and analyses. First, we suggest that universal masking was effective in decreasing infections across HCW but not hospital-acquired patient cases in non-COVID-19 institutes. Secondly, we show that positive cases from non-COVID-19 institutes in general and of HCW-only are more likely to be part of a transmission cluster than those from the COVID-19-only hospital. In addition, we estimate that a smaller proportion of the cases in non-COVID-19 institutes were acquired outside of the hospital. Finally, our genetic analyses further identified some level of virus transmission from non-COVID-19 institutes to the COVID-19-only.

Several studies have shown that HCW are at higher risk of COVID-19 infection than the broader community (38,40–43) and have an important role in seeding and amplifying nosocomial SARS-CoV-2 outbreaks to other HCW and patients (44). However, most of this evidence was generated prior to implementation of universal masking, which is effective in reducing the HCW risk of SARS-CoV-2 hospital-acquired infection (45). In turn, our study period includes the early stages of the pandemic and overlaps the progressive implementation of universal masking. We find that incidence amongst HCW was much higher in non-COVID-19 institutes, especially institute C. Considering that universal masking was implemented first at non-COVID institutes (A and C), and assuming that outside work exposure was the same for HCW from the three institutes, these differences could be explained by differences in behavior and risk perception of SARS-CoV-2 among HCW. Adherence to protective measures is correlated to risk perception and HCW tend to associate a higher risk of exposure to contact with infected patients rather than other HCW (46) (47). However, two studies conducted in São Paulo concluded that HCW who directly provided care to COVID-19 patients were not at higher risk

of infection (48,49). During the first COVID-19 wave in the UK, transmission between HCW was the most common form of nosocomial COVID-19 infection (50). Evidence suggests that after the implementation of universal masking, most of the HCW COVID-19 cases were associated with transmission between HCW rather than contact with an infected patient (50) (51). Thus, one of the possible explanations to our findings is that risk perception was higher amongst HCW from institute B, given the awareness of dealing with COVID-19 patients.

Although universal masking was effective in reducing infection amongst HCW, hospital-associated patient cases were still common in institutes A and C. This suggests that patient-to-patient transmission might have also played a role in the nosocomial transmission dynamics. In fact, evidence suggests that patients with hospital-acquired COVID-19 infections are more likely to get infected through contact with other patients in super-spreading events rather than through contact with HCW (52), especially if contacts also had hospital-acquired infections (53). Moreover, the proportion of patient-to-patient transmission almost doubled in the second wave in the UK and became the most common form of COVID-19 nosocomial transmission (50).

Given that Institute B was a COVID-19-only institute, drawing any conclusions on COVID-19 transmission from epidemiological data alone would be challenging, as patients from this institute were already infected at hospitalization and HCW could have been infected outside the hospital. To overcome this limitation, we used SARS-CoV-2 genome sequences from patients and HCW to infer SARS-CoV-2 transmission dynamics. Transmission clusters were larger in the non-COVID-19 institutes, suggesting that transmission in non-COVID-19 institutes involved sustained onward transmission for longer periods. We also showed that individuals from non-COVID-19 institutes were more likely to be part of a transmission cluster, especially if they were HCW. These findings most likely reflect the within-hospital interactions of HCW, but potentially the social interactions outside work as well. Lunch and smoking breaks are common situations in which unprotected interaction

between HCW has been documented (51,54). HCW cases have also been shown to result from unprotected interactions with other HCW in the community (55). Interestingly, our genetic analysis supports these findings when showing that a higher proportion of cases in non-COVID institutes should be associated with hospital transmission. These differences were observed for both HCW and patients.

While evidence for compartmentalization was observed at the institute level, significant links from non-COVID institutes to institute B were inferred from our genetic analysis. Since HCW could not transit between institutes, these inter-institute transmission events could be explained by patients from Institute A and C being transferred to Institute B and/or by HCW potentially interacting with HCW from other institutes outside of the hospital setting. Transportation of patients from non-COVID-19 hospitals to COVID-19 hospitals is critical and specific protocols should be in place to ensure patient and HCW safety (56,57).

Our study has several limitations. Firstly, this is a retrospective study and, as such, it faced limitations regarding access to samples and full metadata patients and sequences. Secondly, 6.5% (n=234, >90% coverage) of all cases were used for cluster analysis to reduce the chances of poor phylogenetic placement. Although this number represents one in every 15 cases, we have likely missed some transmission clusters, especially smaller ones, and intermediate transmission events. In addition, SARS-CoV-2 has a relatively low mutation rate (30) and it is possible that some phylogenetically related sequences might not be an immediate part of the same transmission chain, especially in clusters to which no epidemiological link could be observed. Finally, given that most of our sequencing sampling dates back to before universal masking was implemented, we were not able to assess the impact of universal masking using genomic data.

By integrating genomic and epidemiological surveillance, hospitals can identify and understand outbreaks and inform targeted infection control interventions. At a time in which new variants are constantly arising and inaccurate risk perception and COVID fatigue become increasingly relevant issues, it is important to emphasize that HCW can become vectors of transmission to other HCW and to non-COVID-19 patients; therefore, interventions towards improving compliance to protective measures should be implemented. Masks should be worn at all times when social distancing is not possible, not only when in contact with COVID-19 patients, including outside and on the way to the hospital. Finally, tighter protective measures (e.g., continuous HCW testing, restriction of visitors, immediate isolation of suspected patients) should also be enforced in non-COVID-19 hospitals, as community introductions can easily become within-hospital outbreaks.

Funding

Medical Research Council-São Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0) (**http://caddecentre.org/**). N.R.F.: Sir Henry Dale Fellowship: 204311/Z/16/Z and Medical Research Council-Sao Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0); and Bill & Melinda Gates Foundation (INV-034540). DSC is supported by the Clarendon Fund, Oxford University Zoology Department and Merton College. FAPESP supports IMC (2018/17176-8), FCSS (2018/25468-9), JGJ (2019/12000-1), TMC (2019/07544-2), CAMS (2019/21301-5) and ECS (18/14389-0). FFMUSP supports MSR (FFMUSP No. 206.706). The Tropical Medicine Institute from the São Paulo University supports ECR, LMS, MCP, LJTA. CAPPES supports PSA (88887.596940/2021-00) and GMF (88887.571465/2020-00).

Contributors

Conception: DSC, IMC, MSR, ASL, ECS, NRF, SFC. **Acquisition:** DSC, IMC, MSR, ALM, FCSS, JGJ, ERM, TMC, CAMS, PSA, GMF, ECR, LMS, MCP, LJTA, CSA, RFS, RZ, CR, MEBS, MFV, RKLI, TG, TMS, EA, IOMS, EF, MSO, ASL, ECS, NRF, SFC. **Analysis:** DSC, IMC, MSR, ALM, BG,

ECS, NRF, SFC. **Interpretation:** DSC, IMC, MSR, ALM, BG, ASL, ECS, NRF, SFC. **Drafting:** DSC, IMC, MSR, ALM, BG, ECS, NRF, SFC. **Revising:** All authors. Data and materials availability: All data, code, and materials used in the analysis are available in a dedicated GitHub Repository.

Declaration of interests

We declare no competing interests.

Acknowledgments

We thank L. Matkin (University of Oxford) for the incredible support throughout the development of this work. We also thank the contributions of the Núcleo de Vigilância Epidemiológica (NUVE) of the HCFMUSP. We thank the reviewers for their comments and suggestions and GISAID for the essential work of sharing genomic data for the global community and making this work possible. A list with full acknowledgments to all the authors publishing genomic data used in this study can be found in Data S1.

Supplementary Material

Supplementary appendix.

References

1. Jesus JG de, Sacchi C, Candido D da S, Claro IM, Sales FCS, Manuli ER, et al. Importation and early local transmission of COVID-19 in Brazil, 2020. Rev Inst Med Trop Sao Paulo. 2020 May 11;62:e30.

2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020 Feb 19;20(5):533–4.

3. Kluytmans-van den Bergh MFQ, Buiting AGM, Pas SD, Bentvelsen RG, van den Bijllaardt W, van Oudheusden AJG, et al. Prevalence and clinical presentation of health care workers with symptoms of coronavirus disease 2019 in 2 dutch hospitals during an early phase of the pandemic. JAMA Netw Open. 2020 May 1;3(5):e209673.

4. Mutambudzi M, Niedwiedz C, Macdonald EB, Leyland A, Mair F, Anderson J, et al. Occupation and risk of severe COVID-19: prospective cohort study of 120 075 UK Biobank participants. Occup Environ Med. 2020 Dec 9;

5. Barrett ES, Horton DB, Roy J, Gennaro ML, Brooks A, Tischfield J, et al. Prevalence of SARS-CoV-2 infection in previously undiagnosed health care workers in New Jersey, at the onset of the U.S. COVID-19 pandemic. BMC Infect Dis. 2020 Nov 16;20(1):853.

6. Jeremias A, Nguyen J, Levine J, Pollack S, Engellenner W, Thakore A, et al. Prevalence of SARS-CoV-2 Infection Among Health Care Workers in a Tertiary Community Hospital. JAMA Intern Med. 2020 Dec 1;180(12):1707–9.

7. Vahidy FS, Bernard DW, Boom ML, Drews AL, Christensen P, Finkelstein J, et al. Prevalence of SARS-CoV-2 Infection Among Asymptomatic Health Care Workers in the Greater Houston, Texas, Area. JAMA Netw Open. 2020 Jul 1;3(7):e2016451.

8. Mani NS, Budak JZ, Lan KF, Bryson-Cahn C, Zelikoff A, Barker GEC, et al. Prevalence of coronavirus disease 2019 infection and outcomes among symptomatic healthcare workers in seattle, washington. Clin Infect Dis. 2020 Dec 17;71(10):2702–7.

9. Lai X, Wang M, Qin C, Tan L, Ran L, Chen D, et al. Coronavirus Disease 2019 (COVID-2019) Infection Among Health Care Workers and Implications for Prevention Measures in a Tertiary Hospital in Wuhan, China. JAMA Netw Open. 2020 May 1;3(5):e209666.

10. El-Boghdadly K, Wong DJN, Owen R, Neuman MD, Pocock S, Carlisle JB, et al. Risks to healthcare workers following tracheal intubation of patients with COVID-19: a prospective international multicentre cohort study. Anaesthesia. 2020 Nov;75(11):1437–47.

11. Eyre DW, Lumley SF, O'Donnell D, Campbell M, Sims E, Lawson E, et al. Differential occupational risks to healthcare workers from SARS-CoV-2 observed during a prospective observational study. eLife. 2020 Aug 21;9.

12. Rickman HM, Rampling T, Shaw K, Martinez-Garcia G, Hail L, Coen P, et al. Nosocomial Transmission of Coronavirus Disease 2019: A Retrospective Study of 66 Hospital-acquired Cases in a London Teaching Hospital. Clin Infect Dis. 2021 Feb 16;72(4):690–3.

13. Kimball A, Hatfield KM, Arons M, James A, Taylor J, Spicer K, et al. Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility - King County, Washington, March 2020. MMWR Morb Mortal Wkly Rep. 2020 Apr 3;69(13):377–81.

14. Rife BD, Mavian C, Chen X, Ciccozzi M, Salemi M, Min J, et al. Phylodynamic applications in 21st century global infectious disease research. Glob Health Res Policy. 2017 May 8;2:13.

15. Aggarwal D, Myers R, Hamilton WL, Bharucha T, Tumelty NM, Brown CS, et al. The role of viral genomics in understanding COVID-19 outbreaks in long-term care facilities. Lancet Microbe. 2021 Sep 29;

16. Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. Progress and challenges in virus genomic epidemiology. Trends Parasitol. 2021 Dec;37(12):1038–49.

17. Walker A, Houwaart T, Finzer P, Ehlkes L, Tyshaieva A, Damagnez M, et al. Characterization of SARS-CoV-2 infection clusters based on integrated genomic surveillance, outbreak analysis and contact tracing in an urban setting. Clin Infect Dis. 2021 Jun 28;

18. Wong RCW, Lee MKP, Siu GKH, Lee LK, Leung JSL, Leung ECM, et al. Healthcare workers acquired COVID-19 disease from patients? An investigation by phylogenomics. J Hosp Infect. 2021 Sep;115:59–63.

19. Myhrman S, Olausson J, Ringlander J, Gustavsson L, Jakobsson HE, Sansone M, et al. Unexpected details regarding nosocomial transmission revealed by whole-genome sequencing of severe acute respiratory coronavirus virus 2 (SARS-CoV-2). Infect Control Hosp Epidemiol. 2021 Aug 20;1–5.

20. Cheng VC-C, Fung KS-C, Siu GK-H, Wong S-C, Cheng LS-K, Wong M-S, et al. Nosocomial outbreak of COVID-19 by possible airborne transmission leading to a superspreading event. Clin Infect Dis. 2021 Apr 14;

21. Løvestad AH, Jørgensen SB, Handal N, Ambur OH, Aamot HV. Investigation of intra-hospital SARS-CoV-2 transmission using nanopore whole-genome sequencing. J Hosp Infect. 2021 May;111:107–16.

22. Lucey M, Macori G, Mullane N, Sutton-Fitzpatrick U, Gonzalez G, Coughlan S, et al. Whole-genome Sequencing to Track Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Transmission in Nosocomial Outbreaks. Clin Infect Dis. 2021 Jun 1;72(11):e727–35.

23. Sikkema RS, Pas SD, Nieuwenhuijse DF, O'Toole Á, Verweij J, van der Linden A, et al. COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. Lancet Infect Dis. 2020 Nov;20(11):1273–80.

24. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated

COVID-19: a prospective genomic surveillance study. Lancet Infect Dis. 2020 Nov;20(11):1263–72.

25. Hamilton WL, Tonkin-Hill G, Smith ER, Aggarwal D, Houldcroft CJ, Warne B, et al. Genomic epidemiology of COVID-19 in care homes in the east of England. eLife. 2021 Mar 2;10.

26. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017 Mar 30;22(13):30494.

27. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. 2017 Jan;1(1):33–46.

28. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. Gisaid's role in pandemic response. China CDC Wkly. 2021 Dec 3;3(49):1049–51.

29. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772–80.

30. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020 Sep 4;369(6508):1255–60.

31. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016 Jan;2(1):vew007.

32. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 2015 May 26;1(1):vev003.

33. Rambaut A, Holmes EC, Hill V, OToole A, McCrone J, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. BioRxiv. 2020 Apr 19;

34. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020 May 1;37(5):1530–4.

35. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun;14(6):587–9.

36. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018 Jan;4(1):vey016.

37. Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, et al. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol. 2012 Jan;61(1):170–3.

38. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005 Mar 1;21(5):676–9.

39. Talbi C, Lemey P, Suchard MA, Abdelatif E, Elharrak M, Nourlil J, et al. Phylodynamics and human-mediated dispersal of a zoonotic virus. PLoS Pathog. 2010 Oct 28;6(10):e1001166.

40. Fenton L, Gribben C, Caldwell D, Colville S, Bishop J, Reid M, et al. Risk of hospital admission with covid-19 among teachers compared with healthcare workers and other adults of working age in Scotland, March 2020 to July 2021: population based case-control study. BMJ. 2021 Sep 1;374:n2060.

41. Shah ASV, Wood R, Gribben C, Caldwell D, Bishop J, Weir A, et al. Risk of hospital admission with coronavirus disease 2019 in healthcare workers and their households: nationwide linkage cohort study. BMJ. 2020 Oct 28;371:m3582.

42. Koh D. Occupational risks for COVID-19 infection. Occup Med (Lond). 2020 Mar 12;70(1):3–5.

43. Riley S, Ainslie KEC, Eales O, Jeffrey B, Walters CE, Atchison CJ, et al. Community prevalence of SARS-CoV-2 virus in England during May 2020: REACT study. medRxiv. 2020 Jul 11;

44. Abbas M, Robalo Nunes T, Cori A, Cordey S, Laubscher F, Baggio S, et al. Explosive nosocomial outbreak of SARS-CoV-2 in a rehabilitation clinic: the limits of genomics for outbreak reconstruction. J Hosp Infect. 2021 Aug 27;117:124–34.

45. Seidelman JL, Lewis SS, Advani SD, Akinboyo IC, Epling C, Case M, et al. Universal masking is an effective strategy to flatten the severe acute respiratory coronavirus virus 2 (SARS-CoV-2) healthcare worker epidemiologic curve. Infect Control Hosp Epidemiol. 2020 Dec;41(12):1466–7.

46. Fakih MG, Sturm LK, Fakih RR. Overcoming COVID-19: Addressing the perception of risk and transitioning protective behaviors to habits. Infect Control Hosp Epidemiol. 2021 Apr;42(4):489–90.

47. Advani SD, Yarrington ME, Smith BA, Anderson DJ, Sexton DJ. Are we forgetting the "universal" in universal masking? Current challenges and future solutions. Infect Control Hosp Epidemiol. 2021 Jun;42(6):784–5.

48. Costa SF, Giavina-Bianchi P, Buss L, Mesquita Peres CH, Rafael MM, Dos Santos LGN, et al. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Seroprevalence and Risk Factors Among Oligo/Asymptomatic Healthcare Workers: Estimating the Impact of Community Transmission. Clin Infect Dis. 2021 Sep 7;73(5):e1214–8.

49. Oliveira MS de, Lobo RD, Detta FP, Vieira-Junior JM, Castro TL de S, Zambelli DB, et al. SARS-Cov-2 seroprevalence and risk factors among health care workers: Estimating the risk of COVID-19 dedicated units. Am J Infect Control. 2021 Sep;49(9):1197–9.

50. Lindsey BB, Villabona-Arenas ChJ, Campbell F, Keeley AJ, Parker MD, Shah DR, et al. Characterising within-hospital SARS-CoV-2 transmission events: a retrospective analysis integrating epidemiological and viral genomic data from a UK tertiary care setting across two pandemic waves. medRxiv. 2021 Jul 19;

51. Richterman A, Meyerowitz EA, Cevik M. Hospital-Acquired SARS-CoV-2 Infection: Lessons for Public Health. JAMA. 2020 Dec 1;324(21):2155–6.

52. Illingworth CJ, Hamilton WL, Warne B, Routledge M, Popay A, Jackson C, et al. Superspreaders drive the largest outbreaks of hospital onset COVID-19 infections. eLife. 2021 Aug 24;10.

53. Mo Y, Eyre DW, Lumley SF, Walker TM, Shaw RH, O'Donnell D, et al. Transmission dynamics of SARS-CoV-2 in the hospital setting. medRxiv. 2021 May 1;

54. Schneider S, Piening B, Nouri-Pasovsky PA, Krüger AC, Gastmeier P, Aghdassi SJS. SARS-Coronavirus-2 cases in healthcare workers may not regularly originate from patient care: lessons from a university hospital on the underestimated risk of healthcare worker to healthcare worker transmission. Antimicrob Resist Infect Control. 2020 Dec 7;9(1):192.

55. Barry M, Robert AA, Temsah M-H, Abdul Bari S, Akhtar MY, Al Nahdi F, et al. COVID-19 Community Transmission among Healthcare Workers at a Tertiary Care Cardiac Center. Med Sci (Basel). 2021 Jun 30;9(3).

56. Yousuf B, Sujatha KS, Alfoudri H, Mansurov V. Transport of critically ill COVID-19 patients. Intensive Care Med. 2020 Aug;46(8):1663–4.

57. Liew MF, Siow WT, Yau YW, See KC. Safe patient transport for COVID-19. Crit Care. 2020 Mar 18;24(1):94.

Chapter 6

Discussion

The main aim of this thesis was to apply genomic and traditional epidemiology approaches to describe and understand the spread and evolution of SARS-CoV-2 in Brazil. This thesis builds upon the recent developments in high-throughput sequencing technologies, especially portable genomic sequencing, and in phylodynamic models to generate real-time insights to guide the COVID-19 pandemic response in Brazil. As such, this thesis is structured according to the different phases of the SARS-CoV-2 epidemic in Brazil: introduction and early response of SARS-CoV-2 (Chapter 1), nationwide spread and genomic diversity in the country's largest population hubs (Chapter 2), identification of the Alpha and Gamma variants of concern in Brazil (Chapter 3), and a retrospective high-resolution investigation of within-hospital transmission in the largest hospital complex in Latin America (Chapter 4).

I will start by summarising the main findings of each Chapter, while also discussing their main strengths and limitations, as well as their public health impact. I will then expand towards a more general discussion of the Brazilian response to the COVID-19 pandemic and the role played by genomic surveillance and epidemiology.

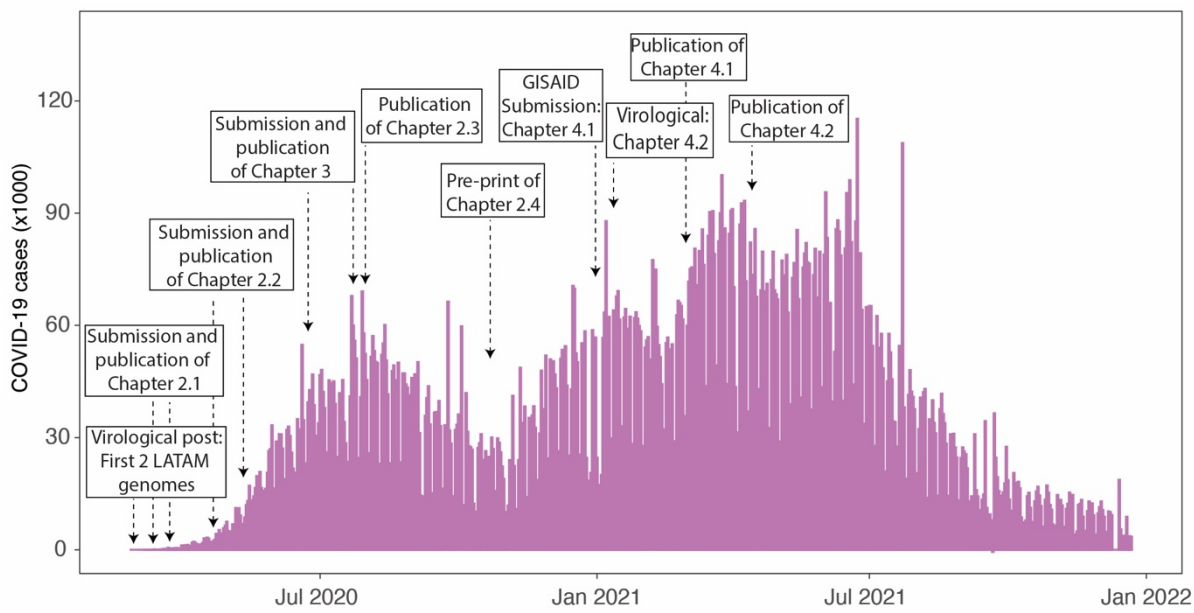


Figure 6.1. Overview of epidemiological scenario and timing of publication for the chapters presented in this thesis. Timeseries of Brazilian COVID-19 cases according to date of reporting. Boxes and arrows indicate the date in which Chapters of this thesis were submitted or published.

6.1 Chapter Summary: Strengths and Limitations

6.1.1. Chapter 2: SARS-CoV-2 importation, initial spread and response in Brazil

Chapter 2 is composed by my contributions to four publications investigating SARS-CoV-2 importation, initial spread and public health response in Brazil. Together with Chapter 3, it offers a valuable overview of the first months of the SARS-CoV-2 epidemic in the country.

In **Chapter 2.1**, "*Routes for COVID-19 importation in Brazil*", I show that most imported SARS-CoV-2 cases would be expected to come from Italy, China and France. In fact, Italy would be the origin location of five of the top 10 SARS-CoV-2 importation routes to Brazil, including the route Italy-São Paulo accounting for approximately one fourth of all potential SARS-CoV-2 imported cases. This manuscript was submitted for publication within 14 days of the first confirmed SARS-CoV-2 cases in Brazil and published 9 days later (Figure 6.1). By doing so, this analysis became the first source of information for health

authorities from all government levels to guide public health policy and response, including priority testing and quarantine of returning travellers from specific countries.

Despite its timeliness and early impact, this analysis has some important limitations especially related to data availability and the limited testing capacity at the early stages of the pandemic. To identify potential routes for SARS-CoV-2 importation in early 2020, data from the International Air Transport Association (IATA) on air travel from 19 countries to all Brazilian airports during February and March 2019 was used. By doing so, I captured the seasonal air travel changes associated with that time of the year, but I also assumed that passenger intensity and proportion would have been similar to pre-COVID-19 times. In addition, air travel data for Portugal, one of Brazil's main air travel corridors was not available, which likely resulted in an overestimation of the proportion of imports from other countries. In fact, some countries, e.g. China, had already limited inbound and outbound air travel from Wuhan since the 23rd January 2020 and several international flight cancelations nationwide from early February 2020 (1, 2).

Moreover, this analysis also relied on SARS-CoV-2 confirmed case counts. However, the limited capacity for SARS-CoV-2 testing and surveillance in many countries likely introduced a bias towards nations with greater surveillance systems. For example, although the United States accounted for more than half of the air passenger flow into Brazil, the US had an initially limited testing capacity (3), which resulted in it being only the 8th ranked country in terms of expected imports. Finally, subsequent analyses from our group investigating the travel history of the first confirmed cases during the first month of SARS-CoV-2 circulation in Brazil, confirmed that at first, most of the imported cases came from Italy, but revealed that the US surpassed Italy's proportion of imported cases by mid-March (4). This highlights the everchanging dynamics of virus outbreaks and public health interventions (including travel bans) during a pandemic, and how such risk importation

analyses should be used for improving early detection of the virus, e.g. by allocating molecular surveillance resources, instead of a prediction of the routes of variant importation throughout the different phases of the pandemic. Moving forward, such limitations could be overcome by refining mathematical models to account for time-changing testing capacity and/or using more reliable data sources such as daily deaths, excess deaths or severe acute respiratory infection (SARI) cases when available. Analyses should also be continuously updated with real-time flight data to provide more realistic measures of human mobility. Similar analyses were subsequently used to generate an estimated importation intensity (EII) measure in the UK and its temporal dynamics have been shown to follow that of imports estimated using genomic data (5-7).

Chapter 2.2 describes the first six genomes of SARS-CoV-2 in Brazil. This Chapter highlighted the importance of collecting travel history early in the pandemic to identify source location of cases and to accurately access local versus imported transmission. SARS-CoV-2's evolutionary rate leads to an average of 2 to 3 substitutions in every month, making it challenging to disentangle phylogenetic relationships between sequences derived from clinical samples collected just a few days apart. All the four returning travellers described in this Chapter reported travelling to Italy, further reinforcing the findings from Chapter 2.1. Part of this work was first published as a report on *virological.org* and figured as the first sequenced SARS-CoV-2 cases form Latin America. Such work was also a defining moment for Brazilian science and the Brazilian response to the SARS-CoV-2 pandemic, as genome sequences and *virological.org* report were available only 48 hours after sample collection, in a time when the average GISAID submission turnaround time was of 2 weeks (Figure 6.1). The impact of this work extended beyond the impact of the scientific discoveries being reported, as Dr Ester Sabino, Dr Jaqueline Góes de Jesus, several other team members and I represented the national response against SARS-CoV-2 in numerous media outlets. This

has increased communication and information transfer between scientists and general population as new findings and measures could be explained in near-real time to the Brazilian population.

The analyses highlighted in **Chapter 2.3** aimed to quantify SARS-CoV-2 transmission prior to the implementation of NPIs in four of the Brazilian states initially most affected by SARS-CoV-2 spread, according to the case count data. This was achieved by estimating the basic reproduction number, R_0, for these four states and Brazil, while putting such transmission into perspective against that of four European countries. In this Chapter, I described how basic transmission levels were very similar across the four Brazilian states from different parts of the country and how transmission in Brazil showed an initial tendency to be slightly higher than that observed in European countries, although with highly overlapping confidence intervals. Such analyses face important limitations. (i) SARS-CoV-2 surveillance in Brazil at the very beginning of the pandemic was very limited, with massively restricted access to molecular diagnostics and potentially delays in reports. Such limitations might partially explain the lower R_0 estimated for Ceará state, together with (ii) limited time points before the implementation of NPIs. (iii) The model used also assumes that delays in notification and testing capacity is the same between different locations, which is clearly a violated assumption. For future analysis, some of these limitations can be overcome by using deaths, excess deaths or SARI cases, rather than absolute case counts and by accounting for different average notification delays and different testing rates. However, even with such limitations, these estimates were similar to other R_0 estimates published at the time (8-12), including more comprehensive analysis performed by the Imperial College London in which a semi-mechanistic Bayesian hierarchical model is used to infer R from death counts and human mobility data from Google, while also accounting for underreporting (13).

Chapter 2.4 finishes the overview of COVID-19 in Brazil by providing an initial analysis to the largest and most comprehensive dataset on the adoption and easing of NPIs across the country, to date. This Chapter shows that NPIs were uniformly adopted across most Brazilian municipalities very early on in the pandemic, March 2020, and even before the detection of the first COVID-19 case in most Brazilian municipalities. However, such uniformity was followed by great fragmentation in the time of easing of NPIs, with neighbouring municipalities presenting completely different easing patterns. Such differences might mean that neighbour cities in Brazil were either experiencing COVID-19 epidemics at very different points in time or that the approaches taken in response to COVID-19 introduction and epidemic spread differed significantly even between municipalities in the same area. In fact, on 15th April 2020 Brazilian mayors and governors were ruled to be autonomous on their response to the COVID-19 pandemic by the Brazilian Supreme Court (14). This decision followed the lack of a clear national response orchestrated by the Brazilian Federal government and preceded the politicisation and polarisation of the COVID-19 response in Brazil, as reviewed by Borges and Renó (15). Brazil's president, Jair Bolsonaro, was internationally infamously recognised as one of the five presidents to downplay COVID-19 and the critical situation their own nations faced. Soon enough, the Brazilian response to COVID-19 was roughly divided between mayors and governors who were Bolsonaro's supporters and as such tended not to follow WHO recommendations and the rest of the country (15). However, it has been shown that regions with strong economic ties and, thus, human mobility connectivity are severely affected by the transmission happening in other regions, supporting the idea that decisions need to be made at central and local levels considering "within" and "between" aspects of transmission patterns (16). The main limitation of this Chapter is its very simple and descriptive analysis. In the future, data on human mobility, cases, deaths, hospitalisations and economic impact

of COVID-19 must be investigated in light of the data presented on this Chapter to understand the overall impact of the fragmented response observed in Brazil.

6.1.2 Chapter 3: Evolution and epidemic spread of SARS-CoV-2 in Brazil

This Chapter was built upon some of the main findings and methodologies presented in Chapter 2 and provided the largest, most comprehensive and the first near-real time published countrywide genomic and epidemiological assessment of SARS-CoV-2 transmission in Brazil. This Chapter used two separate approaches. (i) A semi-mechanistic Bayesian model was used to estimate R_t in the two largest cities in Brazil, São Paulo and Rio de Janeiro, from death counts and human mobility data. It showed that, although SARS-CoV-2 transmission in these cities was dramatically reduced after the implementation of NPIs, transmission levels in both cities plateaued at an R_t consistently >1, indicating insufficiency of the adopted measures to fully control SARS-CoV-2 transmission. (ii) To reconstruct SARS-CoV-2 spread in Brazil and provide insights into its genetic diversity and evolution in the country, this Chapter also applied both discrete and continuous phylogeograpic approaches to what, at the time, was the largest dataset of SARS-CoV-2 complete and near-complete genomes from Brazil, 427 sequences. These analyses revealed that Brazil experienced more than 100 SARS-CoV-2 introductions during the first two months after the first case was reported. From these introductions, three main transmission lineages were identified representing the geographical spread across different areas of the country: clade 1 (B.1.1.28) revealed the spread in the state of São Paulo; clade 2 (B.1.1.33) spread across all states in southeast Brazil and to other regions of the country; and clade 3 spread mostly in the state of Ceará. SARS-CoV-2 introduction across the three clades was timed between late-February and early-march 2020, in line with the reporting of the first cases on 26th February 2020. This Chapter also uncovered different stages of national

spread, with initial transmission events characterising as within-state events, followed by an increase in between-state and between-region transmission events. Such transmission patterns seemed to be linked to the effectiveness of NPIs in reducing within-state spread and the ununiform reduction in flights and air passengers across journeys of different durations: longer journeys were less affected than shorter ones, increasing the average distance travelled per passenger.

The planning and execution of this Chapter proved to be a massive logistic effort and involved several research institutions and collaborators from Brazil and beyond. Analyses, interpretation and writing were performed in real-time as our collaborators at the Universidade de São Paulo, Universidade de Campinas, Universidade Federal de Minas Gerais, Universidade Federal do Rio de Janeiro and Laboratório Nacional de Computação Científica received and sequenced SARS-CoV-2 positive samples from across the country and datasets were updated.

Such large-scale work incurred at least two main challenges. Firstly, creating a dataset which was representative of the transmission happening nationally but also locally was extremely difficult as access to samples from different states was limited and also given the logistical limitations of transporting such samples to the sequencing laboratories based in southeast Brazil. The proportion of SARI cases per state was used as a guide for the number of samples to be sequenced from each state to avoid over and under geographical representation. Although a great strategy on a nationwide basis, the large difference in the case counts of different states, meant that some states were represented by very few sequences, limiting our understanding of the genetic diversity and virus transmission to a more granular level. In fact, this work was followed by other important local studies describing additional international introductions and genetic diversity previously unidentified by the work presented in this Chapter (17-21). However, a recently published

pre-print analysing over 17,000 genome sequences from Brazil between February 2020 and June 2021 found similar results to the ones presented in this Chapter, further validating its results even when using almost 40 times more genomes (22). The study estimated 114 international SARS-CoV-2 introductions into Brazil to have happened by April 2020 and largely coming from Europe. These findings agree with the 102 introductions estimated by in Chapter using genomes collected by the end of April 2020 (22). The pre-print also found most of the between-region movement to have happened from the Southeast (over 40%), similar to the pattern of spread observed in this Chapter (22).

Secondly, the amount of genomic data being produced in response to the SARS-CoV-2 pandemic is the largest ever seen and working with such large datasets required having access to increased computational power. To overcome this challenge, this work was performed by downsampling SARS-CoV-2 genetic diversity from across the globe, which has been the core approach used for all other SARS-CoV-2 genomic epidemiology studies (23). In addition to the challenges listed above, the analysis in this Chapter could have been further improved by formally assessing potential drivers of virus lineage movement across the country and by attempting to infer population dynamics parameters from genomic data as a complementary approach to understand the impact of NPIs on virus lineage movement.

Since its publication, this study has become the main reference literature for understanding the introduction and initial spread of SARS-CoV-2 in Brazil. The timing of publication was also an extremely important factor for the impact and success of this study, as the initial pre-print was made available in early June 2020 and the final version was published in Science in late July of the same year (Figure 6.1), in a time when very few sequences from Brazil were available.

6.1.3 Chapter 4: SARS-CoV-2 variants of concern in brazil

While Chapter 3 described the first two months of SARS-Cov-2 spread in Brazil, the sustained transmission of SARS-CoV-2 in the country and worldwide maintained the necessary conditions for new SARS-CoV-2 lineages to emerge, some of which were later on identified as VOIs and VOCs. Chapter 4 constituted my contributions to two scientific publications on the identification of VOCs in Brazil and was subdivided into two other Chapters.

Chapter 4.1, "Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020", presented a very simple and short phylogenetic analysis focusing on the sequencing of the first 2 Alpha (B.1.1.7) variant cases from Brazil and identified its cryptic local transmission. While both samples were not genetically related to each other, one of the patients reported no international travel, but contact with a relative which had recently been abroad and tested positive for SARS-CoV-2 at the time of return. Much like Chapter 2.2, this Chapter highlights the importance of epidemiological data for contextualising genomic findings. Although Alpha was subsequently reported in at least 17 of the 27 Brazilian Federal units (24), Alpha's introduction in Brazil was met by the simultaneous emergence and spread of Gamma VOC in Manaus and subsequently across the entire country. Alpha would never become the dominant lineage in Brazil, as opposed to what happened in most of the globe (25-28).

In **Chapter 4.2**, "Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil", I presented comprehensive genomic and epidemiological analyses on the identification, emergence and spread of the new VOC P.1/Gamma in Manaus Brazil. Gamma's identification occurred right after Manaus started experiencing a major second wave of SARS-CoV-2 cases in December 2020 (29), and following the identification of the first VOCs, B.1.1.7/Alpha in the UK (30) and B.1.351/Beta in South Africa (31), just a few

weeks earlier. Gamma acquired 17 new amino acid mutations, 10 of which were located in the SARS-CoV-2 spike protein and 3 in its receptor binding domain. Given the large number of mutations acquired in a short period of time and the lack of evidence for a higher evolutionary rate within Gamma, this Chapter employed a Bayesian approach to test the hypothesis of a higher evolutionary rate along the branch leading to Gamma's emergence. This VOC was estimated to have emerged around mid-November 2020. Discrete phylogeographic analyses showed that Gamma emerged in Manaus and spread to other Brazilian metropolitan areas in similar pattern to that observed for air travel destinations from Manaus.

There were two main challenges associated with the detection of the first Gamma cases in Manaus, Amazonas state. First, by the 12 January 2021 only 6 genomes were available from the Amazonas state, five of them reported by our team. The lack of continued and consistent genomic surveillance in Brazil and especially in the north region, including Manaus, greatly limited our early understanding of Gamma's emergence. In addition, the large number of new mutations in Gamma's sequence meant that existing sequencing primers had mismatches in key primer-binding regions, which ultimately led to the generation of a substantial proportion of SARS-CoV-2 genomes with <75% genome coverage. However, we found that our lower coverage genomes could still provide useful information on the frequency of the Gamma VOC over time. Thus, we maximized the information derived our sequence data by using both low coverage datasets (for VOC frequency assessments) and high coverage datasets (for VOC frequency assessment confirmation and Bayesian phylogenetic analysis).

Chapter 4.2 was firstly published on *virological.org* on the 12th of January 2021 (32), and preliminary analysis where shared earlier with the Secretary of Health in Manaus and Pan American Health Organization (10th January) and WHO (11th January 2021). Also on

the 10th January, Japanese authorities released on GISAID the first Gamma genomes from travellers returning from Manaus (33). By early January 2021, eight months after its first epidemic wave which infected nearly three-quarters of Manaus' population (29, 34), Manaus was experiencing a second epidemic wave characterized by a rapid surge of COVID-19 hospitalizations and deaths. Together with independent laboratory assessment of neutralization data associated with spike mutations present in the Gamma VOC, our findings were critical to hypothesize that a significant proportion of cases in Manaus' second wave were being caused by a more transmissibility variant associated with immune escape (29). This was later confirmed by several reports of reinfections caused by the Gamma VOC (35-37). More generally, the rapid identification of Gamma's epidemic growth in Manaus, the identification of its mutations and the assessment of its potential for international spread were crucial for pandemic preparedness nationally and internationally. For example, on the 12 February 2021, after the UK Government become aware of our findings related to Gamma's early spread in Manaus, Boris Johnson implemented travel bans to several Latin American countries.

Subsequent work by researchers from the Oswaldo Cruz Foundation (FioCruz) further contributed to our understanding of the evolutionary history of SARS-CoV-2 in Manaus. For example, retrospective sequencing revealed that the first epidemic wave in Manaus was mainly caused by lineage B.1.195, which was later replaced by lineages B.1.1.28 and B.1.1.33 (18). The Gamma VOC then spread from the Amazonas state to all other Brazilian regions and caused a large second wave in Brazil. According to data from the FioCruz Rede Genômica Dashboard enabled by data available on GISAID (38), Gamma was the dominant lineage nationwide for 6 months, from February (59.7%) to July (87.6%) 2021, when a staggering total of 12.9 million new cases and 401,978 deaths were reported in Brazil. To put it into perspective, in the 12 months prior to Gamma's emergence (February

2020 to January 2021), Brazil reported 9.2 million cases and 154,392 deaths (39). Gamma was subsequently detected in 74 countries in the world, including most South American countries, where it also became the dominant circulating lineage (40, 41). A recent pre-print estimated that the North region seeded 47% of the Gamma infections nationwide and that Brazil exported Gamma on at least 316 occasions, mostly to other South American countries (65%), but also to Europe (14%) and Asia (11%) (22). Several Gamma-descendent lineages were also described during its circulation in Brazil and South America (40).

Since August 2021, Gamma was replaced by Delta and its descendent variants in Brazil and South America. As opposed to Gamma, which emerged before mass COVID-19 vaccination effectively kicked off in Brazil, Delta encountered a population with a different immunity profile, facing a greater proportion of individuals with some level of naturally or vaccine-mediated immunity. COVID-19 vaccination in Brazil started slowly between January and April 2021, and was characterised by a combination of factors, including lack of action from the Brazilian Federal government in securing sufficient and fast delivery of COVID-19 vaccines which resulted in vaccine delivery delays from the suppliers, and the politicisation of COVID-19 vaccines and upsurge of fake news regarding their efficacy and safety (42-45). However, by July 2021, with roughly 50% and 20% of the population vaccinated with one and two doses respectively, the epidemiological situation in Brazil started to change with cases consistently falling from an average of 70,000 to 3,000 new case notifications per day (46).

Encouragingly, mass vaccination may have started reducing death counts in April 2021, decreasing from a 7-day rolling average of over 3,000 daily deaths to as low as 58 deaths in December 2021 (46). At the time of writing, Omicron is the main lineage circulating in Brazil (38). Despite high vaccination rates, daily case count in Brazil are

increased with over 200,000 cases a day, in line with what has been seen in other countries across the globe (46).

6.1.4 Chapter 5: The dynamics of within-hospital SARS-CoV-2 transmission

While the other Chapters in this thesis focused on investigating SARS-CoV-2 spread and transmission at larger spatial scales, from municipality-level to countrywide, Chapter 5 presented analyses with finer spatial granularity. It provided the first comprehensive investigation of hospital-associated SARS-CoV-2 transmission in the early stages of its spread in Brazil. To the best of my knowledge, this is the first investigation of SARS-CoV-2 transmission dynamics between institutes from the same hospitals complex. Here, I provided epidemiological and genomic evidence of more pronounced SARS-CoV-2 transmission in non-COVID-19 institutes compared to COVID-19 institutes, even when mandatory universal masking was the rule across COVID-19 and non- COVID-19 institutes. Phylogenetic analysis further revealed that hospital-associated transmission clusters from non-COVID-19 institutes were larger and of longer duration compared to those from COVID institutes; and suggested that HCW and patients from non-COVID-19 institutes were at higher risk of being part of hospital-associated transmission clusters.

Although this analysis is being submitted for publication after almost 2 years from the early stages of the pandemic in Brazil, its findings remain extremely relevant. Despite of COVID-19 vaccines being now widely available in several countries, vaccination hesitancy amongst HCW ranges between 27.7% to 91.7% across different studies (47-58), and sometimes higher hesitancy rates amongst HCW that are more frequently exposed to COVID-19 and/or at higher risk for severe disease (56). This scenario is also aggravated by the emergence of new VOCs with high capacity of immune evasion, including Omicron (59). It has been estimated that effectiveness against symptomatic disease following a 2-

dose regimen of Pfizer or Astra Zeneca COVID-19 vaccines is reduced to around 30% when challenged against Omicron exposure (60, 61). Thus, adequate risk perception and proper use of PPE by HCW remains essential to avoid within-hospital outbreaks, shortage of HCW and to prevent exposure of at-risk patients and their families. In fact, several countries have been reporting shortage of HCW in hospitals resulting from the high transmission capacity of Omicron (62-64). Such scenario has pushed governments to reduce isolation duration for HCW so health systems can better cope with the large number of Omicron positive cases (65).

This study faced limitations regarding availability of standardised clinical and demographic metadata that was extracted from the hospital data database for this study. In addition, during the early stages of the pandemic, SARS-CoV-2 genome sequencing in Brazil, 39.6% (180/454) of the sequences had <90% genome coverage and were removed from the cluster analysis. However, to understand the impact of sequence removal and investigate the robustness of our findings, I found similar phylogenetic results when analysing datasets comprised by all sequences with genome coverage >75%. Existing clusters became larger, but no new clusters with adequate statistical support were observed (data not shown).

Finally, given the relatively low genetic diversity observed in SARS-CoV-2 datasets obtained during short periods of sampling collection and the relatively low genomic sequencing coverage of SARS-CoV-2 infected HCW and patients, the analysis in this Chapter was focused on general transmission patterns and characteristics of clusters across different institutes, rather than trying to determine infector and infected pairs. An analysis of such granularity could be achieved using a probabilistic framework such as the one used by Illingworth et al (66, 67) with data collected from a prospective study such as the one described by Meredith et al (68) or a retrospective study in a well-structured research-driven

hospital environment. A cohort where testing is routinely conducted in all symptomatic cases and their contacts, with standardised metadata collection and virus sequencing of all individuals identified by contact tracing could provide increased resolution and disentangle transmission within and between healthcare institutes at a finer resolution. In fact, it is possible that some of the clusters observed in this Chapter are not necessarily direct clusters of transmission, meaning direct pairs of infectors and infected, but rather transmission lineages which circulated in the institutes, and are represented by only a fraction of randomly sequenced cases involved in long chains of transmission. Such limitations also prevented us from investigating the transmission hotspots in each institute and directionality of transmission between different hospital occupations. In the future, prospective studies should be designed to further expand our understanding of SARS-CoV-2 associated hospital transmission.

6.2 The Brazilian response to SARS-CoV-2 and future steps

Brazil has been internationally recognized as one of the leading developing countries when it comes to responding to public health threats for its response against HIV/AIDS (69), for its world-renowned national vaccination programme (70), and being the only country with population >100 million to have a universal and free of charge health care system (71). However, Brazil has recently been facing an increasing economic instability and political turmoil. Health, education and research in Brazil have experienced major budget cuts, resulting in the lack of essential research and human resources, including the cancellation of several ongoing research scholarships, freezing of academic positions in Federal institutions and a diaspora of academic minds of unprecedented levels (72). Such a debilitating political and research scenario has had limited resources to tackle large epidemics caused by Zika, chikungunya, dengue and yellow fever viruses, and the SARS-CoV-2 pandemic.

The COVID-19 pandemic response in Brazil was led by a federal government leadership which undermines science and lacks a clear centralized strategy to mitigate virus spread and its toll on Brazil's public health system (73, 74). This led to a fragmented response at the state and municipality levels (75-77), as shown in Chapter 2.4. Brazil's president has been reported to undermine the severity of the disease by calling it a "little cold" (73), publicly arguing with the Minister of Health against the adoption of NPIs, ignoring scientists (73, 78), spreading fake news (76), not sharing data transparently (77, 79) and encouraging the use of scientifically proved ineffective therapies, such as hydroxychloroquine, while discouraging the administration of vaccines (80, 81).

Despite the vulnerable scenario for research and health, Brazilian scientists were able to very rapidly detect SARS-CoV-2 local circulation within just 48 hours of first cases, and swiftly identify new variants of concern in the country, as shown in Chapters 2.2, Chapter 3 and Chapter 4. For comparison, ZIKV circulated cryptically for up to 18 months before it was first detected in Northeast Brazil (82). As highlighted in Chapter 2.4, although the adoption of NPIs often occurred before most Brazilian municipalities had even identified their first cases, our data shows strong discoordination in the easing of NPI's even in municipalities within the same state and within the same region, despite unrestricted movement between Brazilian states during this period (71). Interestingly, COVID-19 deaths were found to geographically cluster approximately 1 month before the geographical clustering of cases became apparent, suggesting that case detection was limited by insufficient diagnostic capacity (71). Inequalities in access to COVID-19 diagnosis, vaccination and treatment have also been highlighted in other studies (4, 83). For example, a study built upon the results described in Chapter 2.3 found an association between higher income and access to COVID-19 diagnosis in the metropolitan area of São Paulo (4). Low-income and non-white populations from São Paulo were found more likely to be hospitalised

and die, and patients hospitalised in public hospitals were also more likely to die than patients hospitalised in private hospitals (83).

Work by Castro et al. published in April 2021 showed that the duration of SARS-CoV-2 clusters of cases and deaths did not reduce over time (71), indicating limited effectiveness of mitigation strategies countrywide. This is in line our findings on epidemic transmission in São Paulo and Rio de Janeiro cities (Chapter 3) and with an independent report from the Imperial College London (13). São Paulo was identified as the main super spreader city in Brazil accounting for 85% of the importations in the entire country (84), and that transmission of COVID-19 became more intense in the north and northeast regions after the first cases were reported in São Paulo (71). Such spatial trend is consistent with my phylogenetic analysis published in September 2020 (see Chapter 3), in which we showed an increase in between-state and between-region virus lineage movements after a short period dominated by within-state lineage transitions. Most of the long-distance virus migrations were originating from southeast Brazil region, which includes São Paulo and Rio de Janeiro states, the most populated and well-connected hubs in the country.

As of 28th January 2022, Brazil reported over 24.8 million cases and 625 thousand deaths, rendering it the third highest ranked country in number of cases and second in number of deaths globally (46). Currently, 69% of its population has received the 2-dose vaccine regime (85). According to the Global Health Security (GHS) Index 2019, Brazil was one of the best placed countries in terms of preparedness for responding to a biological threat. Out of 195 countries evaluated, Brazil was placed number 22nd in the overall score and top 20 in three of the six categories in which the index is organised: 16th in prevention of the emergence or release of pathogens, 12nd for early detection & reporting for epidemics of potential international concern and 9th in rapid response to and mitigation of the spread of an epidemic (86). In 2021, GHS ranked Brazil as number 43, after increasing its previous

overall score by 0.2, while losing 8.5 points in the "rapid response" category. However, assessment of the COVID-19 response of 98 countries performed by the Lowy Institute, as of January 2021, placed the Brazilian response as the worst in the world (98th) (87). In fact, studies have found GHS not to predict COVID-19 response and vaccine rollout (88-90). While the association of variables measured by the GHS index changed over time, COVID-19 outcomes were significantly associated to variables not directly measured by the index such as social cohesion, reduction in social polarisation and reduced perceptions of corruption. Future versions of such indexes must consider the inclusion of other sociodemographic, political and governance variables.

6.2.1 Metasurveillance of SARS-CoV-2 genomic sequencing in Brazil

To better understand diversity of SARS-CoV-2 viral lineages and contextualize the role of genomic surveillance in the response to the COVID-19 pandemic in Brazil, I present and discuss a preliminary analysis on 87,324 Brazilian genomes available on GISAID (available as of the 23rd of December 2021). In 2020, only 1,809 genomes were made available, with a median of 94 genomes/month (range 1-615). These numbers massively increased in 2021 with a total of 85,522 published genomes and a median of 6,699 genomes a month (range 631-20,381) (Figure 6.2 A). Most genomes belonged to samples collected in Southeast Brazil (58,060, 66.5%), especially from the states of São Paulo (44,834, 51.3% of Brazilian sequences) and Rio de Janeiro (10,405, 11.9%) (Figure 6.2 A and B). Although Southeast, North and Northeast Brazil figure as the regions with the highest sequencing rates (number of genomes sequences per 100,000 cases), when disaggregated by federal states, sequencing capacity in absolute numbers is positively correlated with state GDP (Figure 6.2 B, Rho = 0.8, p-value = 0.91, as recently seen worldwide (91). The median sequencing coverage per state in Brazil is 170.7 sequences/100,000 cases or 0.17%. São Paulo is the

state with the highest sequencing rate 1,007.5 sequences/100,000 cases or 1% of all cases, while Piauí has the lowest sequencing rate, 53.3 or 0.05%. Overall, sequencing rate in Brazil was 392.9 sequences/100,000 cases [excluding February 2020 when only 2 cases and 1 sequence (see Chapter 2.2) were reported], or 0.4% of all cases.

The median turnaround time, defined as the number of days between sample collection and GISAID submission, was 58 days (range 1-632). There was a non-statistically significant negative correlation tendency between GDP and turnaround times per states (Pearson rho = -0.34, p-value = 0.08796), with some states with higher GDP, e.g. Minas Gerais and Distrito Federal, performing worse than states with lower GDP, e.g. Alagoas, Paraíba and Pará (data not shown). The transition from 2020 to 2021 also saw a statistically significant decrease in the turnaround times (p-value < 2.2×10^{-16}), with a reduction in the median turnaround time from approximately 118 to 57 days, respectively (Figure 6.2 C). In fact, there was a constant increase in the median turnaround time per month all throughout the year of 2020, reaching a peak of approximately 222 days in December (2020 range 3-288 days), which was followed by a major reduction to a median of 95 days in January 2021 (range 1-632 days) and a nadir of 39 days in April of the same year (Figure 6.2 D). Such reduction was accompanied by an increase in the proportion of sequenced cases from July 2021 onwards, when COVID-19 incidence in Brazil started to decline, reaching a peak of 8,080 sequences/100,000 cases or 8% in December 2021(Figure 6.2 D). When disaggregated by year, the median sequencing rate increased 26-fold between 2020 and 2021, increasing from 15.3 (range 0.5-612) to 391 sequences/100,000 cases per month (range 41.3-8081) (Figure 6.2 D).

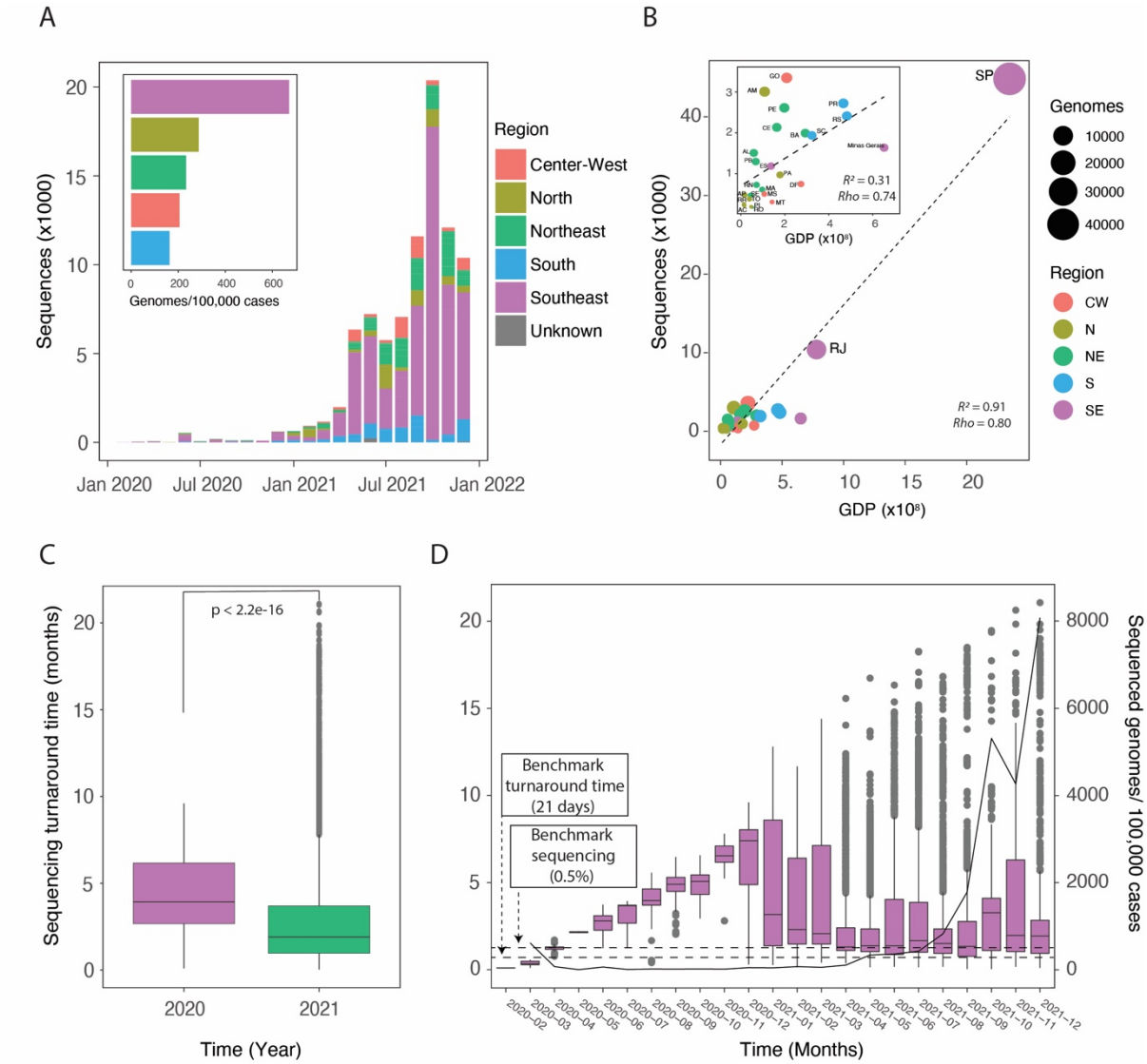


Figure 6.2. Overview of Brazil's SARS-COV-2 genomic surveillance performance. Metadata for all SARS-CoV-2 genome sequence GISAID entries from Brazil (n=87,324), regardless of their size, were downloaded on the 23rd December 2021. Case count data was downloaded from Brazil.io and Our World In Data on the 23rd December 2020. (A) Timeseries of all Brazil SARS-CoV-2 genome sequences according to submission date and stratified by Brazilian region of sample collection: Centre-West (red), North (yellow), Northeast (green), South (blue) and Southeast (pink). Inset shows genome sequencing effort per Brazilian region as per genomes/100,000 cases. Region are coloured coded similarly to the main figure. (B) Correlation between genomes sequenced and gross domestic product (GDP) by Brazilian Federative Unit (states). Inset shows a similar correlation plot excluding the states of São Paulo and Rio de Janeiro. Colours are coded according to figure XB. Circles are sized according to number of genome sequences. (C) Turnaround times per year of sequence submission (time between date of sample collection and date of GISAID submission). (D) Time series of turnaround times and sequencing efforts (sequences/100,000 cases) in Brazil according to month of sequence submission.

The descriptive analyses presented in Chapter 1 showed limited genome sequencing in Brazil across the last two decades, with a generation of <300 genome sequences/year (or

<25 per month) when not accounting for the pandemic years of 2020 and 2021, and a turnaround time ranging between 1.3 and 2.6 years since 2016. In comparison to the pre-pandemic scenario, Brazil produced over 285 times more viral genome sequences in 2021 alone. In addition, turnaround times were reduced by at least 8 times in 2021 when compared to the lowest yearly pre-pandemic median, in 2016. From this perspective, Brazilian research and surveillance made an incredible effort and demonstrated an impressive capacity to adapt and respond while the country was going through not only a global public health emergency, but also a period of great economic, social and political instability.

On the other hand, despite the huge improvements in genome sequencing capacity when compared to the country's own historical numbers, COVID-19 sequencing efforts in Brazil can be considered less than optimal when put into global perspective. A recent pre-print has found that the number of sequenced genomes is highly correlated to the number of lineages identified in each location (91). This finding shows the impact that limited genome sequencing can have in the identification of new SARS-CoV-2 lineages, especially VOCs, which are lineages with higher transmissibility and/or immune escape. This study also estimated that a sequencing proportion of 0.5% of all COVID-19 cases with a turnaround time of 21 days would lead to a 20% probability of early identification of a newly circulating variant (before reaching 100 cases) and would be a good benchmark to guide sequencing efforts (91). Considering such estimates as the ideal minimum target for adequate COVID-19 genomic surveillance, Brazil would have only met the target sequencing proportion at the very beginning of the pandemic and from August 2021 onwards, when the number of COVID-19 cases were declining (< 1 million cases a month). As for the target turnaround time of 21 days, it was only met at the very beginning of the pandemic, in March 2020. Median turnaround times in the following months were at least 2 times higher than the target, but as high as 10 times higher in December 2020, for example.

At the state-level, only São Paulo, Rio de Janeiro, Amazonas and Alagoas have met the 0.5% sequencing proportion target when considering the total number of sequences and cases to date. No states have met the median turnaround time target of 21 days. Moreover, the regional disparities in sequencing capacity are also a major concern to be considered, as such target estimates assume weekly random sampling across the country, rather than selectively sequencing larger proportion of cases in higher income states, especially São Paulo, which accounts for over half of the sequences from Brazil. States such as Piauí and Mato Grosso, two of the states with the lowest sequencing rates and turnaround times, would have a probability of detecting a newly circulating lineage before reaching 100 cases of less than 2.6%, according to the same pre-print study. For instance, by 11th January 2021, when Gamma was first reported, there were only 7 genomes from Manaus available on GISAID (32), 6 of which had been published as part of Chapter 3 of this thesis. In fact, Gamma was first reported in returning travelers in Japan that had visited Manaus (33, 92), although Brazilian studies describing the lineage followed in the next couple of days (32). The state of Amazonas, to which Manaus is the capital, is now the 4th in total number of sequences and the third in sequencing rate. Most of this increase in sequencing capacity was likely encouraged by the discovery of Gamma and the targeted sequencing of samples from Amazonas to investigate Gamma's emergence, local spread and characteristics. A considerable share of these samples was probably not sequenced locally. However, other states in North Brazil have much lower sequencing rates, which limits our understanding of the evolution and spread of new variants in the region (93). The emergence of Gamma is also likely to have influenced the increase in sequencing and decrease in turnaround times observed nationwide in 2021 when compared to 2020, similar to what occurred globally after the discovery Alpha and Beta (91).

However, Brazil is far from being the only country in a suboptimal genomic surveillance situation (Figure xxx). Only 16 countries worldwide have been able to generate genomes of 5% or more of its COVID-19 cases (91). Some of the barriers faced by LMICs include: (i) genome sequencing might become an additional and a lower priority expense for countries which are already struggling with medical care and diagnostics (91); (ii) reagents and equipment for genome sequencing are usually not locally produced and importation can take a long time and be very expensive, as it has been shown for diagnostics (94, 95) – for example, importation process in Brazil can take more than 3 months between order placement and arrival of reagents; (iii) lack of local specialized human resources, especially when it comes to bioinformatics (96, 97); (iv) use of older sequencing technologies that allow for lower throughput (98); (v) overly centralized sequencing systems (98), (vi) the need to stablish new collaborations and arrangements to strengthen sequencing (99), (vi) the hesitancy of publishing data prior to it been published (100), (vii) and fear of the consequences of finding new VOIs and VOCs (101). In Brazil, this is further aggravated when considering sample transportation limitations. For example, in the Amazon region human mobility is highly fluvial in this area which presents challenges for the conservation of RNA in biological samples. However, even if some of these problems are primarily faced by LMIC countries, virus spread does not respect geopolitical borders, which means VOC emerging in poorly connected regions in the Amazon region may rapidly become a global issue. “To be as effective as possible, surveillance needs to be widespread, standardized and embedded in national pandemic-prevention programmes” (102).

Limitations regarding genome sequencing capacity and submission turnaround times have encouraged nations across the globe to create initiatives to ramp up their sequencing performances (98, 102). This includes the COVID-19 Genomics UK Consortium (COG-UK) in United Kingdom (103), the Indian SARS-CoV-2 Genomic Consortia (INSA-COG)

in India, the Network For Genomic Surveillance In South Africa (NGS-SA) in South Africa, the Canadian COVID Genomics Network (CanCOGeN) in Canada, and the SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance (SPHERES) in the United States (98). In Latin America, the Pan America Association of Health (PAHO) led the creation of a continent-wide network for local support of SARS-CoV-2 genome sequencing, COVID-19 Genomic Surveillance Regional Network (COVGEN). The network is composed by countries which can perform genome sequencing locally and two regional centers which receive samples from countries without local sequencing capacity, Fundação Oswaldo Cruz/FIOCRUZ - Brazil and Instituto de Salud Publica/ISPCH-Chile (40).

In Brazil, several important nationwide initiatives were created to increase genome sequencing capacity in the country. The Rede Genômica FioCruz was created by the Oswaldo Cruz Foundation (FioCruz) and includes 12 local branches of FioCruz across Brazil and the Instituto Adolf Lutz in São Paulo. It performs genomic surveillance, capacitation and technical support for the entire country and also to other nations in Latin America. It also provides technical reports and updates an online dashboard on the spread of SARS-CoV-2 variants across Brazil (38). The Rede Corona-Ômica BR MCTI from the Brazilian Ministry of Science, Technology and Innovation is a national network for genomics, transcriptomics and exomics of SARS-CoV-2 and it includes specialists from several research institutions across Brazil (104). The Rede Nacional de Sequenciamento Genético from the Brazilian Ministry of Health was created in February 2021 and includes all Central Laboratories for public health (LACENS) of each state in Brazil (105). Some of these initiatives were timed with the emergence of Gamma in Brazil and certainly contributed the massive increase in genome sequencing capacity in Brazil from early 2021 onwards. Following WHO guidelines for SARS-CoV-2 sequencing in early January 2021

(106), Brazil released a national guideline for SARS-CoV-2 sequencing to strengthen SARS-CoV-2 genome sequencing in the country (107).

Although not specifically created in response to the COVID-19 pandemic, the Brazil-UK Centre for (Arbo)virus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) has been one of the main research and surveillance initiatives on COVID-19 genomics and epidemiology in Brazil. CADDE was founded in 2019 following the success of the Zika in Brazil Real Time Analysis (ZiBRA), which travelled between LACENS in Northeast Brazil in a mobile laboratory in 2016 and generated ZIKV genomes using portable genome sequencing for the first time in Brazil (82). Since then, ZiBRA has also been responsible for capacity building in genome sequencing and phylogenetics across Brazil and for generating genomic data on some of the major virus outbreaks in Brazil, including the large 2016–2019 yellow fever outbreak in Southeast Brazil (108). Although initially focused on arbovirus research, CADDE was responsible for the generation of the first Latin American SARS-CoV-2 genomes and phylogenetic analysis in only 48 hours and for the earliest identification of the Gamma variant in Manaus, as presented on Chapters 2, 3 and 4 of this thesis.

6.2.2 Future directions for genomic sequencing and epidemiology

The improvement of sequencing capacity and the reduction of turnaround times will be essential for the future of genome surveillance programmes across Brazil. However, fast interpretation and contextualisation of well characterized genomic data, as well as data sharing and accurate and concise communication of research findings, are also critical to inform public health responses (109, 110). Such challenges could be ameliorated by strengthening local capacity, employing representative sampling strategies, and by collating and sharing appropriate metadata associated with clinical or environmental samples (111).

One of key limitations for genomic epidemiological studies is the availability and the quality of the metadata associated with sequenced samples. According to the WHO Guidance for surveillance of SARS-CoV-2 variants, there are three tiers of metadata for genomic surveillance based on their collection priority (111). Highest priority metadata includes the basic information needed for tracking virus movement, time and place of collection, and information on laboratories collecting and processing samples and sequencing. The second priority metadata includes patient and epidemiological characteristics such as age, gender, race and ethnicity, date of exposure and onset. Finally, the third tier includes information relevant to characterise the potential public health impact of new variants such as cycle threshold (Ct) values, travel history, vaccination status, clinical severity, hospitalisation status, outcome and etc (111).

Databases for metadata associated with clinical samples and patients are usually not interconnected in Brazil. In addition, testing facilities usually do not often collect relevant metadata such as recent travel history, race and ethnicity, comorbidities or vaccination status. Centralizing complete metadata would optimize human resources and speed up sequencing analysis and data sharing. Issues surrounding metadata collection, sharing and quality are a longstanding global problem, as highlighted by the Genomics Standards Consortium (112). Guidelines on which metadata is necessary and important to be shared improved include the use of MIxS (Minimum Information about any (x) Sequence) checklists, MIxS records and the FAIR sharing principles. However, a pre-print analysis showed that out of 75,000 GISAID entries, 68.81% had unknown gender information, 69.12% had unknow age information, with >10,000 and >1,000 missing entries regarding specimen source and country fields, respectively (113). Misspelling errors were also found to be fairly common in fields such as originating and submitting lab, 9.8% and 11.6%, respectively (114). These findings highlight an pressing need for understanding why

collection and/or sharing of genomic metadata have been so often neglected during the COVID19 response. Implementation of mechanisms that can increase metadata collection, standardisation, sharing and quality assurance will be essential to improve our understanding of virus spread and the emergence of new variants and their potential to cause severe disease globally (112).

Sampling strategy is also an important aspect of genome surveillance, as different research questions will be better answered by specific sampling strategies to avoid the introduction of sampling biases. For example, unbalanced sampling strategies can result into locations being over or underrepresented in internal nodes of phylogeographic analysis, and lead to biased conclusions on virus geographical origins and movement, overly-sized transmission clades and incorrect diffusion rates (23, 115). In such cases, downsampling strategies may be used to reduce sampling bias, although such approaches may sometimes be less efficient and costlier than employing adequate sampling at first (23).

International guidelines for genomic surveillance recognise two general categories of sampling: representative sampling from surveillance systems and targeted sampling in specific settings or populations (111). Representative sampling should include criteria such as geographical and temporal distribution, age, sex and clinical spectrum. As per sample size, strategies may include sequencing of a fixed proportion of cases or of a fixed number of cases. The former will result in increased sensitivity for the detection of new variants but can become demanding and impractical in periods of high virus transmission and incidence (111). The latter is more feasible as sequencing capacity does not need to be changed according to current incidence, however, it presents lower sensitivity. Targeted sampling strategies can be used to identify new variants or understanding virus diversity and spread in specific scenarios and populations (111). Common examples of targeted-sampling strategies are based on specimen characteristics (e.g. Ct values or failure of detection in

specific assays), individual characteristics (disease outcome, breakthrough infections, etc), environment characteristics (wastewater studies) or outbreak characteristics (outbreak surveillance) (111). In fact, GISAID has recently include a new nonmandatory "sampling strategy" field, which could be of great importance in reducing sampling bias in future SARS-CoV-2 analysis, including correct assessments of VOC frequency (116).

In Brazil, as probably in most countries, genome sequencing started using a convenience sampling strategy given the limited sample access and sequencing capacity. Once sequencing capacity improved and genomic surveillance increased, sampling strategies likely shifted towards a more representative sampling strategy nationally, varying between fixed proportion and fixed number approaches as it can be seen in Figure 6.2. Targeted sampling has also been used to understanding vaccine breakthrough infections, spread in specific populations and characterisation of new variants. In Chapter 3, given the aim to understand virus introduction and spread at a countrywide level, I used a representative sampling using a number of genomes proportional to the number of SRAG cases per state to characterise geographical and temporal spread across the country. Such approach was likely very effective in its attempt of understanding countrywide introduction and spread, however of limited effectiveness for understanding local spread in some Brazilian regions given the small number of samples. For Chapters 4.1, 4.2 and 5, targeted sampling was used for the identification and characterisation of new variants and to understanding virus spread within and between HCW and hospital patients. In Chapter 4.1, two samples positive for SARS-CoV-2 infection but presenting S-gene negative PCR results were sequenced under the suspicion of Alpha variant infection. For Chapter 4.2, samples from Manaus with collection date between November 2020 and January 2021 were sequenced to understand the potential causes of the case upsurge in Manaus and to characterise the newly emerging Gamma variant. In Chapter 5, a specific proportion of

SARS-CoV-2 positive samples from three specific medical institutes from the same hospital complex were sequenced to provide insights on within- and between-hospital SARS-CoV-2 transmission.

Globally, the fast increase in the availability of SARS-CoV-2 genome sequences and the development of novel methods and technologies have increased the depth in which genome sequences can be used to inform and evaluate public health responses during epidemics. For instance, travel histories, human mobility and epidemiological data can be integrated to phylodynamic frameworks to overcome limitations regarding unsampled locations, sampling biases and low inference accuracy (5, 117-119). Moreover, genomic data can also been used to estimate SARS-CoV-2 population-level epidemiological parameters such as R_0, R_t and the overdispersion parameter k, and to assess the impact of interventions in virus spread (120-125). Covariates can also be input into phylogenetic generalised linear models (GLMs) to identify potential factors associated with virus spread (126-128). In Brazilian SARS-CoV-2 studies, some of these approaches have rarely if ever been used (129, 130), and most Brazilian publications rely on the description of newly generated sequences, evolutionary description of new variants, genomic diversity, or phylogeographic reconstruction of virus spread (25, 26, 131-138). Although genomic surveillance in Brazil remains uncoordinated and highly unbalanced at the national level, the state of São Paulo has been able to generate a large number of viral sequences during the ongoing COVID-19 pandemic. Such datasets could be used to understand factors affecting virus spread, the impact of NPIs and vaccination in the largest interconnected urban area of the Southern hemisphere. In addition, it will also be important to integrate population structure data to genomic analysis, including time-changing vaccination rates and spatially heterogenous contact networks.

6.3. Concluding remarks

This thesis provided invaluable information on the early SARS-CoV-2 importation routes and transmission, and importantly, informed on the early identification and characterization of newly circulating VOCs in Brazil. The work presented here highlights the public health impact of real-time and retrospective genomic surveillance across distinct temporal and geographical scales during a global epidemic threat in a LMIC. This thesis also provided evidence for limited effectivity of short-lived NPIs implemented across Brazil. On a smaller geographical scale, this thesis also showed that SARS-CoV-2 hospital transmission in the largest hospital complex in Brazil was higher in institutes believed to be SARS-CoV-2-free, such as non-COVID-19 hospitals and wards. This is of special importance in the context of the emergence on novel immunity-evading variants, such as Omicron. Overall, this thesis represents an overview of the main aspects of SARS-CoV-2 importation, spread, evolution and public health response in Brazil.

The findings presented here also highlight the challenges and limitations faced by the genomic surveillance community in Brazil and other LMIC, and their impact in sequencing rates and turnaround times. Although Brazil has made an impressive effort and achieved great genomic surveillance capacity in the second year of the pandemic, there is still an urgent need for increasing surveillance capacity, especially in the most deprived regions of the country, and to standardise metadata collection. Lack of reagents, technological-dependency on developed countries and limited local human resources in LMIC are only some of the challenges that need to be globally addressed for genomic surveillance to be truly universal, COVID-19 variant tracking and public health response to be timely and effective, and for humanity to be prepared for the next pandemic Disease X.

6.4. References

1. Kraemer MUG, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science. 2020;368(6490):493-7.
2. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020;368(6489):395-400.
3. Manabe YC, Sharfstein JS, Armstrong K. The Need for More and Better Testing for COVID-19. JAMA. 2020;324(21):2153-4.
4. de Souza WM, Buss LF, Candido DDS, Carrera JP, Li S, Zarebski AE, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat Hum Behav. 2020;4(8):856-65.
5. du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021;371(6530):708-12.
6. McCrone JT, Hill V, Bajaj S, Pena RE, Lambert BC, Inward R, et al. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. medRxiv. 2021:2021.12.14.21267606.
7. Kraemer MUG, Hill V, Ruis C, Dellicour S, Bajaj S, McCrone JT, et al. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. Science. 2021;373(6557):889-95.
8. D'Arienzo M, Coniglio A. Assessment of the SARS-CoV-2 basic reproduction number,. Biosaf Health. 2020;2(2):57-9.
9. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med. 2020;382(13):1199-207.
10. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet. 2020;395(10225):689-97.
11. Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Euro Surveill. 2020;25(4).
12. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis. 2020;20(5):553-8.
13. Mellan TA, Hoeltgebaum HH, Mishra S, Whittaker C, Schnekenberg RP, Gandy A, et al. Report 21: Estimating COVID-19 cases and reproduction number in Brazil. medRxiv. 2020.
14. de Souza Santos AA, Candido DDS, de Souza WM, Buss L, Li SL, Pereira RHM, et al. Dataset on SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities. Sci Data. 2021;8(1):73.
15. Fernandez M, Machado C. COVID-19's political challenges in Latin America. Springer; 2021.
16. Giuliani D, Dickson MM, Espa G, Santi F. Modelling and predicting the spatio-temporal spread of cOVID-19 in Italy. BMC Infect Dis. 2020;20(1):700.
17. Paiva MHS, Guedes DRD, Docena C, Bezerra MF, Dezordi FZ, Machado LC, et al. Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil. Viruses. 2020;12(12).

18. Naveca FG, Nascimento V, de Souza VC, Corado AL, Nascimento F, Silva G, et al. COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. Nat Med. 2021;27(7):1230-8.
19. Franceschi VB, Caldana GD, de Menezes Mayer A, Cybis GB, Neves CAM, Ferrareze PAG, et al. Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. BMC Genomics. 2021;22(1):371.
20. de Medeiros Oliveira M, Schemberger MO, Suzukawa AA, Riediger IN, Debur MdC, Becker G, et al. Genomic surveillance of SARS-CoV-2 in the state of Paraná, Southern Brazil, reveals the cocirculation of the VOC P.1, P.1-like-II lineage and a P.1 cluster harboring the S:E661D mutation. medRxiv. 2021:2021.07.14.21260508.
21. Xavier J, Giovanetti M, Adelino T, Fonseca V, Barbosa da Costa AV, Ribeiro AA, 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. 2020;9(1):1824-34.
22. Giovanetti M, Slavov SN, Fonseca V, Wilkinson E, Tegally H, Patané JSL, et al. Genomic epidemiology reveals how restriction measures shaped the SARS-CoV-2 epidemic in Brazil. medRxiv. 2021:2021.10.07.21264644.
23. Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. Progress and challenges in virus genomic epidemiology. Trends Parasitol. 2021;37(12):1038-49.
24. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22(13).
25. Moreira FRR, D'arc M, Mariani D, Herlinger AL, Schiffler FB, Rossi Á, et al. Epidemiological dynamics of SARS-CoV-2 VOC Gamma in Rio de Janeiro, Brazil. Virus Evol. 2021;7(2):veab087.
26. Moreira FRR, Bonfim DM, Zauli DAG, Silva JP, Lima AB, Malta FSV, et al. Epidemic Spread of SARS-CoV-2 Lineage B.1.1.7 in Brazil. Viruses. 2021;13(6).
27. Slavov SN, Bezerra RDS, Rodrigues ES, Santos EV, Borges JS, de la Roque DGL, et al. Genomic monitoring of the SARS-CoV-2 B1.1.7 (WHO VOC Alpha) in the Sao Paulo state, Brazil. Virus Res. 2022;308:198643.
28. Hodcroft EB. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. 2021 [Available from: https://covariants.org/.
29. Sabino EC, Buss LF, Carvalho MPS, Prete CA, Jr., Crispim MAE, Fraiji NA, et al. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. Lancet. 2021;397(10273):452-5.
30. Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations: Virological; 2020 [Available from: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
31. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021;592(7854):438-43.
32. Faria N, Claro I, Candido D, Moyses Franco L, Andrade P, Coletti T, et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. 2021. 2021.
33. NIID-Japan. Brief report: New Variant Strainof SARS-CoV-2Identified in Travelers from Brazil: NIID-Japan; 2021 [Available from: https://www.niid.go.jp/niid/images/epi/corona/covid19-33-en-210112.pdf.

34. Buss LF, Prete CA, Abrahim CMM, Mendrone A, Salomon T, de Almeida-Neto C, et al. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. Science. 2021;371(6526):288-92.
35. Naveca F, da Costa C, Nascimento V, Souza V, Corado A, Nascimento F, et al. Three SARS-CoV-2 reinfection cases by the new Variant of Concern (VOC) P. 1/501Y. V3. 2021.
36. Naveca F, da Costa C, Nascimento V, Souza V, Corado A, Nascimento F, et al. SARS-CoV-2 reinfection by the new Variant of Concern (VOC) P. 1 in Amazonas, Brazil. virological org. 2021.
37. Romano CM, Felix AC, Paula AV, Jesus JG, Andrade PS, Cândido D, et al. SARS-CoV-2 reinfection caused by the P.1 lineage in Araraquara city, Sao Paulo State, Brazil. Rev Inst Med Trop Sao Paulo. 2021;63:e36.
38. Fiocruz. Dashboard Rede Genômica Fiocruz: Fiocruz; 2021 [Available from: http://www.genomahcov.fiocruz.br/grafico/.
39. Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, et al. Coronavirus Pandemic (COVID-19) *OurWorldInData.org*2020 [Available from: https://ourworldindata.org/coronavirus.
40. PAHO. COVID-19 Genomic Surveillance Regional Network Pan American Health Organization2021 [Available from: https://www.paho.org/en/topics/influenza-and-other-respiratory-viruses/covid-19-genomic-surveillance-regional-network.
41. O'Toole Á, Hill V, Pybus OG, Watts A, Bogoch II, Khan K, et al. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. Wellcome Open Res. 2021;6:121.
42. Knaul FM, Touchton M, Arreola-Ornelas H, Atun R, Anyosa RJC, Frenk J, et al. Punt Politics as Failure of Health System Stewardship: Evidence from the COVID-19 Pandemic Response in Brazil and Mexico. Lancet Reg Health Am. 2021;4:100086.
43. Alves L. Health experts welcome Brazil COVID-19 inquiry findings. Lancet. 2021;398(10312):1674-5.
44. Furlan L, Caramelli B. The regrettable story of the "Covid Kit" and the "Early Treatment of Covid-19" in Brazil. Lancet Reg Health Am. 2021;4:100089.
45. Daniels JP. Health experts slam Bolsonaro's vaccine comments. Lancet. 2021;397(10272):361.
46. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20(5):533-4.
47. Arslanca T, Fidan C, Daggez M, Dursun P. Knowledge, preventive behaviors and risk perception of the COVID-19 pandemic: A cross-sectional study in Turkish health care workers. PLoS One. 2021;16(4):e0250017.
48. Paudel S, Palaian S, Shankar PR, Subedi N. Risk Perception and Hesitancy Toward COVID-19 Vaccination Among Healthcare Workers and Staff at a Medical College in Nepal. Risk Manag Healthc Policy. 2021;14:2253-61.
49. Li M, Luo Y, Watson R, Zheng Y, Ren J, Tang J, et al. Healthcare workers' (HCWs) attitudes and related factors towards COVID-19 vaccination: a rapid systematic review. Postgrad Med J. 2021.
50. Toth-Manikowski SM, Swirsky ES, Gandhi R, Piscitello G. COVID-19 vaccination hesitancy among health care workers, communication, and policy-making. Am J Infect Control. 2022;50(1):20-5.
51. Paris C, Bénézit F, Geslin M, Polard E, Baldeyrou M, Turmel V, et al. COVID-19 vaccine hesitancy among healthcare workers. Infect Dis Now. 2021;51(5):484-7.

52. Aemro A, Amare NS, Shetie B, Chekol B, Wassie M. Determinants of COVID-19 vaccine hesitancy among health care workers in Amhara region referral hospitals, Northwest Ethiopia: a cross-sectional study. Epidemiol Infect. 2021;149:e225.
53. Ledda C, Costantino C, Cuccia M, Maltezou HC, Rapisarda V. Attitudes of Healthcare Personnel towards Vaccinations before and during the COVID-19 Pandemic. Int J Environ Res Public Health. 2021;18(5).
54. Papagiannis D, Rachiotis G, Malli F, Papathanasiou IV, Kotsiou O, Fradelos EC, et al. Acceptability of COVID-19 Vaccination among Greek Health Professionals. Vaccines (Basel). 2021;9(3).
55. Holzmann-Littig C, Braunisch MC, Kranke P, Popp M, Seeber C, Fichtner F, et al. COVID-19 Vaccination Acceptance and Hesitancy among Healthcare Workers in Germany. Vaccines (Basel). 2021;9(7).
56. Abuown A, Ellis T, Miller J, Davidson R, Kachwala Q, Medeiros M, et al. COVID-19 vaccination intent among London healthcare workers. Occup Med (Lond). 2021;71(4-5):211-4.
57. Verger P, Scronias D, Dauby N, Adedzi KA, Gobert C, Bergeat M, et al. Attitudes of healthcare workers towards COVID-19 vaccination: a survey in France and French-speaking parts of Belgium and Canada, 2020. Eurosurveillance. 2021;26(3):2002047.
58. Thorneloe R, Wilcockson HE, Lamb M, Jordan C, Arden MA, editors. Willingness to receive a COVID-19 vaccine among adults at high-risk of COVID-19: a UK-wide survey2020.
59. Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Lessells RJ, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. medRxiv. 2021.
60. Buchan SA, Chung H, Brown KA, Austin PC, Fell DB, Gubbay JB, et al. Effectiveness of COVID-19 vaccines against Omicron or Delta infection. medRxiv. 2022:2021.12.30.21268565.
61. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Effectiveness of COVID-19 vaccines against the Omicron (B.1.1.529) variant of concern. medRxiv. 2021:2021.12.14.21267615.
62. Nebehay S. Global shortage of nurses set to grow as pandemic enters third year - group Reuters.com: Reuters; 2021 [Available from: https://www.reuters.com/business/healthcare-pharmaceuticals/global-shortage-nurses-set-grow-pandemic-enters-third-year-group-2021-12-10/.
63. Kim L. These 18 States Are Grappling With Critical Hospital Worker Shortages As Covid Hospitalizations Surge Forbes.com: Forbes; 2021 [Available from: https://www.forbes.com/sites/lisakim/2022/01/08/these-18-states-are-grappling-with-critical-hospital-worker-shortages-as-covid-hospitalizations-surge/?sh=301b27012e23.
64. What is the cure for the global healthcare worker shortage? ft.com: Financial times; 2021 [Available from: https://www.ft.com/partnercontent/j-and-j/what-is-the-cure-for-the-global-healthcare-worker-shortage.html.
65. CDC. Interim Guidance for Managing Healthcare Personnel with SARS-CoV-2 Infection or Exposure to SARS-CoV-2 cdc.gov2021 [Available from: https://www.cdc.gov/coronavirus/2019-ncov/hcp/guidance-risk-assesment-hcp.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fhcp%2Freturn-to-work.html.
66. Illingworth CJ, Hamilton WL, Jackson CH, Popay A, Meredith L, Houldcroft CJ, et al. A2B-COVID: a method for evaluating potential SARS-CoV-2 transmission events. medRxiv. 2021:2020.10. 26.20219642.

67. Illingworth CJ, Hamilton WL, Warne B, Routledge M, Popay A, Jackson C, et al. Superspreaders drive the largest outbreaks of hospital onset COVID-19 infections. Elife. 2021;10.
68. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis. 2020;20(11):1263-72.
69. Berkman A, Garcia J, Muñoz-Laboy M, Paiva V, Parker R. A critical analysis of the Brazilian response to HIV/AIDS: lessons learned for controlling and mitigating the epidemic in developing countries. Am J Public Health. 2005;95(7):1162-72.
70. Valente DS, Zanella RK. Brazil's COVID-19 response. Lancet (London, England). 2020;396(10254):e32-e.
71. Castro MC, Kim S, Barberia L, Ribeiro AF, Gurzenda S, Ribeiro KB, et al. Spatiotemporal pattern of COVID-19 spread in Brazil. Science. 2021;372(6544):821-6.
72. Kowaltowski AJ. Brazil's scientists face 90% budget cut. Nature. 2021;598(7882):566.
73. Hallal PC. SOS Brazil: science under attack. Lancet. 2021;397(10272):373-4.
74. Galvão-Castro B, Cordeiro RSB, Goldenberg S. Brazilian science under continuous attack. Lancet. 2022;399(10319):23-4.
75. Benítez MA, Velasco C, Sequeira AR, Henríquez J, Menezes FM, Paolucci F. Responses to COVID-19 in five Latin American countries. Health Policy Technol. 2020;9(4):525-59.
76. Barberia LG, Costa SF, Sabino EC. Brazil needs a coordinated and cooperative approach to tackle COVID-19. Nat Med. 2021;27(7):1133-4.
77. Idrovo AJ, Manrique-Hernández EF, Fernández Niño JA. Report From Bolsonaro's Brazil: The Consequences of Ignoring Science. Int J Health Serv. 2021;51(1):31-6.
78. Taylor L. 'We are being ignored': Brazil's researchers blame anti-science government for devastating COVID surge. Nature. 2021;593(7857):15-6.
79. Dyer O. Covid-19: Bolsonaro under fire as Brazil hides figures. BMJ. 2020;369:m2296.
80. Ventura D, Reis R. An unprecedented attack on human rights in Brazil: the timeline of the federal government's strategy to spread Covid-19. Offprint. Bulletin Rights in the Pandemic n 10. São Paulo, Brazil: CEPEDISA/USP and Conectas Human Rights; 2021.
81. Brum E. Study finds that Brazil's Jair Bolsonaro carried out an 'institutional strategy to spread the coronavirus': El Pais; 2021 [Available from: https://english.elpais.com/americas/2021-01-29/study-finds-that-brazils-jair-bolsonaro-carried-out-an-institutional-strategy-to-spread-the-coronavirus.html.
82. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546(7658):406-10.
83. Li SL, Pereira RHM, Prete CA, Zarebski AE, Emanuel L, Alves PJH, et al. Higher risk of death from COVID-19 in low-income and non-White populations of São Paulo, Brazil. BMJ Glob Health. 2021;6(4).
84. Nicolelis MAL, Raimundo RLG, Peixoto PS, Andreazzi CS. The impact of super-spreader cities, highways, and intensive care availability in the early stages of the COVID-19 epidemic in Brazil. Sci Rep. 2021;11(1):13001.
85. Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, et al. A global database of COVID-19 vaccinations. Nat Hum Behav. 2021;5(7):947-53.

86. GHS. 2019 Global Health Security index. Nuclear Threat Initiative (NTI) and the Johns Hopkins Center for Health Security; 2019.
87. Lowy institute Covid Performance Index 2021 [Available from: https://interactives.lowyinstitute.org/ features/covid-performance.
88. Haider N, Yavlinsky A, Chang YM, Hasan MN, Benfield C, Osman AY, et al. The Global Health Security index and Joint External Evaluation score for health preparedness are not correlated with countries' COVID-19 detection response time and mortality outcome. Epidemiol Infect. 2020;148:e210.
89. Khalifa BA, Abbey EJ, Ayeh SK, Yusuf HE, D Nudotor R, Osuji N, et al. The Global Health Security Index is not predictive of vaccine rollout responses among OECD countries. Int J Infect Dis. 2021;113:7-11.
90. Abbey EJ, Khalifa BAA, Oduwole MO, Ayeh SK, Nudotor RD, Salia EL, et al. The Global Health Security Index is not predictive of coronavirus pandemic responses among Organization for Economic Cooperation and Development countries. PLoS One. 2020;15(10):e0239398.
91. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. medRxiv. 2021:2021.08.21.21262393.
92. Reuters. Japan finds new coronavirus variant in travellers from Brazil reuters.com: Reuters; 2021 [Available from: https://www.reuters.com/article/us-health-coronavirus-japan-variant-idUSKBN29F08R.
93. Wadman M. Blind spots thwart global coronavirus tracking. Science. 2021;372(6544):773-4.
94. ECLAC. Restrictions on the export of medical products hamper efforts to contain coronavirus disease (COVID-19) in Latin America and the Caribbean: ECLAC; 2020 [
95. Rubin R, Abbasi J, Voelker R. Latin America and Its Global Partners Toil to Procure Medical Supplies as COVID-19 Pushes the Region to Its Limit. JAMA. 2020;324(3):217-9.
96. Maxmen A. One million coronavirus sequences: popular genome site hits mega milestone. Nature. 2021;593(7857):21.
97. Grubaugh ND, Hodcroft EB, Fauver JR, Phelan AL, Cevik M. Public health actions to control new SARS-CoV-2 variants. Cell. 2021;184(5):1127-32.
98. Kalia K, Saberwal G, Sharma G. The lag in SARS-CoV-2 genome submissions to GISAID. Nat Biotechnol. 2021;39(9):1058-60.
99. Agrawal A. India's COVID crisis flags need to forecast variants. Nature. 2021;594(7861):9.
100. Van Noorden R. Scientists call for fully open sharing of coronavirus genome data. Nature. 2021;590(7845):195-6.
101. Mallapaty S. Omicron-variant border bans ignore the evidence, say scientists. Nature. 2021;600(7888):199.
102. Cyranoski D. Alarming COVID variants show vital role of genomic surveillance. Nature. 2021;589(7842):337-8.
103. Nicholls SM, Poplawski R, Bull MJ, Underwood A, Chapman M, Abu-Dahab K, et al. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. Genome Biol. 2021;22(1):196.
104. Rede Nacional de ômicas de COVID-19 para identificação de fatores associados à dispersão da epidemia e severidade. 2020 [Available from: **http://www.corona-omica.br-mctic.lncc.br/**.
105. Gandra A. Rede investigará mutações do novo coronavírus em circulação no país: Agência Brasil; 2021 [Available from:

https://agenciabrasil.ebc.com.br/saude/noticia/2021-02/rede-nacional-investigara-mutacoes-do-sars-cov-2-em-circulacao-no-pais.
106. WHO. Genomic Sequencing of SARS-CoV-2: a Guide to Implementation for Maximum Impact on Public Health. Geneva: WHO; 2021.
107. Brasil. Vigilância Genômica do vírus SARS-CoV-2no âmbito da SVS/MS Brasilia: Ministério da Saúde. Secretaria; 2021 [Available from: https://bvsms.saude.gov.br/bvs/publicacoes/vigilancia_genomica_SARS-CoV-2_ambito_SVS.pdf.
108. Faria NR, Kraemer MUG, Hill SC, Goes de Jesus J, Aguiar RS, Iani FCM, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science. 2018;361(6405):894-9.
109. The L. Genomic sequencing in pandemics. Lancet. 2021;397(10273):445.
110. Romano CM, Melo FL. Genomic surveillance of SARS-CoV-2: A race against time. Lancet Reg Health Am. 2021;1:100029.
111. WHO. Guidance for surveillance of SARS-CoV-2 variants: Interim guidance, 9 August 2021: WHO; 2021 [Available from: https://apps.who.int/iris/rest/bitstreams/1361901/retrieve.
112. Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci Data. 2020;7(1):188.
113. Velazquez A, Bustria M, Ouyang Y, Moshiri N. An analysis of clinical and geographical metadata of over 75,000 records in the GISAID COVID-19 database. medRxiv. 2020:2020.09.22.20199497.
114. Gozashti L, Corbett-Detig R. Shortcomings of SARS-CoV-2 genomic metadata. BMC Res Notes. 2021;14(1):189.
115. Kalkauskas A, Perron U, Sun Y, Goldman N, Baele G, Guindon S, et al. Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. PLoS Comput Biol. 2021;17(1):e1008561.
116. Scott L, Hsiao NY, Moyo S, Singh L, Tegally H, Dor G, et al. Track Omicron's spread with molecular data. Science. 2021;374(6574):1454-5.
117. Butera Y, Mukantwari E, Artesi M, D'Arc Umuringa J, O'Toole ÁN, Hill V, et al. Genomic Sequencing of SARS-CoV-2 in Rwanda: evolution and regional dynamics. medRxiv. 2021:2021.04.02.21254839.
118. Lemey P, Hong SL, Hill V, Baele G, Poletto C, Colizza V, et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nat Commun. 2020;11(1):5110.
119. Lemey P, Ruktanonchai N, Hong SL, Colizza V, Poletto C, Van den Broeck F, et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. Nature. 2021;595(7869):713-7.
120. Miller D, Martin MA, Harel N, Tirosh O, Kustin T, Meir M, et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nat Commun. 2020;11(1):5518.
121. Smith MR, Trofimova M, Weber A, Duport Y, Kühnert D, von Kleist M. Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020. Nat Commun. 2021;12(1):6009.
122. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 2021;593(7858):266-9.

123. Douglas J, Mendes FK, Bouckaert R, Xie D, Jiménez-Silva CL, Swanepoel C, et al. Phylodynamics reveals the role of human travel and contact tracing in controlling the first wave of COVID-19 in four island nations. Virus Evol. 2021;7(2):veab052.
124. Müller NF, Wagner C, Frazar CD, Roychoudhury P, Lee J, Moncla LH, et al. Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. Sci Transl Med. 2021;13(595).
125. Ragonnet-Cronin M, Boyd O, Geidelberg L, Jorgensen D, Nascimento FF, Siveroni I, et al. Genetic evidence for the association between COVID-19 epidemic severity and timing of non-pharmaceutical interventions. Nat Commun. 2021;12(1):2188.
126. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544(7650):309-15.
127. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014;10(2):e1003932.
128. Nunes MR, Palacios G, Faria NR, Sousa EC, Pantoja JA, Rodrigues SG, et al. Air travel is associated with intracontinental spread of dengue virus serotypes 1-3 in Brazil. PLoS Negl Trop Dis. 2014;8(4):e2769.
129. Naveca F, Nascimento V, Souza V, Corado A, Nascimento F, Silva G, et al. Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein. Virological org. 2021;1:1-8.
130. Resende PC, Delatorre E, Gräf T, Mir D, Motta FC, Appolinario LR, et al. Evolutionary Dynamics and Dissemination Pattern of the SARS-CoV-2 Lineage B.1.1.33 During the Early Pandemic Phase in Brazil. Front Microbiol. 2020;11:615280.
131. Botelho-Souza LF, Nogueira-Lima FS, Roca TP, Naveca FG, de Oliveria Dos Santos A, Maia ACS, et al. SARS-CoV-2 genomic surveillance in Rondônia, Brazilian Western Amazon. Sci Rep. 2021;11(1):3770.
132. Voloch CM, da Silva Francisco R, de Almeida LGP, Cardoso CC, Brustolini OJ, Gerber AL, et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. J Virol. 2021.
133. Nascimento VAD, Corado ALG, Nascimento FOD, Costa Á, Duarte DCG, Luz SLB, et al. Genomic and phylogenetic characterisation of an imported case of SARS-CoV-2 in Amazonas State, Brazil. Mem Inst Oswaldo Cruz. 2020;115:e200310.
134. Campos KR, Sacchi CT, Abbud A, Caterino-de-Araujo A. SARS-CoV-2 variants in severely symptomatic and deceased persons who had been vaccinated against COVID-19 in São Paulo, Brazil. Rev Panam Salud Publica. 2021;45:e126.
135. Mir D, Rego N, Resende PC, Tort F, Fernández-Calero T, Noya V, et al. Recurrent Dissemination of SARS-CoV-2 Through the Uruguayan-Brazilian Border. Front Microbiol. 2021;12:653986.
136. Tosta S, Giovanetti M, Brandão Nardy V, Reboredo de Oliveira da Silva L, Kelly Astete Gómez M, Gomes Lima J, et al. Short Report: Early genomic detection of SARS-CoV-2 P.1 variant in Northeast Brazil. PLoS Negl Trop Dis. 2021;15(7):e0009591.
137. Oliveira MM, Schemberger MO, Suzukawa AA, Riediger IN, do Carmo Debur M, Becker G, et al. Re-emergence of Gamma-like-II and emergence of Gamma-S:E661D SARS-CoV-2 lineages in the south of Brazil after the 2021 outbreak. Virol J. 2021;18(1):222.
138. de Souza UJB, Dos Santos RN, Campos FS, Lourenço KL, da Fonseca FG, Spilki FR, et al. High Rate of Mutational Events in SARS-CoV-2 Genomes across Brazilian Geographical Regions, February 2020 to June 2021. Viruses. 2021;13(9).

Supplementary Information

Routes for COVID-19 importation in Brazil

Running title: COVID-19 importation in Brazil

Darlan Da S Candido, MSc[1], Alexander Watts, PhD [2,3], Leandro Abade, DPhil [1], Moritz UG Kraemer, DPhil [1,4,5], Prof Oliver G Pybus, DPhil [1,6], Prof Julio Croda, MD, PhD [7,8,9], Wanderson de Oliveira, PhD [7], Kamran Khan, MD, MPH [2,3], Prof Ester C Sabino, PhD [10], Prof Nuno R Faria, PhD[1,10]

Correspondence to Nuno Rodrigues Faria (nuno.faria@zoo.ox.ac.uk)

Historical air travel data

Historical air travel data was obtained for 29 of the countries with the highest number of SARS-Cov-2 cases as of 5th March 2010 (see list below). Data from the International Air Transport Association (IATA) was used to estimate the number of international travellers departing from each of the 29 countries and having Brazil as a final destination. IATA data corresponds to 90% of all worldwide trips on commercial flights from February to March 2019, and market intelligence was used to model the remaining data for the same period[1]. No information on possibly interrupted journeys was available.

SARS-CoV-2 Incidence

SARS-CoV-2 incidence for each of the 29 of the countries was calculated using the confirmed SARS-CoV-2 number of cases reported by World Health Organization as of 9th March 2020[2] and the total population for each country for 2019 from the United Nations

World Population Prospects 2019 database {UN, 2019 #4876}. A total of 29 countries used in this study: Algeria, Australia, Canada, China, Croatia, Denmark, Ecuador, Finland, France, Germany, Greece, Indonesia, Israel, Italy, Japan, Lebanon, Malaysia, Netherlands, Norway, Singapore, South Korea, Spain, Sweden, Switzerland, Thailand, United Arab Emirates, United Kingdom, United States of America, Viet Nam.

SARS-CoV-2 importation estimates

Estimates on the proportion of expected importations (E) for each air travel route were calculated using the incidence for each route (i) and the number of passengers (p) (historical air travel data) following $E = i * p / \sum i$ *100. The expected proportion of importations per country of origin (e.g. Italy) was calculated as the sum of (E) for all routes for starting at that specific country. Finally, the expected proportion of importations per Brazilian destination (e.g. Sao Paulo) was estimated as the sum of (E) for all routes ending at that specific destination, regardless of the country of origin.

Correlation between estimation and cases

To assess the accuracy of our estimates, we ran a correlation analysis between the estimated number of imported cases per final destination in Brazil (e.g. Sao Paulo) and the actual number of imported cases as reported by the Brazilian Ministry of Health on the 9th of March 2020. We fitted a simple linear regression model using RStudio Version 1.2.1335.

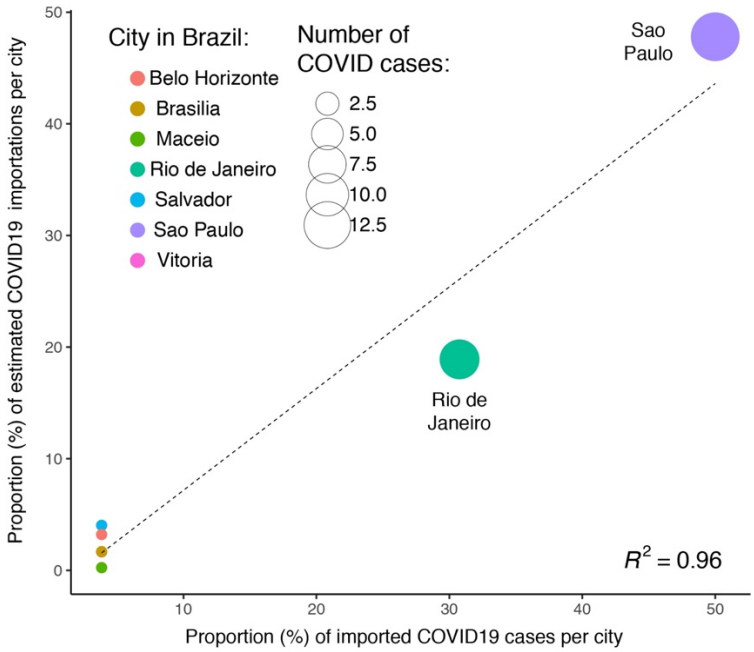


Figure S1. Correlation between the estimated proportion of imported COVID-19 cases per Brazilian city and the proportion of COVID importations reported Brazilian Ministry of Health (as of 9th March 2020). Circles are coloured according to location in Brazil and sized according to the number of COVID-19 imported cases reported Brazilian Ministry of Health (as of 9th March 2020). A linear regression model was fitted to the data using RStudio Version 1.2.1335.

References

1. Passenger Intelligence Services (PaxIS). Montreal: International Air Transport Association; 2017. Available from http://www.iata.org/services/statistics/intelligence/paxis/Pages/index.aspx [cited 2018 Jun 11].

2. WHO. Coronavirus disease (COVID-2019) situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports; 2020.

nature research

https://doi.org/10.1038/s41562-020-0928-4

Supplementary information

Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil

In the format provided by the authors and unedited

Supplementary information

Geospatial analysis

We adopted a Bayesian hierarchical model to compute relative risk for each census tract, due to the following reasons: (i) there is a large number of census tracts (n=30,815), (ii) there is substantial heterogeneity in the size of census tracts, and (iii) small counts in each tract obscure the spatial distribution of observed cases. The number of observed cases in census tract i is modelled using a Poisson distribution $Y_i = Poisson(\lambda_i)$ with mean $\lambda_i = E_i\,\mu_i$ where E_i is the expected number of cases under a null model in which cases are uniformly distributed among the population. For example, the total number of cases in the MRSP multiplied by the proportion of the population in the census tract $E_{it} = \frac{\sum_i Y_i}{\sum_i pop_i} \times pop_i$. The factor of-$\mu_i$ describes tract specific risk and models the additional variation in the observation process[1]. A log-linear model is used to estimate the relative risk μ_i. For example, the log relative risk is expressed as a sum of an intercept α, which represents the overall relative risk (in our case, the global relative risk is zero), and random effects (Z_i):

$$\log(\mu_i) = \alpha + Z_i$$

We used a Besag-York-Mollié model (BYM)[2] to separate the random effects into a spatially structured U_i, and independent random effects, V_i, so ($Z_i = U_i + V_i$). In the BYM model, a conditional autoregressive (CAR) process is used to introduce correlation among the U_i for each tract. Given the U_i of neighbouring tracts, the U_i has a normal distribution with mean equal to the average of the neighbours' U_i, and variance $s_i^2 = \frac{1}{\#N(i)\tau_U}$ where $\#N(i)$ is the number of tracts that share boundaries with tract i and τ_U is a precision parameter. The random effect, V_i follows a zero mean normal distribution with unknown precision, τ_V=$\frac{1}{\sigma_{v^2}}$ (where σ_{V^2} is the variance). Both random effects in the model capture extra-Poisson variability, and were expressed as the following:

$$U_i|\, U_{j\neq i} \sim Normal\left(m_i, s_i^2\right), \quad V_i \sim N(0, \sigma_v^2)$$

$$m_i = \frac{\sum_{j\in N(i)} U_j}{\#N(i)} \quad , \quad s_i^2 = \frac{\sigma_U^2}{\#N(i)} = \frac{1}{\#N(i)\tau_U}$$

The log of the precision parameters, τ_U and τ_V, follows a gamma distribution with shape 1 and rate 0.0005. These are the default priors used by R-INLA and are minimally informative[3]. The prior default distributions in R-INLA were used for the precision parameters of both U_i and V_i. These are minimally informative and are the recommended settings [4].

To quantify the uncertainty in the point estimates of the mean relative risk estimates, we mapped the posterior probability of elevated relative risk in each census tract (**Extended Data Fig. 9**). This is the posterior probability, which a tract has an elevated risk of observing cases, formally, this is Prob(μ_i > 1| data). For instance, a probability of 0.6 in a census tract indicates a 60% chance that this census tract is at greater risk of observing cases relative to the rest of the MRSP.

Analysis of the relationship between income per capita and final diagnostic category in the Metropolitan Region of Sao Paulo (MRSP)

We evaluated the relationship between final diagnostic category (COVID-19 or SARI cases with unknown aetiology) and socioeconomic status in the subset of cases in the MRSP with geocoded residential information. We focused on the cases in epidemiological weeks 12, 16 and 22, where the census tracts that reported cases varied across weeks. In each of the three weeks, if a census tract reported any COVID-19 or SARI cases with unknown aetiology with unknown aetiology, we calculated the proportion of the number of COVID-19 cases. Since most census tracts reported only one case each week, the proportion of COVID-19 of each census tract were mostly either 0 or 1 in a given week. Based on this observation and let i index the census tracts, we subsequently defined the binary outcome Y_i of census tract i, where (i) $Y_i = 0$ if census tract i only reported SARI cases with unknown aetiology with unknown aetiology, i.e. no COVID-19 cases, (ii) $Y_i = 1$ if census tract i reported at least one COVID-19 case in the week. Logistic regression models were applied to investigate the association between this binary outcome and the log(X+1) transformed income per capita. The analyses were adjusted by the logarithm of the population sizes. In addition, the census tracts were grouped by their geographic locations using cluster analysis, and the groupings were used as the random effect in the logistic regressions to account for potential spatial autocorrelation. The number of clusters was chosen based on the AIC/BIC values of the logistic regression models. The analysis was performed individually for each of epidemiological weeks 12, 16 and 22.

A likelihood ratio test (LRT) is applied to each analysis to examine whether the log(X+1) transformed income per capita provides information in addition to the information from the log population size and the random effects. The regression coefficients and LRT P-values of income are presented in (**Supplementary Table S3**).

Estimating basic reproduction number (R_0)

Since SARS-CoV-2 is a novel virus, and we are subsetting data to avoid the impact of either non-pharmaceutical interventions or depletion of the susceptible pool, we deemed it reasonable to model the incidence of infection with an exponential approximation to the early behaviour of an SIR model, i.e., the incidence grows exponentially [5]. This model makes several strong assumptions about the dynamics of the epidemic: (i) the populations under consideration mix homogeneously, (ii) the proportion of the population that is susceptible stays close to 100%, (ii) the proportion of infections that are observed, and the basic reproduction number are constant throughout time, and (iv) the delay between infection, and notification is a constant. Although there are obvious violations of these assumptions, they provide a convenient starting point for estimating the basic reproduction number. Ignoring the delay between infection and observation will on average only translate the results in time by the incubation period and the delay from infection to diagnosis.

Under the assumptions outlined above, the expected number of daily cases, $\mu(n)$ on day n is given by the following equation: $\mu(n) = \rho \mathrm{R}_0 \gamma i_0 e^{(\mathrm{R}0-1)\gamma.n}$ where ρ is the probability of an infection being counted in the time series, R_0, is the basic reproduction number, γ is the rate at which individuals cease to be infectious and i_0, is the proportion of the population that was infectious at the start of the observations. We assume that the observed number of cases on day n was drawn from a negative binomial observation where the mean is $\mu(n)$ and the variance, $\sigma = \mu + \mu^2/k$, with fixed size parameter, k (*dispersion parameter*). The product of ρ and i_0 is denoted ξ. Since the probability of being observed and the initial condition only appear as the product ξ in the likelihood, there is an

identifiability problem preventing the estimation of ρ and i_0 individually, consequently we only consider their product, ξ. Although in this model it is theoretically possible to estimate both R_0 and γ, in practice this is difficult so we will use an informative prior to constrain γ to a priori plausible values.

Regarding prior distributions, for R_0 we used a uniform prior over values from 1 to 10. The removal rate, γ, was given an informative prior distribution: a normal distribution with mean (1/5 + 1/14) / 2 = 0.1357, leading to an average duration 7.4 days during which an individual is infectious. Moreover, the average duration of infectivity is constrained to be between the extremes of 5 and 14 days. These values for the infective duration were found in the literature [6,7]. The standard deviation of the prior distribution for γ is (1/5 - 1/14) / 4 = 0.03124, this ensures that 95% of the prior probability lay within these bounds. For the parameter ξ, we used a log-normal prior with a log mean of 0.0 and a log standard deviation of 1.0. For the size parameter of the negative binomial, k, a log-normal distribution was used with a log-mean of 0.0 and log-standard deviation of 1.0 to enable this parameter to have a large range of values.

Samples from the posterior distribution were obtained using MCMC running 4 chains from random initial conditions using the mcmc library available on CRAN2 and using coda for diagnostics[8,9]. Trace plots of the posterior samples suggested that the chain had converged and mixed, and there was an effective size of at least several hundred for each of the 4 parameters of this model. The prior and posterior distributions were checked to ensure that (beyond the removal rate) each parameter was being informed by the data. Each data set: Brazil and Europena countries (Italy, the United Kingdom, France, and Spain) or Brazilian states (São Paulo, Rio de Janeiro, Amazonas, and Ceará) were run as independent analyses, the model fit from the point estimate along with the corresponding trace plots and prior/posterior comparisons is shown in **Extended Data Figs. 5 and 6.**

References

1 Lawson, A. B. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiolog*. (2008).

2 Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1-20, doi:10.1007/BF00116466 (1991).

3 Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 319-392, doi:10.1111/j.1467-9868.2008.00700.x (2009).

4 Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology* **7**, 39-55, doi:10.1016/j.sste.2013.07.003 (2013).

5 Brauer, F., van den Driessche, P. & Wu, J. *Mathematical Epidemiology*. (Springer-Verlag Berlin Heidelberg, 2008).

6 Wolfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature*, doi:10.1038/s41586-020-2196-x (2020).

7 European Centre for Disease Prevention and Control. Novel coronavirus (SARS-CoV-2). (2020).

8 Geyer, C.J., & Jonson, L. T. mcmc: Markov Chain Monte Carlo. R package version 0.9-6 (https://CRAN.R-project.org/package=mcmc, 2019).

9 Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7-11 (2006).

science.sciencemag.org/cgi/content/full/science.abd2161/DC1

Supplementary Materials for

Evolution and epidemic spread of SARS-CoV-2 in Brazil

Darlan S. Candido*, Ingra M. Claro*, Jaqueline G. de Jesus*, William M. Souza*, Filipe R. R. Moreira*, Simon Dellicour*, Thomas A. Mellan*, Louis du Plessis, Rafael H. M. Pereira, Flavia C. S. Sales, Erika R. Manuli, Julien Thézé, Luiz Almeida, Mariane T. Menezes, Carolina M. Voloch, Marcilio J. Fumagalli, Thaís M. Coletti, Camila A. M. da Silva, Mariana S. Ramundo, Mariene R. Amorim, Henrique H. Hoeltgebaum, Swapnil Mishra, Mandev S. Gill, Luiz M. Carvalho, Lewis F. Buss, Carlos A. Prete Jr, Jordan Ashworth, Helder I. Nakaya, Pedro S. Peixoto, Oliver J. Brady, Samuel M. Nicholls, Amilcar Tanuri, Átila D. Rossi, Carlos K.V. Braga, Alexandra L. Gerber, Ana Paula de C. Guimarães, Nelson Gaburo Jr, Cecila Salete Alencar, Alessandro C. S. Ferreira, Cristiano X. Lima, José Eduardo Levi, Celso Granato, Giulia M. Ferreira, Ronaldo S. Francisco Jr, Fabiana Granja, Marcia T. Garcia, Maria Luiza Moretti, Mauricio W. Perroud Jr, Terezinha M. P. P. Castiñeiras, Carolina S. Lazari, Sarah C. Hill, Andreza Aruska de Souza Santos, Camila L. Simeoni, Julia Forato, Andrei C. Sposito, Angelica Z. Schreiber, Magnun N. N. Santos, Camila Zolini de Sá, Renan P. Souza, Luciana C. Resende-Moreira, Mauro M. Teixeira, Josy Hubner, Patricia A. F. Leme, Rennan G Moreira, Maurício L. Nogueira, Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic Network, Neil M. Ferguson, Silvia F. Costa, José Luiz Proenca-Modena, Ana Tereza R. Vasconcelos, Samir Bhatt, Philippe Lemey, Chieh-Hsi Wu, Andrew Rambaut, Nick J. Loman, Renato S. Aguiar, Oliver G. Pybus, Ester C. Sabino†, Nuno Rodrigues Faria†

*These authors contributed equally to this work.
†Corresponding author. Email: sabinoec@usp.br (E.C.S.); nuno.faria@zoo.ox.ac.uk (N.R.F.)

Published 23 July 2020 on *Science* First Release
DOI: 10.1126/science.abd2161

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/cgi/content/full/science.abd2161/DC1)

Fig. S8 (PDF file)
Data S1 (CSV file) and S2 (Excel file)
MDAR Reproducibility Checklist (PDF file)

Materials and Methods

Ethical statement

Residual nasopharyngeal, tracheal and bronchial aspirate samples testing positive for SARS-CoV-2 by RT-qPCR were obtained from public health and private medical diagnostics laboratories (**Table S1**). All samples were de-identified before receipt by the researchers. Ethical approval for this study was confirmed by the national ethical review board (Comissão Nacional de Ética em Pesquisa), protocol number CAAE 30127020.0.0000.0068.

Epidemiological data

We analysed case counts and deaths from the *Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP-Gripe)*. The *SIVEP-Gripe* was created in 2009 for the H1N1 influenza pandemic and centralizes the notification of severe acute respiratory infection (SARI) cases for the Brazilian Ministry of Health. This database contains mostly hospitalized cases, while the non-hospitalized cases are mostly notified in the e-*Sistema Único de Saúde* (eSUS) *Vigilância Epidemiológica* database that is not available for public consultation. The SARI case definition used since 2012 includes hospitalized patients of any age, with a flu-like syndrome (fever and cough or throat pain) that present dyspnoea or O2 saturation <95% or respiratory discomfort. Registered deaths due to SARI are also included independent of hospitalization. The SARI database has been made publicly available on a daily basis and can be retrieved at https://opendatasus.saude.gov.br/dataset/bd-srag-2020 (accessed 1 June 2020). COVID-19 daily case counts for confirmed cases and deaths were downloaded from the official database of the Brazilian Ministry of Health (https://covid.saude.gov.br/). The 2-letter ISO 3166-1 codes for the 27 federal units in Brazil (26 federal states and 1 federal district) are as follows: AC=Acre, AL=Alagoas, AM=Amazonas, AP=Amapá, BA=Bahia, CE=Ceará, ES=Espírito Santo, DF=Distrito Federal, GO=Goiás, MA=Maranhão, MG=Minas Gerais, MS=Mato Grosso do Sul, MT=Mato Grosso, PA=Pará, PB=Paraíba, PE=Pernambuco, PI=Piauí, PR=Paraná, RJ=Rio de Janeiro, RN=Rio Grande do Norte, RO=Rondônia, RR=Roraima, RS=Rio Grande do Sul, SC=Santa Catarina, SE=Sergipe, SP=São Paulo, and TO=Tocantins.

Estimates of human mobility flows

Openly-available Google Community Mobility Reports for São Paulo and Rio de Janeiro city were used to obtain an aggregated estimate of daily percent changes in mobility and include changes in visits to places compared to baseline values for a 5-week period between 3 January to 6 February, 2020 (available at: https://www.google.com/covid19/mobility/) (*41*). We compare the Google Community Mobility Reports to temporally-aggregated anonymised mobile phone mobility data that is freely provided by the Brazilian company *InLoco* which gathers data from more than 60 million mobile devices spread in all areas of Brazil (available at: https://mapabrasileirodacovid.inloco.com.br/pt/) (*42*). Data consists of pairs of origin-destination trips within and between states of Brazil, with approximately 10 to 12 million trip records of more than 1km per day (*42*). Data was anonymized and pre-processed by the company, which has wide national coverage across Brazil, sampling approximately one fourth of the Brazilian population, focused on smartphone users, as previously described (*18, 42, 43*). In this case, daily population-level mobility was measured for São Paulo and Rio de Janeiro from 1 January to 30 April 2020 (baseline 1 January to 29 February).

Reproduction number using epidemiological and mobility data

A Bayesian semi-mechanistic model was used to estimate transmission intensity and attack rates of COVID-19 conditional on the reported number of deaths (available at https://opendatasus.saude.gov.br/dataset/bd-srag-2020). Considering both Rio de Janeiro and São Paulo cities, four covariates related to mobility are considered. These describe the reduction or increase in mobility in parks ($k = 1$), transit stations ($k = 2$) and the average of groceries and pharmacies, retail and recreational areas, and workplaces ($k = 3$). Their average was calculated because these variables are collinear.

In addition, a social isolation index provided by *InLoco* geolocation company and recorded at city level (*42*, *43*) (see previous section) was also introduced in the model as an additional covariate ($k = 4$). The time-varying reproduction number *R* (or R_t) is modelled as a function of Google mobility variables and the city isolation index. The approach is similar to that described by Mellan *et al*. (*22*) but replaces the Google mobility residential covariate with the *InLoco* isolation index, on the basis that a city level index is preferred over state level to model deaths at city level. In the end, we observed that the choice of using the *InLoco* isolation index over the Google residential index has a marginal effect on the R_t predictions (not shown).

Furthermore, to account for residual variation beyond that included in the mobility parameterization of R_t, a second order autoregressive (AR2) random process was included in the model. The AR2 process accounts for residual correlation structure fitting random effects on a weekly basis, from the start of the epidemic in each city, up to the three weeks before the final time reported. For further details on implementation, see (*20*). Model parameters were jointly estimated for both cities using partial pooling. Denote $I_{k,t,m}$ as the k-th Google mobility indicator, at time t for city m. The time-varying reproduction number for city m, $R_{t,m}$, is modelled by:

$$R_{t,m} = R_{0,m} \cdot 2\lambda^{-1}\left(-\sum_{k=1}^{4}(\alpha_k + \beta_{m,k})(I_{k,t,m} + B_k) - \varepsilon_{m,w_m(t)}\right)$$

where λ^{-1} denotes the logistic function, α_k the effects shared between M cities and $\beta_{m,k}$ city-specific effects. $\varepsilon_{m,w_m(t)}$ denotes a weekly (AR2) process that accounts for fitting extra variation not captured by the designated covariates. B_k denotes noise in the baseline and is set to $B_k \sim Normal(0,0.25)$. Prior distributions for the partial pooling model were set as:

$$\alpha_k \sim Normal(0,0.5)$$

$$\beta_{m,k} \sim Normal(0,\gamma), \text{ with } \gamma \sim N(0,0.5)$$

while the prior distribution for $R_{0,m}$ was chosen to be:

$$R_{0,m} \sim Normal(3.28,|\kappa|)$$
$$\kappa \sim Normal(0,0.5)$$

with κ being the same between both cities to share information about the variability of $R_{0,m}$. The value of 3.28 was used in (*21*, *22*) based on (*44*).

We note that identifying the outcome NPI relaxation is inherently challenging due to the asymmetric nature of the changes in population behaviour that NPIs affect - typically an initial step-like response followed by slow release. For example, mobility indicators suggest a non-negligible fraction of the population began to relax their behaviour almost shortly after the interventions (that were not implemented strictly), as evidenced by the slight upward trend in *R* for São Paulo (see **Fig. 1C**, **Table S1**), while a substantial proportion of the population effectively continues to observe mandated NPIs even after relaxation.

Sample and metadata collection

SARS-CoV-2 convenience samples were collected between March 5 and April 30, 2020, from patients residing in 18 of the 27 Brazilian federal states. Positive samples were provided for diagnostic, confirmatory testing or genome sequencing purposes by public and private laboratories. SARS-CoV-2 diagnosis was performed using the Charité and/or the Centre for Disease Control real-time quantitative polymerase chain reaction (RT-qPCR) assays (*24*, *45*). Residual samples were processed for genome sequencing at the Institute of Tropical Medicine-University of São Paulo, National Laboratory for Scientific Computation and University of Campinas (**Table S1**). Minimum metadata for processed samples included date of sample collection (day, month and year), sex, age, municipality and state of residence in Brazil. To guide our sequencing efforts and to maximize data representativity, the number of new samples processed per state was selected to maximize its correlation with cumulative number of confirmed cases per state, according to up to date case count information (see section on Epidemiological Data).

Virus multiplex PCR amplification

SARS-CoV-2 positive samples were sequenced using a targeted multiplex PCR amplicon approach with the MinION sequencing platform (Oxford Nanopore Technologies, ONT, UK). RNA was converted to cDNA using the Protoscript II First Strand cDNA synthesis Kit (New England Biolabs, UK) and random hexamers or SuperScriptIV First-Strand Synthesis System (Thermo Fisher Scientific, USA). Whole-genome amplification was performed with multiplex PCR amplification using the SARS-CoV-2 primer scheme (V1 to V3) and Q5 High-Fidelity DNA polymerase (New England Biolabs, UK) (*46*) (https://artic.network/ncov-2019). PCR products were cleaned-up using AmpureXP purification beads (Beckman Coulter, United Kingdom) and quantified using fluorimetry with the Qubit dsDNA High Sensitivity assay on the Qubit 3.0 instrument (Life Technologies, USA). Amplicons from each sample were normalised and pooled in an equimolar fashion and barcoded using the EXP-NBD104 (1–12) and EXP-NBD114 (13–24) Native Barcoding Kits (Oxford Nanopore Technologies, UK).

Whole genome sequencing and genome assembly

Sequencing libraries were generated using the SQK-LSK109 Kit (ONT, UK) and were loaded onto an R9.4.1 flow-cell (ONT, UK). RAMPART software from the ARTIC Network (https://artic.network/ncov-2019) was used to monitor the sequencing run in real-time to estimate the depth of coverage (target of 200-fold) across the genome for each barcoded sample (https://artic.network/rampart) and samples were sequenced between 8 to 48 hours. After the completion of the sequencing runs, fast5 files were basecalled, demultiplexed, and trimmed

using Guppy software v2.2.7 (ONT, UK). The consensus genomes were obtained by the mapping of fastq files to the reference genome of SARS-CoV-2 isolate Wuhan-Hu 1 (GenBank Accession Number MN908947) using minimap2 v2.28.0 and converted to a sorted BAM file using SAMtools (*47*). Length filtering and the quality test was performed for each barcode using ARTIC guppyplex (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html). The genome statistics were obtained from SAMtools and the Tablet viewer (*48*) and to recover consensus sequences, called variants were detected with nanopolish. Individual nanopore sequencing statistics can be found in **Data S1**. Genome regions with a depth of <20-fold were not included in final consensus sequences, and these positions are represented with N characters.

Quality control of genome consensus sequences

To ensure the quality of our downstream analyses, we undertook stringent quality control steps on a total of 499 genomes generated for this study. Firstly, only sequences with genome coverage above a given cut-off were included to guarantee the highest possible phylogenetic accuracy of the resulting datasets. Given the current lack of genome coverage thresholds for SARS-CoV-2 phylogenetic studies, we used a conservative cut-off of 75.0% to guarantee phylogenetic accuracy. Thus, we removed 72 partial genomes (mean genome coverage of 46.9%, range 0.1 to 74.6%) from our data. We used MAFFT automatic algorithm to build a multiple sequence alignment of the resulting dataset (*49*). We estimated maximum likelihood phylogenies using an alignment of 427 near-complete and complete genomes using a Hasegawa-Kishino-Yano (HKY + Γ) nucleotide substitution model (*50*) with a gamma distribution to describe among-site variation in the rate of nucleotide substitution (*51*) in IQTree v.2 (*52*). We then regressed root-to-tip genetic divergence against sampling dates to investigate the temporal signal of our datasets and to identify sequences with low data quality (e.g. with assembling issues, sequencing and alignment errors, data annotation errors and sample contamination) (*53*). No obvious outliers were identified in this step. Finally, we assessed genome sequence quality through quality control scores and identified virus lineages using Pangolin (https://github.com/cov-lineages/pangolin) (*6*) and CoV-GLUE (cov-glue.cvr.gla.ac.uk/) (*54*). All sequences passed the quality control steps.

Collation of SARS-CoV-2 global datasets

Our genome dataset contains 427 near-complete and complete new genomes from 18 out of 27 Brazilian states. We appended this data to 63 other Brazilian genomes available in GISAID (*29*) until May 5, 2020 (https://www.gisaid.org), generating a dataset of 490 Brazilian genomes that covers 21 out of the 27 Brazilian states. Sampling collection dates of Brazilian sequences ranged from February 25 2020 [first reported case in Brazil (*55*)] to 30 April 2020. The Brazilian datasets represent approximately 1 sequence for every 200 cases (0.5%) notified up to April 30, 2020; including 3.6% of all cases notified in the city of São Paulo as of 5 April 2020 (the day of our most recent sequence from São Paulo), 1.27% of all cases notified in the city of Rio de Janeiro as of 24 April 2020 and 3.13% of all cases notified in the city Fortaleza as of 6 April 2020 (the day of our most recent sequence from Fortaleza). Representativity of the genome data with regards to the number of cumulative SARS-CoV-2 cases in each state up until the date of the last sequenced sample (30 April 2020) can be found in **Fig. 2A** and **Fig. S2**. We appended Brazilian data to two global datasets prepared from genome data deposited in GISAID (https://www.gisaid.org). The first global dataset contains 710 subsampled sequences to include one genome per country per day (based on sample collection day) available until April 24, 2020

(named hereafter as "subsampled dataset"). The second dataset ("full dataset") contains 13,406 sequences made available until May 5, 2020.

In silico analysis of molecular diagnostic assays

The presence of frequently identified mismatches in several diagnostic RT-qPCR assays suggests that certain assays may be less appropriate for use in Brazil than other diagnostic assays in which no mismatches were identified. We used a custom Python script to analyse mismatches between sequences of Brazilian SARS-CoV-2 genomes to sequences of primers and probes used in 13 common assays (*24, 56-61*).The relevant binding site sequence in every genome was compared to each primer or probe sequence, and the position and bases of any mismatching sites were recorded. If any unknown bases (i.e., Ns) were present in the binding site sequence, that genomic sequence was excluded in analyses of the relevant primer or probe. If other (i.e., non-N) ambiguous bases were present in the primer or probe or within the genomic binding site, the genome was not excluded. In such cases, a mismatch is recorded if the set of bases allowable by the primer/probe sequence does not intersect with the set of bases allowable by the genomic binding site sequence. For each primer, we plotted the proportion of mismatching bases at each site in the genomes considered from the alignment, and coloured by the mismatching base in the Brazilian sequence (**Fig. S5**).

Phylogenetic analysis of SARS-CoV-2 in Brazil

Maximum likelihood phylogenies were estimated for the global subsampled and full datasets in IQTree v.2 (*52*) using an HKY + Γ (*50, 51*) as described above. As recombination is a relatively frequent evolutionary mechanism in coronaviruses, we screened our datasets for recombination using the Phi-test approach (*62*) in SplitsTree (*63*) and all available methods in RDP4 (*64*). No evidence of recombination was found in either datasets. Dated phylogenies were estimated under HKY + Γ nucleotide substitution model and a strict molecular clock in BEAST v.1.10.4 (*65*) that assumes constant evolutionary rates throughout the phylogeny. Bayesian analyses were run using BEAGLE (*66*) in duplicate for a length of 250 million Markov chain Monte Carlo (MCMC) steps using both a parametric exponential growth tree prior and a non-parametric skygrid tree prior (*67*). For the skygrid model we used 24 grid points that corresponded to the approximate number of weeks between the x-intercept in the root-to-tip distance correlation with sampling dates obtained from the ML tree of the global subsampled dataset. A non-informative continuous time Markov Chain (CTMC) prior (*68*) was used for the clock rate. Convergence of the MCMC chains was inspected using Tracer v.1.7.1 (*69*). After removal of 10% burn-in, log and tree files were combined and resampled using LogCombiner v.1.10.4. (*65*) to obtain a posterior sample of 1,000 dated phylogenetic trees. Maximum clade credibility (MCC) summary trees were generated using TreeAnnotator v.1.10.4 (*65*). Brazilian clades were defined as clades identified in the Bayesian MCC trees with three or more sequences sampled in Brazil falling in the same clade with >75% of the sequences in that clade collected in Brazil (**Fig. S10**). Clade ages were obtained directly from MCC trees.

Temporal phylogeography

To reconstruct geographic history of SARS-CoV-2 in Brazil we modelled instantaneous transitions between locations in the global subsampled dataset using a discrete asymmetric phylogeographic approach (*70*). To enhance computational time, analyses were run on an empirical distribution posterior dated trees as previously described (*71*). We considered several

discretization schemes. First, taxa were assigned to two locations, "Brazil" and "Others" (*k*=2, scheme A). Second, taxa were assigned to the following locations: "North America", "Europe", "Asia", "Oceania", "Africa", and to the five Brazilian regions of "Southeast", "Northeast", "North", "Centre-West", "South" (*k*=10, scheme B). Thirdly, we considered movement across states in Brazil (*k*=21, scheme C). We then estimated the number of migration events over time on a branch-by-branch basis using a Markov jumps (*72-74*) approach implemented in BEAST v.1.10.4 (*65*). Discretization scheme A was used to quantify the number of virus lineage introductions in Brazil (see annotated tree in **Fig. 3A** and **Fig. S9**). To count the number of migrations (i) from locations outside Brazil to any Brazilian location and (ii) from one Brazilian location to another Brazilian location, relative to the total number of transitions during the same time interval we used scheme B (**Fig. 3B** and **Fig. S11**). Finally, to investigate source-sink dynamics of virus spread within Brazil, we estimate the number of transitions into and out of each state considering locations discretized as in scheme C (**Fig. S10**).

To model phylogenetic diffusion of Brazilian lineages across the country, we used a flexible relaxed random walk (RRW) diffusion model that accommodates branch-specific variation in rates of dispersal with a Cauchy distribution (*75*). We focus on Brazilian clades with three or more strains as inferred in our discrete phylogeographic analysis. To enhance computational time, we use a fixed time-georeferenced MCC tree and we estimate the evolution of number of virus lineage movements within a given federal unit and between federal units in Brazil across its evolutionary history, as described recently (*37*). In brief, for each sequence, latitude and longitude were attributed to a point randomly sampled within the patient's municipality of residence (samples were derived from a total of 100 municipalities out of the 5,570 municipalities in Brazil, Data S1). MCMC chains were run for >10 million generations and sampled every 1000th step, with convergence assessed using Tracer v1.7 (*68*). We used the R package "seraphim" (*76*, *77*) to extract and map spatiotemporal information embedded in posterior trees.

Air travel mobility data

Data on air passenger flows were analysed using public data produced by Brazil's Civil Aviation Agency (ANAC). These data provided detailed information, including the number of passengers and connections, for every international flight to and from Brazil, as well as national flights within the country. The data set is available at https://www.anac.gov.br/assuntos/setor-regulado/empresas/envio-de-informacoes/base-de-dados-estatisticos-do-transporte-aereo. Using ANAC's data, we calculated the daily number of passengers who disembarked or had a national connection in each Brazilian city between January and April in 2019 and 2020 disaggregated by airport and country of origin. When aggregating international passenger flows, we considered the final destination where they disembarked whenever this information was available. Information was missing for 23,254 (60%) international flights that arrived in Brazil between January and March of 2019 and 2020. As a result, the numbers of passengers arriving in international airport hubs in Brazil might be overestimated. A related limitation of this database is that it does not track individual trajectories. Consequently, passengers who make connections on flights with different numbers are counted in the origin-destination pair of the last leg of the trip.

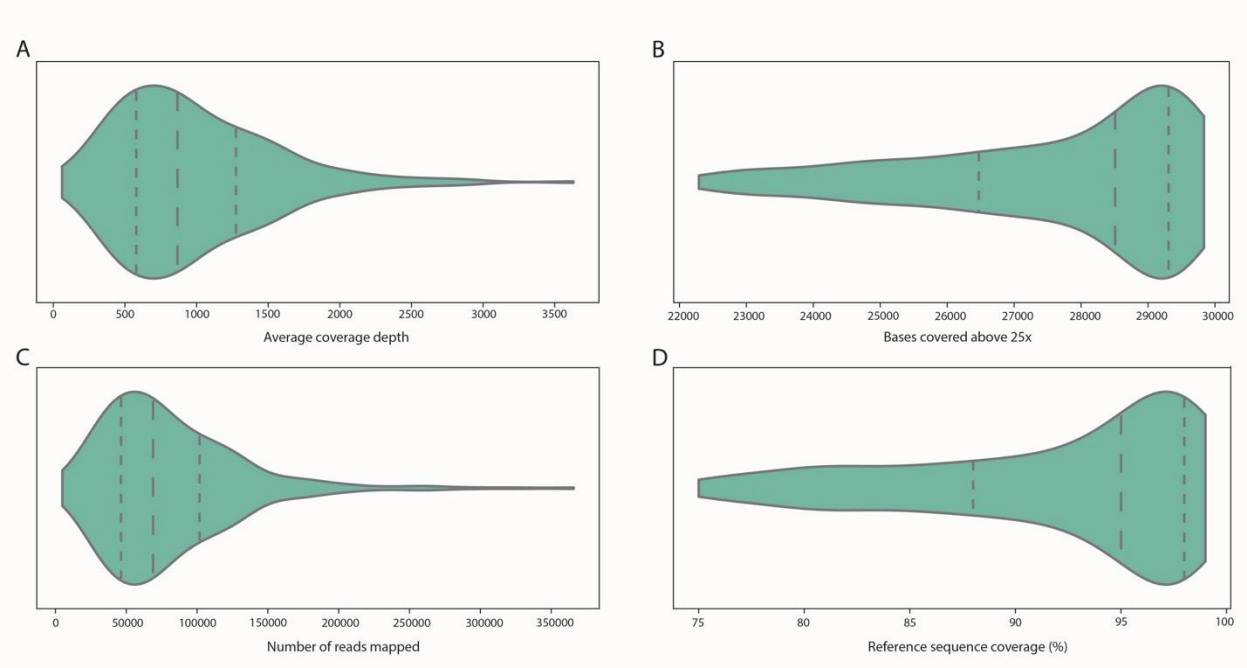


Fig. S1

Sequencing statistics. (A) Average coverage depth, (B) number of bases whose coverage is above 25x, (C) genomic distribution of the number of mapped reads and (D) percentage of covered bases for sequences generated by this study.

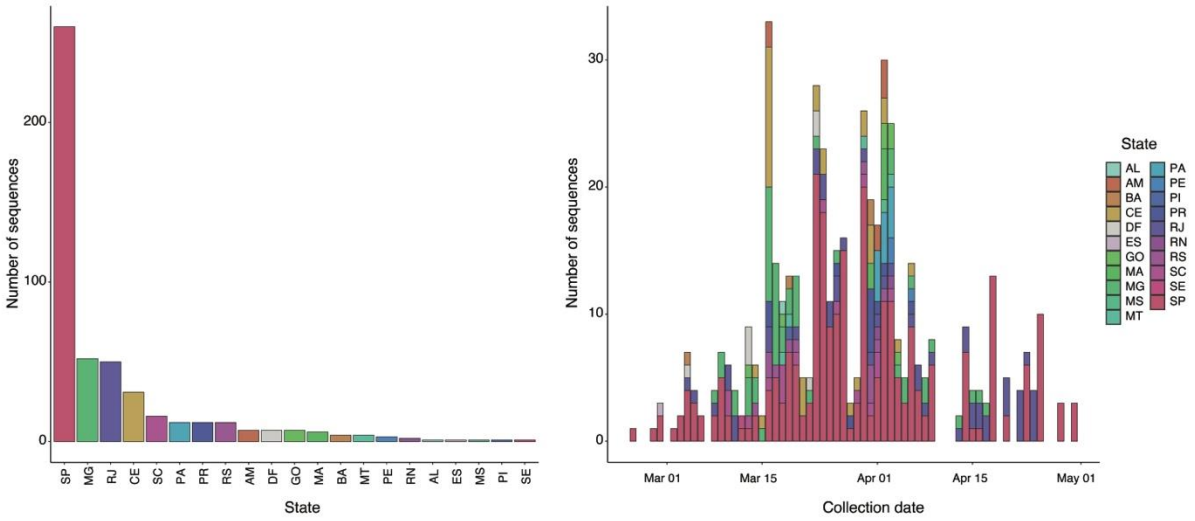


Fig. S2

Distribution of Brazilian SARS-CoV-2 genomes (*n*=427) by state and collection date. (A) SARS-CoV-2 genomes are grouped according to federal state of sample collection. Sampling per state was proportional to the number of severe acute respiratory illness (SARI) cases for each state. (B) Date of SARS-CoV-2 genomes grouped according to federal state of collection. Colours represent federal state of origin. Dates range from the 25 February (first reported Brazilian case) to the 30 April 2020.

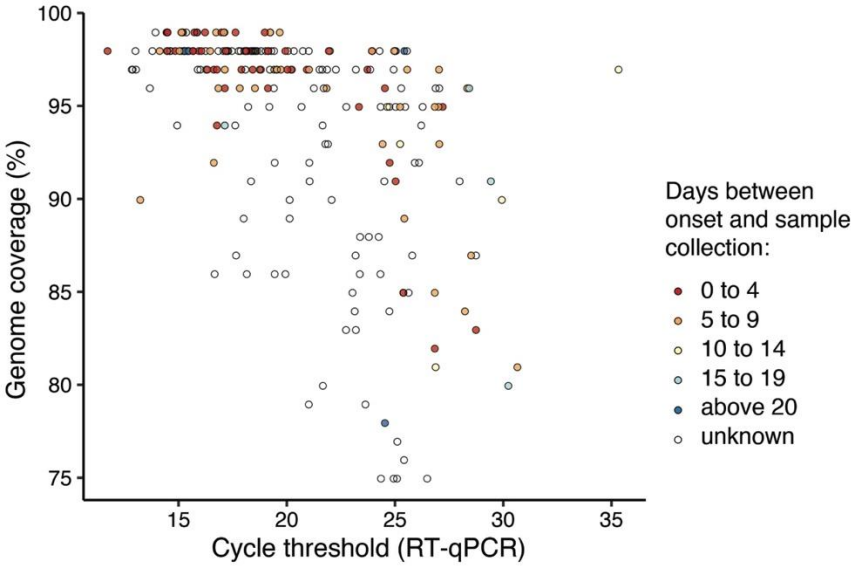


Fig. S3

Genome coverage plotted against RT-qPCR cycle threshold value. Each circle corresponds to a sequenced genome with 20-fold coverage >75%. Each sample is coloured by the number of days between onset of symptoms (red to blue) to date of sample collection. Circles with no colour indicate samples for which information on onset of symptoms was unavailable.

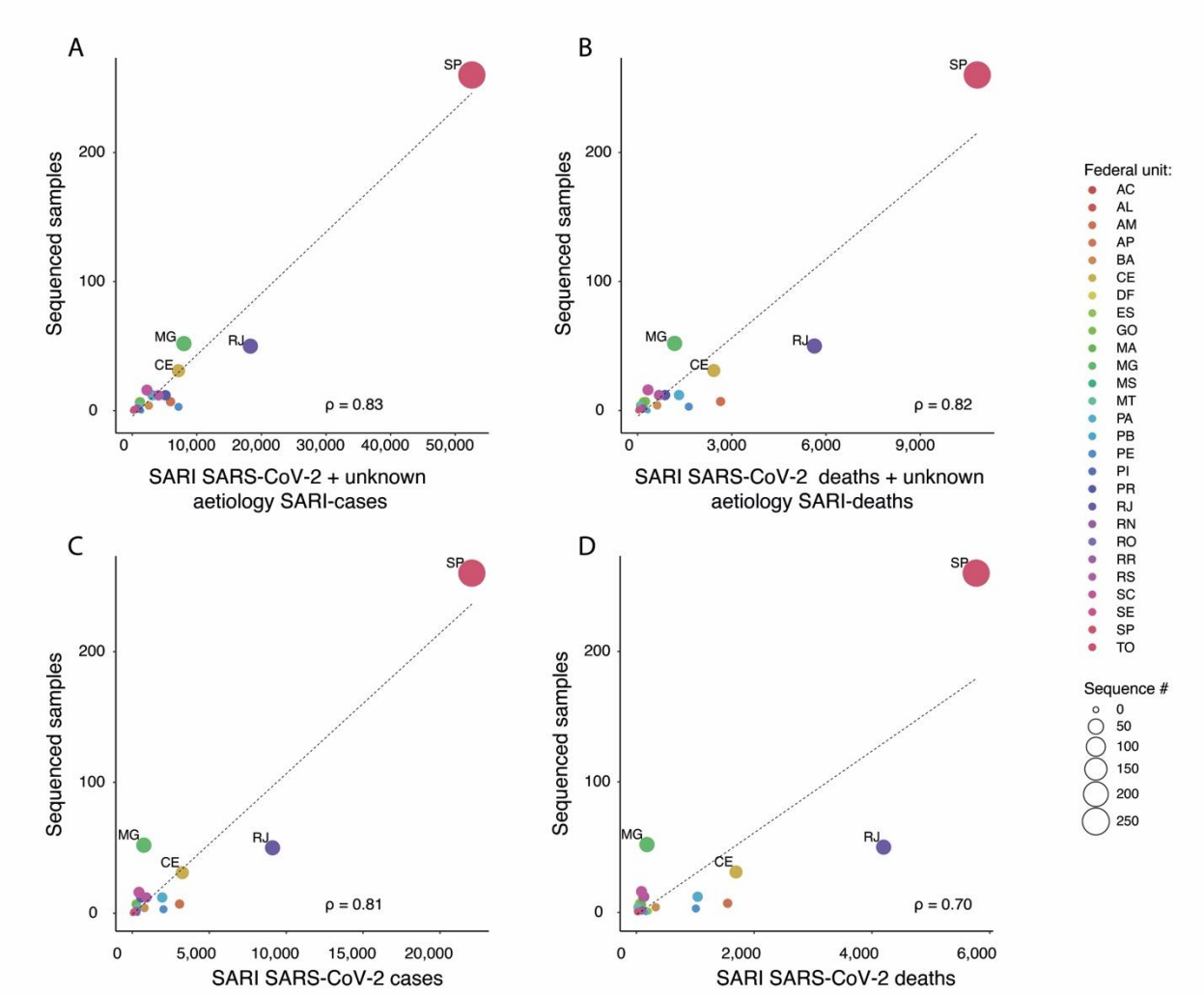


Fig. S4

Spatial representativity of the genome data generated in this study (n=427) and publicly available sequences from Brazil available in GISAID (n=63). SARI cases (A) and deaths (B) of SARS-CoV-2 confirmed cases plus cases with unknown aetiology (excluding those cases that tested positive for other respiratory pathogens). SARS-CoV-2 SARI cases (C) and deaths (D) correspond to cases and deaths confirmed to be SARS-CoV-2 positive using molecular, clinical and epidemiological criteria. Epidemiological data is available at https://opendatasus.saude.gov.br/dataset/bd-srag-2020. Cumulative cases until the 30 April 2020 (date of the most recent genome sequence) were used for the correlation analysis.

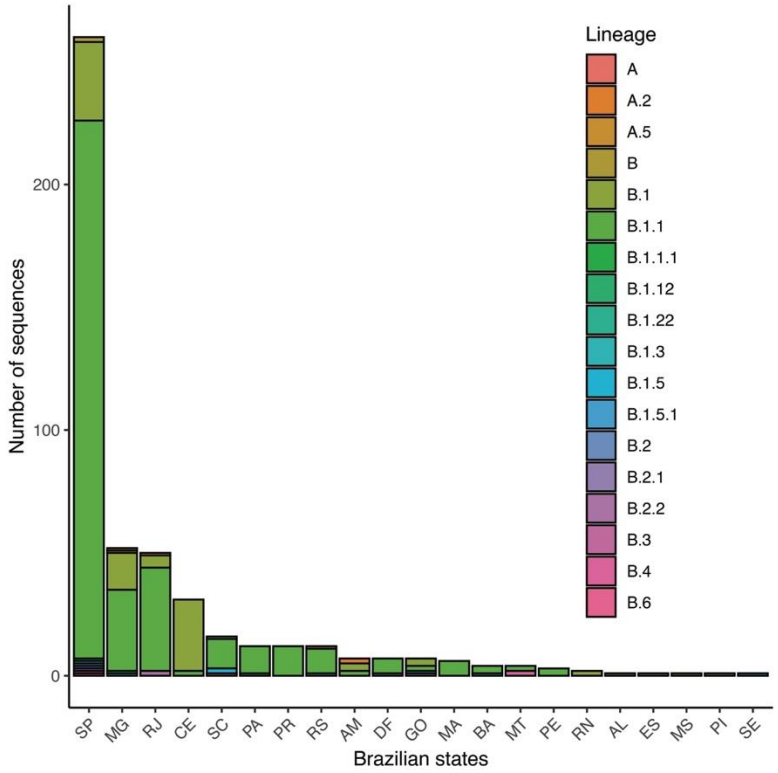


Fig. S5

Distribution of SARS-CoV-2 lineages per Brazilian state. SARS-CoV-2 lineage identification for 490 Brazilian genomes was performed using Pangolin (github.com/hCoV-2019/pangolin) (*6*) and CoV-GLUE (cov-glue.cvr.gla.ac.uk/) (*54*) tools. States are defined by their 2-letter ISO 3166-1 codes.

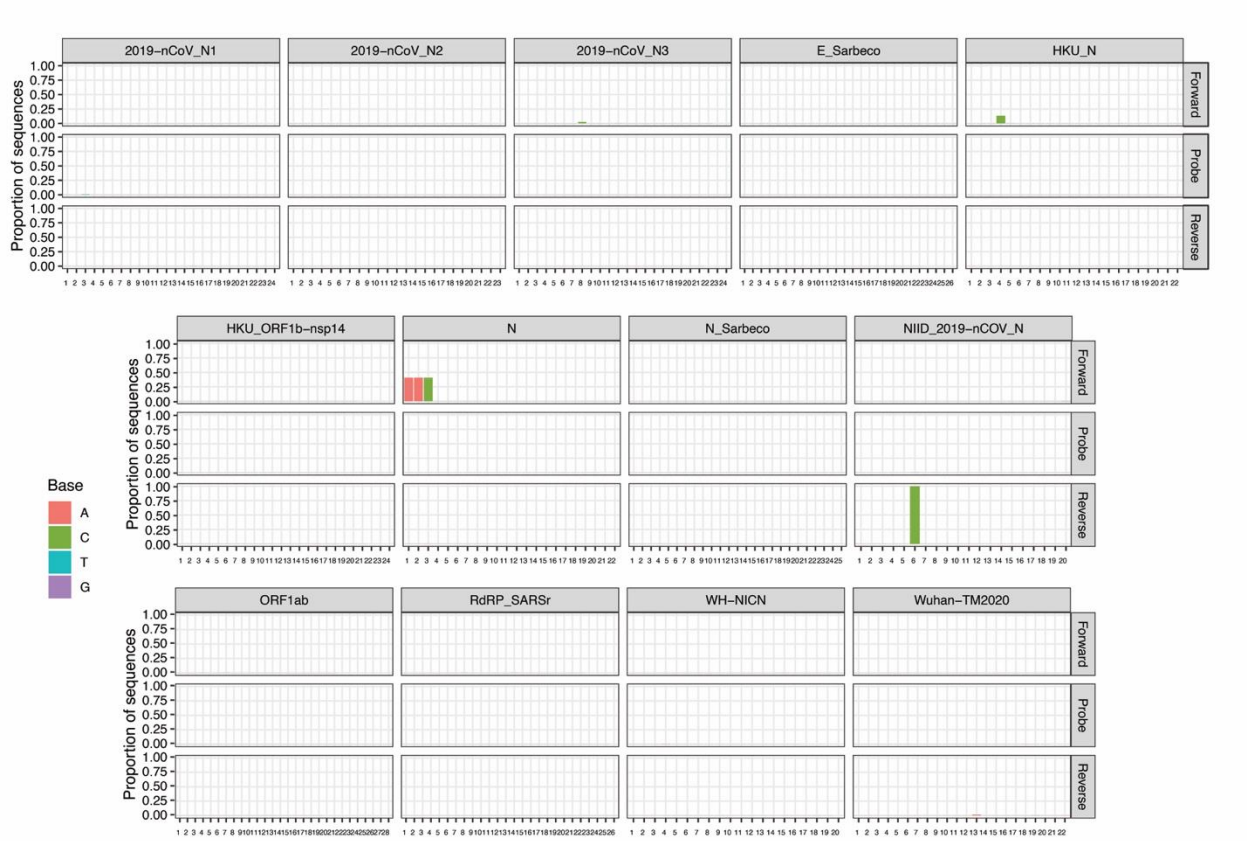


Fig. S6

In silico analysis of primer/probe mismatches to Brazilian strains. Proportion of Brazilian SARS-CoV-2 genomes used on this study that are not exactly complementary to primer/probe bases for each primer/probe site. The number of genomes considered here is variable for each primer or probe because genomes with Ns in primer or probe binding sites were excluded from analyses (as detailed in Materials and Methods). This number, and the sequence information for each primer or probe, is given in Table S2. Site positions are given relative to the primer/probe sequence (5' to 3'). Colours represent bases in the Brazilian SARS-CoV-2 genomes that do not match the primer/probe sequence. Note that the number of positions displayed on the x-axis for each assay is given as the length of the longest primer or probe in that assay, and therefore not all marked site positions with lack of mismatches displayed are relevant.

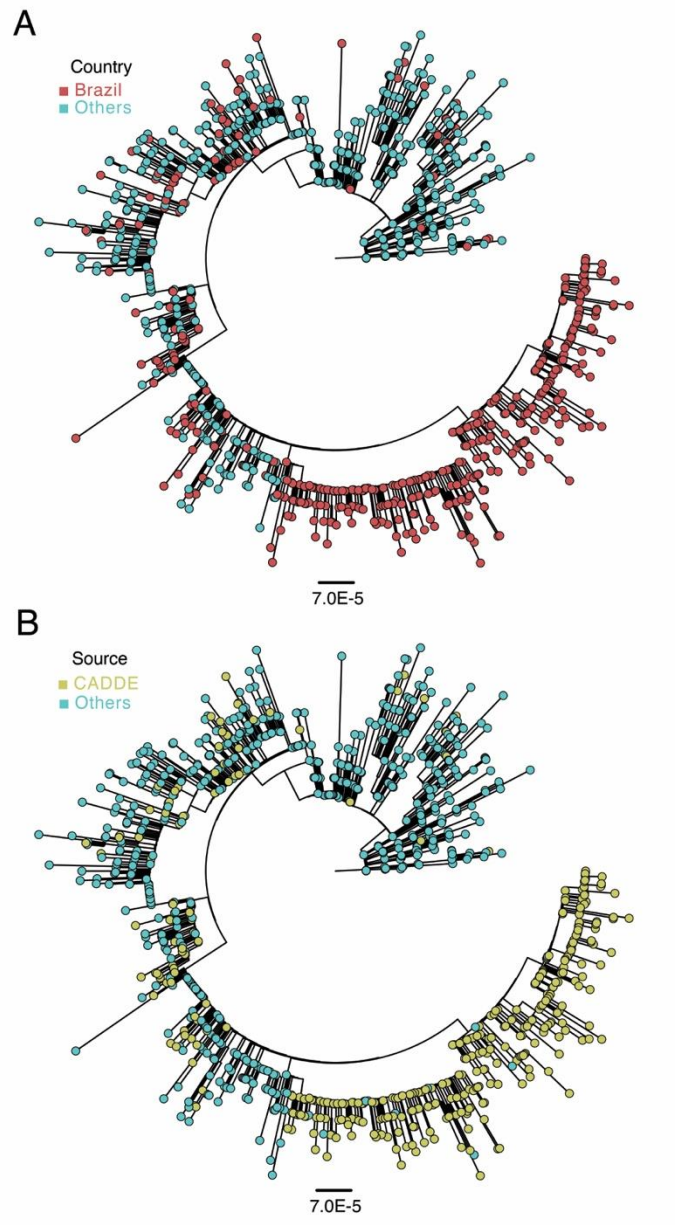


Fig. S7

Maximum likelihood phylogenies estimated using the global subsampled dataset (*n*=1,182) were coloured according to two schemes: (A) "Brazil" *(n*=490) and "Others" (*n*=692) (*k*=2, scheme A) and (B) "CADDE study" *(n*=427) and "Other studies" *(n*=755) (*k*=2).

Fig. S8 (separate PDF file)

Detailed annotated maximum clade phylogenetic tree. Time-resolved maximum clade credibility phylogeny of 1,182 SARS-CoV-2 sequences, 490 from Brazil (red) and 692 from outside Brazil (blue). The largest Brazilian clusters are highlighted in grey (Clade 1, Clade 2 and Clade 3). States are defined by their 2-letter ISO 3166-1 codes. The MCC tree can be found at our Dryad repository (see Data Availability).

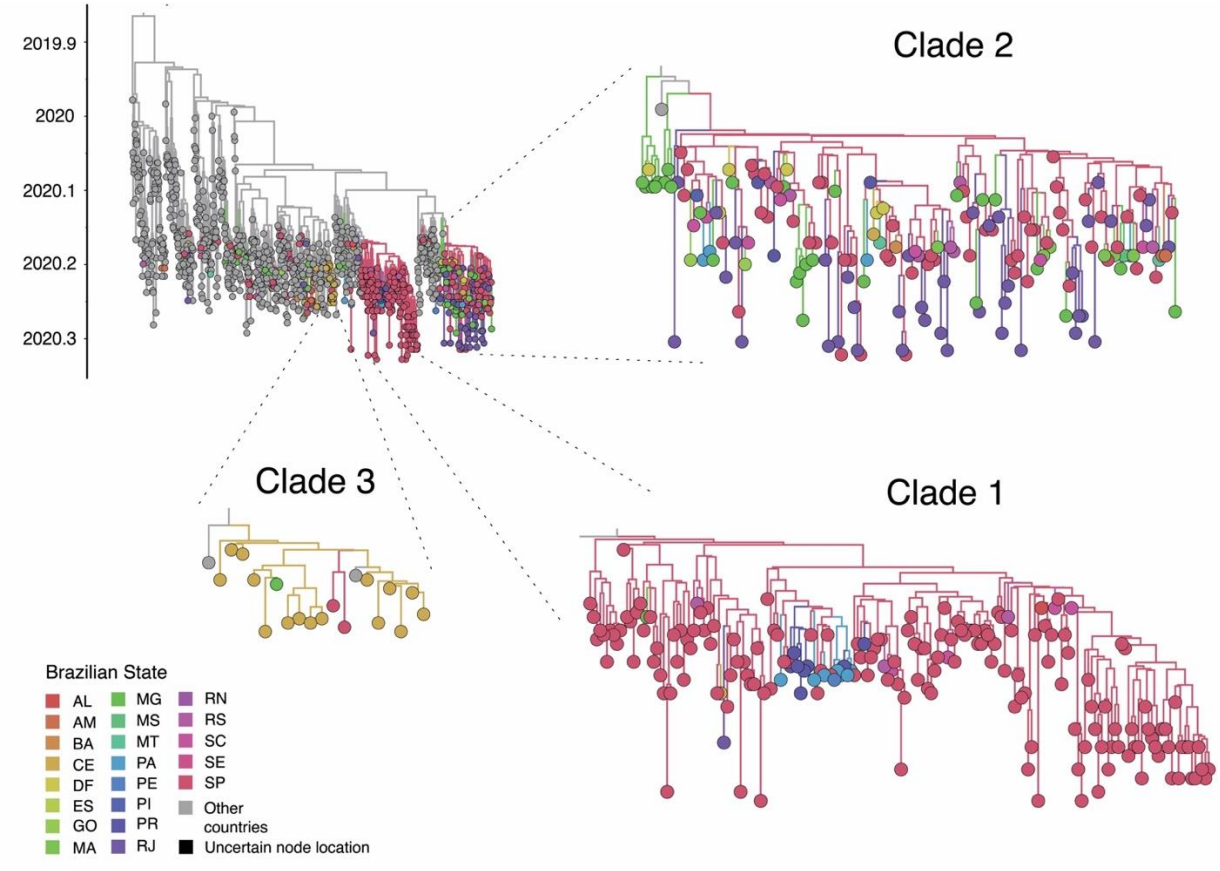


Fig. S9

Detailed view of main Brazilian phylogenetic clades. Maximum clade credibility tree was generated from 490 Brazilian genomes plus 692 genomes from other countries under a phylogeographic discrete trait analysis implemented in BEAST v1.10.14 (*65*) (for details, see Methods section). Colours are assigned according to Brazilian state. Black colour was assigned to branches with high location uncertainty. States are defined by their 2-letter ISO 3166-1 codes. The phylogeographic MCC tree can be found in our Dryad repository (see Data Availability).

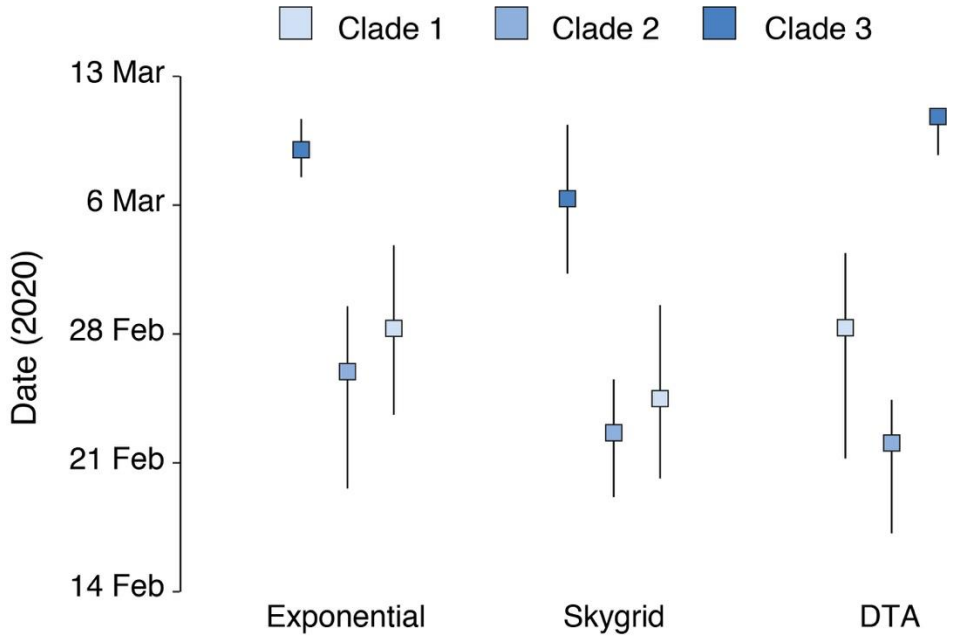


Fig. S10

Estimated dates of emergence of SARS-CoV-2 main clades in Brazil were summarized from MCC trees (median and 95% BCIs) estimated using a parametric exponential growth (left), a non-parametric Skygrid model (centre) and a discrete trait analysis (DTA) as presented in fig. 3A (right). Dated phylogenies were estimated under HKY + Γ nucleotide substitution model and a strict molecular clock that assumes constant evolutionary rates throughout the phylogeny. XMLs used for the Bayesian analyses can be found in our Dryad repository (see Data Availability).

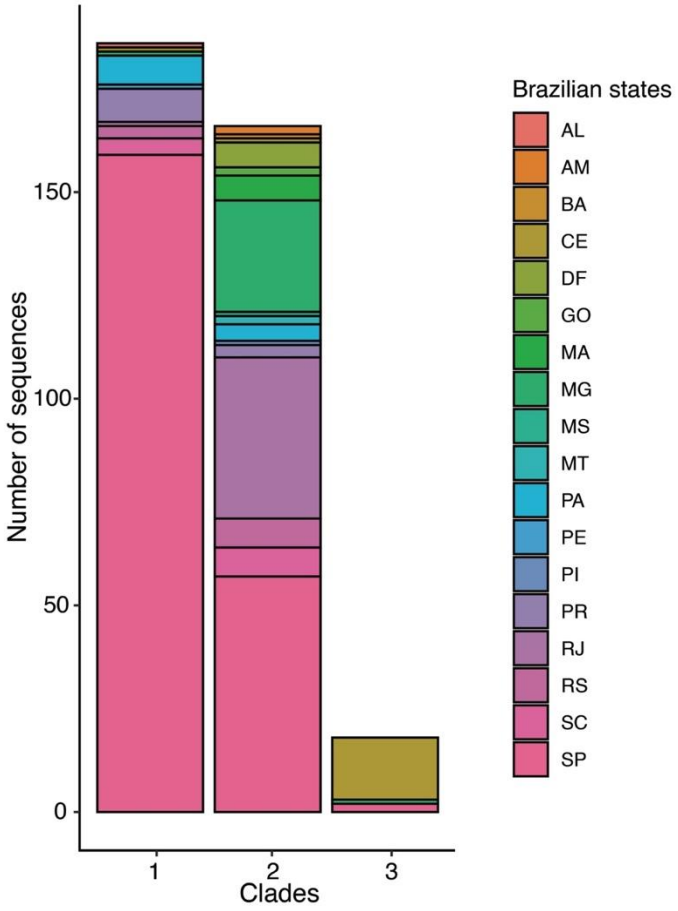


Fig. S11

Geographic distribution of Brazilian SARS-CoV-2 clusters. 490 SARS-CoV-2 Brazilian sequences were grouped according to Brazilian state of collection and SARS-CoV-2 phylogenetic cluster as identified from the Maximum Clade Credibility (MCC) tree generated under a Bayesian phylogenetic approach using BEAST v1.10.14 (*65*) (for details see methods section). States are defined by their 2-letter ISO 3166-1 codes.

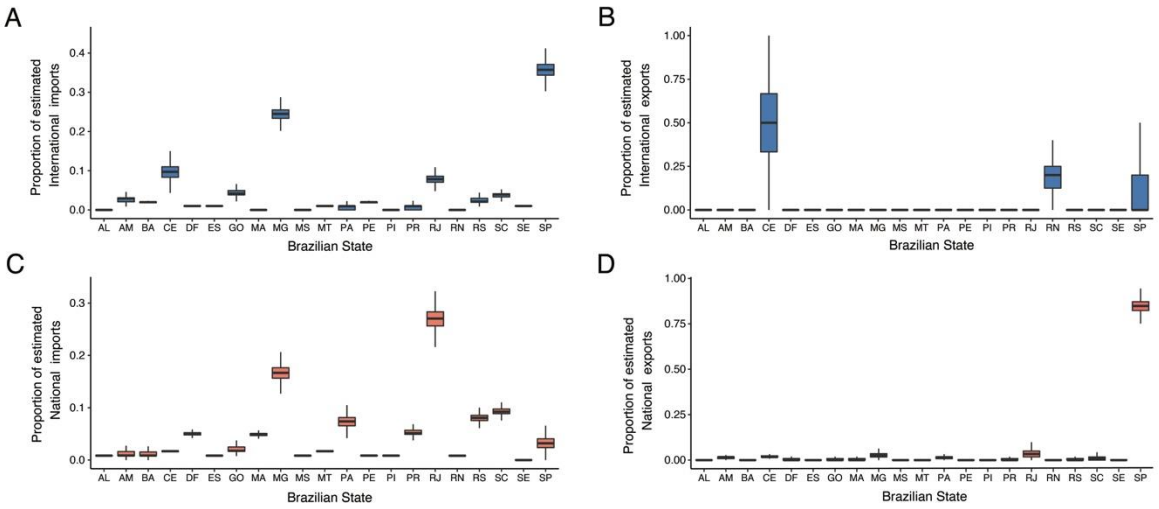


Fig. S12

Estimated proportions of geographic transition events for each Brazilian state. Transitions have been estimated using a phylogeographic approach with Markov jumps implemented in BEAST v1.10.14 (*65*) (see methods for details). For each state the proportion of international (imports and exports) and national (imports and exports) were calculated from the total estimated for each event type. Blue denotes international events, while pink denotes national events. The XMLs used for these analyses can be found at our Dryad repository (see Data Availability).

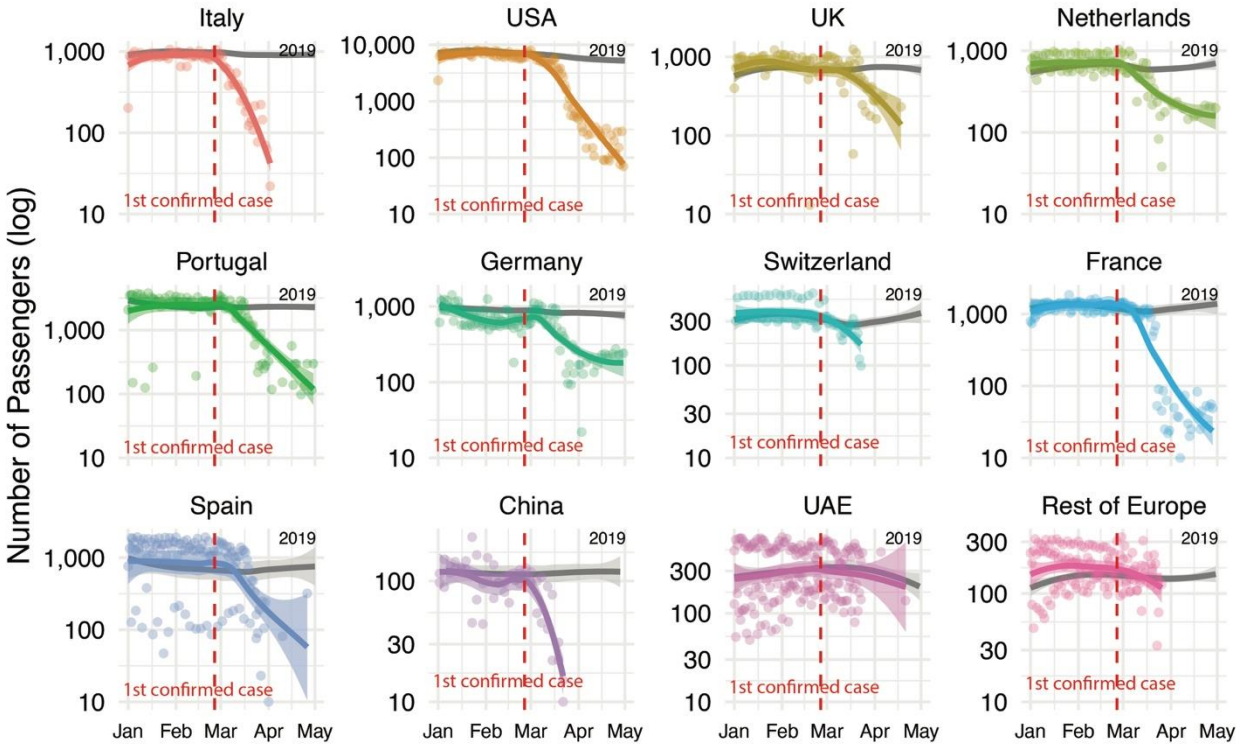


Fig. S13

Impact of the COVID-19 epidemic on international travel to Brazil. Number of inbound international passengers flying to Brazil for the top countries of origin as made available by the National Civil Aviation Agency of Brazil (ANAC). Dark grey dots represent the daily number of passengers in 2019. Coloured plots and lines represent the number of passengers and the trend line for 2020, respectively. Dotted grey line shows the day of the first confirmed case in Brazil.

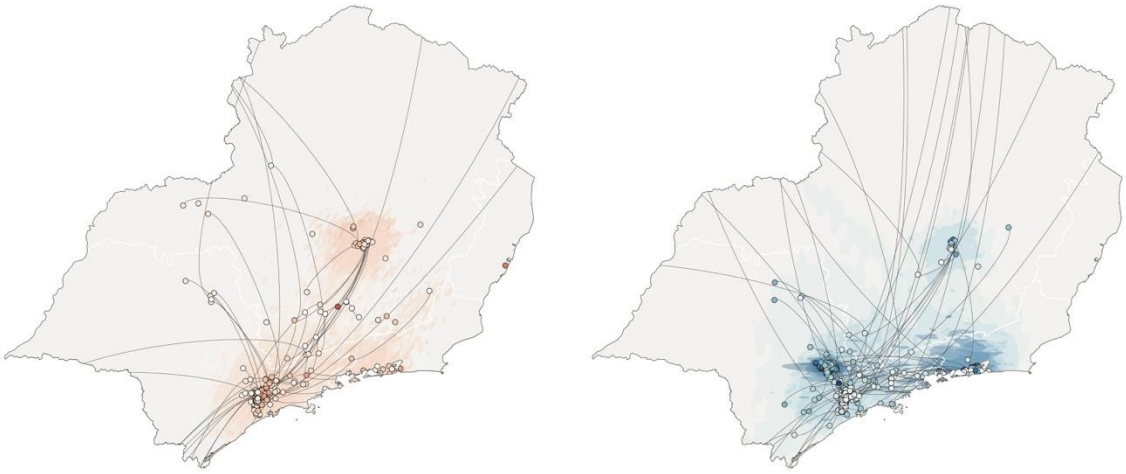


Fig. S14

SARS-CoV-2 spread in southeast Brazil before (left) and after (right) 21 March 2020. This is an expanded version of the maps in fig. 4A of main text. Circles represent nodes of the MCC phylogeny and are coloured according to their inferred time of occurrence. Shaded areas represent the 80% high posterior density (HPD) interval and depict the uncertainty of the phylogeographic estimates for each node. Solid curved lines denote the links between sequences and the directionality of movement. In addition to the clusters included in the continuous analysis, sequences belonging to clusters with less than 3 sequences were also plotted in the map with no lines connecting them.

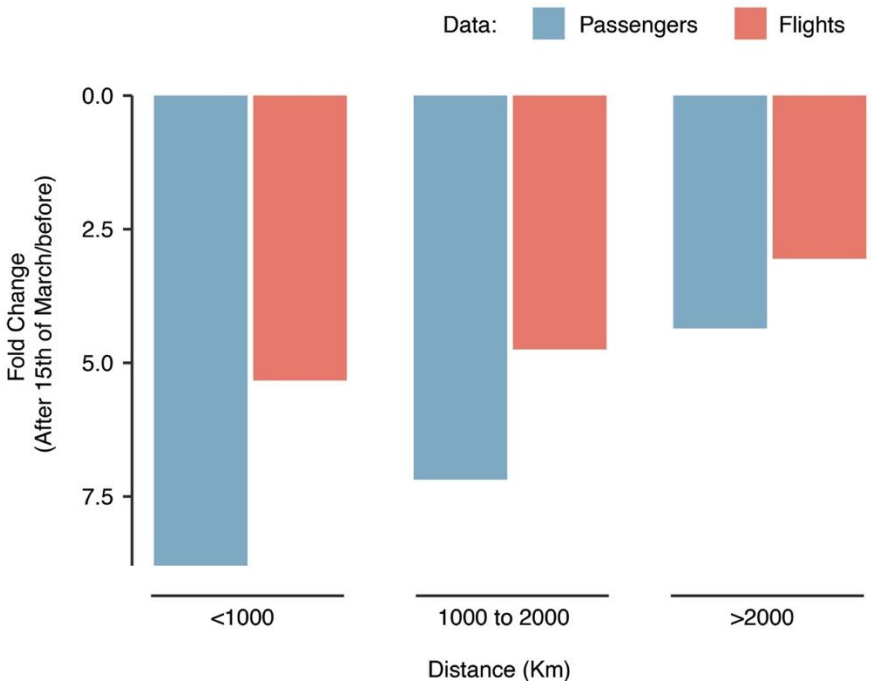


Fig. S15

Impact of the COVID-19 epidemic in domestic air travel in Brazil. Fold change was calculated using the number of domestic flights and passengers before and after 15 March as made available by the National Civil Aviation Agency of Brazil (ANAC). Data was grouped according to flight distance.

Table S1.

City-level estimates of time-varying reproduction number *R* for São Paulo and Rio de Janeiro based on deaths reported in the SARI SARS-CoV-2 dataset (available at https://opendatasus.saude.gov.br/dataset/bd-srag-2020, accessed 1 June 2020). *R* is estimated on 4 May 2020 with 95% Bayesian Credible Intervals (BCIs). 7-day mean values are also given from 27 April 2020 to 4 May 2020.

City	***R* 95% BCI**	***R* 95% BCI 7-day average**
São Paulo	1.3 (1.0, 1.6)	1.2 (0.9, 1.7)
Rio de Janeiro	1.3 (1.0, 1.6)	1.2 (0.9, 1.5)

Table S2.

Genomic laboratories and protocols involved in this study. Ct = real-time quantitative polymerase chain reaction (RT-qPCR) cycle threshold. No. = number. Individual sample level information on sequenced data can be found in Data S1. See also Table S2 for detailed information on the assays used here for molecular diagnostic.

Sequencing Institution	Sample Type	Diagnostic Protocol	Sequencing Protocol	No. generated genomes
IMT-USP	NPS/BAL (n=32), NPS (n=112), NPS/OPS (n=21), OPS (n=3), TS (n=6)	Charité	ARTIC V1 (n=7), V2 (n=210) and V3 (n=56)	273
UFRJ-LNCC	NPS (n=36), NPS/OPS (n=52)	Charité and CDC USA	ARTIC V3 (n=88)	88
UNICAMP	BAL (n=1), NPS (n=64), TA (n=1)	Charité	ARTIC V3 (n=66)	66

Table S3.

Assay, and sequence information for each primer and probe included in the *in-silico* analysis. N represents the number of Brazilian sequences (total of 490) with sufficient information at given primer of probe binding sites.

Assay	Primer/probe	Sequence	N
2019-nCoV_N1	2019-nCoV_N1-F	GACCCCAAAATCAGCGAAAT	479
2019-nCoV_N1	2019-nCoV_N1-P	ACCCCGCATTACGTTTGGTGGACC	478
2019-nCoV_N1	2019-nCoV_N1-R	TCTGGTTACTGCCAGTTGAATCTG	478
2019-nCoV_N2	2019-nCoV_N2-F	TTACAAACATTGGCCGCAAA	435
2019-nCoV_N2	2019-nCoV_N2-P	ACAATTTGCCCCCAGCGCTTCAG	434
2019-nCoV_N2	2019-nCoV_N2-R	GCGCGACATTCCGAAGAA	434
2019-nCoV_N3	2019-nCoV_N3-F	GGGAGCCTTGAATACACCAAAA	473
2019-nCoV_N3	2019-nCoV_N3-P	AYCACATTGGCACCCGCAATCCTG	478
2019-nCoV_N3	2019-nCoV_N3-R	TGTAGCACGATTGCAGCATTG	478
E_Sarbeco	E_Sarbeco_F1	ACAGGTACGTTAATAGTTAATAGCGT	489
E_Sarbeco	E_Sarbeco_P1	ACACTAGCCATCCTTACTGCGCTTCG	479
E_Sarbeco	E_Sarbeco_R2	ATATTGCAGCAGTACGCACACA	479
HKU_N	HKU-NF	TAATCAGACAAGGAACTGATTA	434
HKU_N	HKU-NP	GCAAATTGTGCAATTTGCGG	435
HKU_N	HKU-NR	CGAAGGTGTGACTTCCATG	434
HKU_ORF1b-nsp14	HKU-ORF1b-nsp14F	TGGGGYTTTACRGGTAACCT	481
HKU_ORF1b-nsp14	HKU-ORF1b-nsp14P	TAGTTGTGATGCWATCATGACTAG	484
HKU_ORF1b-nsp14	HKU-ORF1b-nsp14R	AACRCGCTTAACAAAGCACTC	485
N	N_F	GGGGAACTTCTCCTGCTAGAAT	463
N	N_P	TTGCTGCTGCTTGACAGATT	470
N	N_R	CAGACATTTTGCTCTCAAGCTG	460
N_Sarbeco	N_Sarbeco_F1	CACATTGGCACCCGCAATC	478
N_Sarbeco	N_Sarbeco_P1	ACTTCCTCAAGGAACAACATTGCCA	466
N_Sarbeco	N_Sarbeco_R1	GAGGAACGAGAAGAGGCTTG	471
NIID_2019-nCOV_N	NIID_2019-nCOV_N_F2	AAATTTTGGGGACCAGGAAC	436
NIID_2019-nCOV_N	NIID_2019-nCOV_N_P2	ATGTCGCGCATTGGCATGGA	435
NIID_2019-nCOV_N	NIID_2019-nCOV_N_R2	TGGCAGCTGTGTAGGTCAAC	438
ORF1ab	ORF1ab_F	CCCTGTGGGTTTTACACTTAA	488
ORF1ab	ORF1ab_P	CCGTCTGCGGTATGTGGAAAGGTTATGG	298
ORF1ab	ORF1ab_R	ACGATTGTGCATCAGCTGA	295
RdRP_SARSr	RdRP_SARSr-F	GTGARATGGTCATGTGTGGCGG	490

RdRP_SARSr	RdRP_SARSr-P2	CAGGTGGAACCTCATCAGGAGATGC	490
RdRP_SARSr	RdRP_SARSr-R	CARATGTTAAAWACACTATTAGCATA	490
WH-NICN	WH-NICN-F	CGTTTGGTGGACCCTCAGAT	478
WH-NICN	WH-NICN-P	CAACTGGCAGTAACCA	478
WH-NICN	WH-NICN-R	CCCCACTGCGTTCTCCATT	477
Wuhan-TM2020	Wuhan-TM2020For	TCGTGCTACAACTTCCTCAAG	467
Wuhan-TM2020	Wuhan-TM2020Probe	CCGCCTCTGCTCCCTTCTGC	470
Wuhan-TM2020	Wuhan-TM2020Rev	CTGCCWGGAGTTGAATTTCTTG	471

Data S1 (separate CSV file)

Detailed metadata on all 1,182 sequences used in this study. File contains information on epidemiology, demography, location, diagnostics, sequencing statistics and evolution of 427 SARS-CoV-2 sequences generated in this study and 755 sequences downloaded from GISAID.

Data S2 (separate Excel file)

Acknowledgment GISAID table. File contains Accession ID, collection date, originating and submitting lab and authors, from https://www.gisaid.org.

Members of the Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic-Network

Cynthia Chester Cardoso, Orlando da Costa Ferreira Jr., Rodrigo Moraes Brindeiro, Diana Mariani, Alice Laschuk Herlinger, André Felipe Andrade dos Santos, Anna Carla Pinto Castineiras, Camila de Almeida Velozo, Camila Nacif, Camille Victoria Leal Correia da Silva, Caroline Macedo Nascimento, Cassia Cristina Alves Gonçalves, Cíntia Policarpo, Débora Souza Faffe, Ekaterine Simões Goudoris, Elaine Sobral, Elisangela Costa da Silva, Érica Ramos dos Santos Nascimento, Fabio Hecht Castro Medeiros, Fábio Luís Lima Monteiro, Fernando Luz de Castro, Francine Bittencourt Schiffler, Guilherme Sant'Anna de Lira, Helena Keito Toma, Huang Ling Fang, Ingrid Camelo da Silva, Isabela de Carvalho Labarba, Isabela de Carvalho Leitão, Jessica Maciel de Almeida, Joissy Aprigio de Oliveira, Juliana Cazarin de Menezes, Juliana Tiemi Sato Fortuna, Karyne Ferreira Monteiro, Lendel Correia da Costa, Lídia Theodoro Boullosa, Liliane Tavares de Faria Cavalcante, Lucas Matos Millioni, Luciana Jesus da Costa, Marcelo Calado de Paula Tôrres, Matheus Augusto Calvano Cosentino, Mayla Gabryele Miranda de Melo, Mirela D'arc Ferreira da Costa, Pedro Henrique Costa da Paz, Pedro Telles Calil, Rafael de Mello Galliez, Richard Araujo Maia, Sergio Lisboa Machado, Thamiris dos Santos Miranda, Victor Akira Ota, Viviane Guimarães Gomes, Gislaine Celestino Dutra Silva, Marilia Mazzi Moraes, Danielle Alves Gomes Zauli, Joice do Prado Silva, Ana Carolina Fialho Dias, Anna Sara Shafferman Levin, Harrison James Westgarth.

References and Notes

1. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020). doi:10.1038/s41591-020-0820-9 Medline

2. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020). doi:10.1038/s41586-020-2008-3 Medline

3. World Health Organization, *Coronavirus Disease (COVID-2019) Situation Reports* (2020); www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

4. H. Tian, Y. Liu, Y. Li, C.-H. Wu, B. Chen, M. U. G. Kraemer, B. Li, J. Cai, B. Xu, Q. Yang, B. Wang, P. Yang, Y. Cui, Y. Song, P. Zheng, Q. Wang, O. N. Bjornstad, R. Yang, B. T. Grenfell, O. G. Pybus, C. Dye, An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638–642 (2020). doi:10.1126/science.abb6105 Medline

5. M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, S. V. Scarpino; Open COVID-19 Data Working Group, The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020). doi:10.1126/science.abb4218 Medline

6. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* (2020). doi:10.1038/s41564-020-0770-5 Medline

7. T. W. Russell, J. Hellewell, C. I. Jarvis, K. van Zandvoort, S. Abbott, R. Ratnayake, S. Flasche, R. M. Eggo, W. J. Edmunds, A. J. Kucharski; Cmmid Covid-Working Group, Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Euro Surveill.* **25**, 2000256 (2020). doi:10.2807/1560-7917.ES.2020.25.12.2000256 Medline

8. R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, N. M. Ferguson, Estimates of the severity of coronavirus disease 2019: A model-based analysis. *Lancet Infect. Dis.* **20**, 669–677 (2020). doi:10.1016/S1473-3099(20)30243-7 Medline

9. J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. de Salazar, B. J. Cowling, M. Lipsitch, G. M. Leung, Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* **26**, 506–510 (2020). doi:10.1038/s41591-020-0822-7 Medline

10. M. M. Arons, K. M. Hatfield, S. C. Reddy, A. Kimball, A. James, J. R. Jacobs, J. Taylor, K. Spicer, A. C. Bardossy, L. P. Oakley, S. Tanwar, J. W. Dyal, J. Harney, Z. Chisty, J. M. Bell, M. Methner, P. Paul, C. M. Carlson, H. P. McLaughlin, N. Thornburg, S. Tong, A.

Tamin, Y. Tao, A. Uehara, J. Harcourt, S. Clark, C. Brostrom-Smith, L. C. Page, M. Kay, J. Lewis, P. Montgomery, N. D. Stone, T. A. Clark, M. A. Honein, J. S. Duchin, J. A. Jernigan; Public Health–Seattle and King County and CDC COVID-19 Investigation Team, Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *N. Engl. J. Med.* **382**, 2081–2090 (2020). doi:10.1056/NEJMoa2008457 Medline

11. L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, C. Fraser, Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936 (2020). doi:10.1126/science.abb6936 Medline

12. E. Lavezzo, E. Franchin, C. Ciavarella, G. Cuomo-Dannenburg, L. Barzon, C. Del Vecchio, L. Rossi, R. Manganelli, A. Loregian, N. Navarin, D. Abate, M. Sciro, S. Merigliano, E. De Canale, M. C. Vanuzzo, V. Besutti, F. Saluzzo, F. Onelia, M. Pacenti, S. Parisi, G. Carretta, D. Donato, L. Flor, S. Cocchio, G. Masi, A. Sperduti, L. Cattarino, R. Salvador, M. Nicoletti, F. Caldart, G. Castelli, E. Nieddu, B. Labella, L. Fava, M. Drigo, K. A. M. Gaythorpe, A. R. Brazzale, S. Toppo, M. Trevisan, V. Baldo, C. A. Donnelly, N. M. Ferguson, I. Dorigatti, A. Crisanti, Suppression of COVID-19 outbreak in the municipality of Vo, Italy. *Nature* Medline (2020). doi:10.1038/s41586-020-2488-1

13. K. Mizumoto, K. Kagaya, A. Zarebski, G. Chowell, Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro Surveill.* **25**, 2000180 (2020). doi:10.2807/1560-7917.ES.2020.25.10.2000180 Medline

14. Brazilian Ministry of Health, *Painel de Casos de Doença Pelo Coronavírus 2019 (COVID-19) No Brasil Pelo Ministério da Saúde* (2020); http://covid.saude.gov.br.

15. W. M. de Souza, L. Fletcher Buss, D. da Silva Candido, J. P. Carrera, S. Li, A. Zarebski, M. Vincenti-Gonzalez, J. Messina, F. C. da Silva Sales, P. dos Santos Andrade, C. A. Prete Jr., V. H. Nascimento, F. Ghilardi, R. H. Moraes Pereira, A. A. de Souza Santos, L. Abade, B. Gutierrez, M. U. G. Kraemer, R. Santana Aguiar, N. Alexander, P. Mayaud, O. J. Brady, I. O. M. de Souza, N. Gouveia, G. Li, A. Tami, S. Barbosa Oliveira, V. B. Gomes Porto, F. Ganem, W. Ferreira Almeida, F. Fontana Sutile Tardetti Fantinato, E. Marques Macario, W. Kleber Oliveira, O. Pybus, C.-H. Wu, J. Croda, E. Cerdeira Sabino, N. R. Faria, Epidemiological and clinical characteristics of the early phase of the COVID-19 epidemic in Brazil. medRxiv 10.1101/2020.04.25.20077396 [Preprint]. 29 April 2020; https://doi.org/10.1101/2020.04.25.20077396 .

16. J. Croda, W. K. Oliveira, R. L. Frutuoso, L. H. Mandetta, D. C. Baia-da-Silva, J. D. Brito-Sousa, W. M. Monteiro, M. V. G. Lacerda, COVID-19 in Brazil: Advantages of a socialized unified health system and preparation to contain cases. *Rev. Soc. Bras. Med. Trop.* **53**, e20200167 (2020). doi:10.1590/0037-8682-0167-2020 Medline

17. J. Croda, L. Garcia, Immediate health surveillance response to COVID-19 epidemic [in Portuguese]. *Epidemiol. Ser. Saúde* **29**, e2020002 (2020). doi:10.5123/S1679-49742020000100021 Medline

18. S. B. Oliveira, V. Bertollo Gomes Porto, F. Ganem, F. Macedo Mendes, M. Almiron, W. Kleber de Oliveira, F. Fontana Sutile Tardetti Fantinato, W. Aparecida Ferreira de Almeida, A. Pereira de Macedo Borges, H. Natan Batista Pinheiro, R. dos Santos

Oliveira, J. R. Andrews, N. R. Faria, M. Barreto Lopes, W. Araujo, F. A. Diaz-Quijano, H. I. Nakaya, J. Croda, Monitoring social distancing and SARS-CoV-2 transmission in Brazil using cell phone mobility data. medRxiv 2020.04.30.20082172 [Preprint] (5 May 2020); https://doi.org/10.1101/2020.04.30.20082172.

19. S. M. Kissler, Reductions in commuting mobility predict geographic differences in SARS-CoV-2 prevalence in New York City (Harvard DASH Repository, 2020); https://dash.harvard.edu/bitstream/handle/1/42665370/Kissler_etal_NYC_mobility.pdf?sequence = 1&isAllowed = y.

20. H. J. T. Unwin, S. Mishra, V. C. Bradley, A. Gandy, M. Vollmer, T. Mellan, H. Coupland, K. Ainslie, C. Whittaker, J. Ish-Horowicz, S. Filippi, X. Xi, M. Monod, O. Ratmann, M. Hutchinson, F. Valka, H. Zhu, I. Hawryluk, P. Milton, M. Baguelin, A. Boonyasiri, N. Brazeau, L. Cattarino, G. Charles, L. V. Cooper, Z. Cucunuba, G. Cuomo-Dannenburg, B. Djaafara, I. Dorigatti, O. J. Eales, J. Eaton, S. van Elsland, R. FitzJohn, K. Gaythorpe, W. Green, T. Hallett, W. Hinsley, N. Imai, B. Jeffrey, E. Knock, D. Laydon, J. Lees, G. Nedjati-Gilani, P. Nouvellet, L. Okell, A. Ower, K. V. Parag, I. Siveroni, H. A. Thompson, R. Verity, P. Walker, C. Walters, Y. Wang, O. J. Watson, L. Whittles, A. Ghani, N. M. Ferguson, S. Riley, C. A. Donnelly, S. Bhatt, S. Flaxman, *Report 23: State-Level Tracking of COVID-19 in the United States (21-05-2020)* (Imperial College London, 2020); https://doi.org/10.25561/79231.

21. S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, M. Monod, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. C. Vollmer, N. M. Ferguson, L. C. Okell, S. Bhatt; Imperial College COVID-19 Response Team, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* (2020). doi:10.1038/s41586-020-2405-7 Medline

22. T. A. Mellan, H. H. Hoeltgebaum, S. Mishra, C. Whittaker, R. P. Schnekenberg, A. Gandy, H. J. T. Unwin, M. A. C. Vollmer, H. Coupland, I. Hawryluk, N. Rodrigues Faria, J. Vesga, H. Zhu, M. Hutchinson, O. Ratmann, M. Monod, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, N. Brazeau, G. Charles, L. V. Cooper, Z. Cucunuba, G. Cuomo-Dannenburg, A. Dighe, B. Djaafara, J. Eaton, S. L. van Elsland, R. FitzJohn, K. Fraser, K. Gaythorpe, W. Green, S. Hayes, N. Imai, B. Jeffrey, E. Knock, D. Laydon, J. Lees, T. Mangal, A. Mousa, G. Nedjati-Gilani, P. Nouvellet, D. Olivera, K. V. Parag, M. Pickles, H. A. Thompson, R. Verity, C. Walters, H. Wang, Y. Wang, O. J. Watson, L. Whittles, X. Xi, L. Okell, I. Dorigatti, P. Walker, A. Ghani, S. Riley, N. M. Ferguson, C. A. Donnelly, S. Flaxman, S. Bhatt, *Report 21: Estimating COVID-19 Cases and Reproduction Number in Brazil* (2020); https://doi.org/10.25561/78872.

23. Y.-Z. Zhang, E. C. Holmes, Novel 2019 coronavirus genome, Virological (2020); https://virological.org/t/novel-2019-coronavirus-genome/319.

24. V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. W. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, D. G. J. C. Mulders, B. L. Haagmans, B. van der Veer, S. van den Brink, L. Wijsman, G. Goderski, J.-L. Romette, J. Ellis, M. Zambon, M. Peiris, H. Goossens, C. Reusken, M. P. G. Koopmans, C. Drosten, Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill.* **25**,2000045 (2020). doi:10.2807/1560-7917.ES.2020.25.3.2000045 Medline

25. T. Thi Nhu Thao, F. Labroussaa, N. Ebert, P. V'kovski, H. Stalder, J. Portmann, J. Kelly, S. Steiner, M. Holwerda, A. Kratzel, M. Gultom, K. Schmied, L. Laloli, L. Hüsser, M. Wider, S. Pfaender, D. Hirt, V. Cippà, S. Crespo-Pomar, S. Schröder, D. Muth, D. Niemeyer, V. M. Corman, M. A. Müller, C. Drosten, R. Dijkman, J. Jores, V. Thiel, Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. *Nature* **582**, 561–565 (2020). doi:10.1038/s41586-020-2294-9 Medline

26. P. C. Resende, E. Delatorre, T. Gräf, D. Mir, F. do Couto Motta, L. Reis Appolinario, A. C. Dias da Paixão, M. Ogrzewalska, B. Caetano, M. Cordeiro dos Santos, J. de Almeida Ferreira, E. Costa Santos Junior, S. Patroca da Silva, S. Bianchini Fernandes, L. A. Vianna, L. da Costa Souza, J. F. G. Ferro, V. B. Nardy, J. Croda, W. K. Oliveira, A. Abreu, G. Bello, M. M. Siqueira, Genomic surveillance of SARS-CoV-2 reveals community transmission of a major lineage during the early pandemic phase in Brazil. bioRxiv 020.06.17.158006 [Preprint] (2020); https://doi.org/10.1101/2020.06.17.158006.

27. J. Xavier, M. Giovanetti, T. Adelino, V. Fonseca, A. V. Barbosa da Costa, A. Aparecida Ribeiro, K. Nascimento Felicio, C. Guerra Duarte, M. V. Ferreira Silva, A. Salgado, M. Teixeira Lima, R. de Jesus, A. Fabri, C. Franco Soares Zoboli, T. Gutemberg Souza Santos, F. Iani, A. M. Bispo de Filippis, M. Agudo Mendonca Teixeira de Siqueira, A. L. de Abreu, V. de Azevedo, D. Brock Ramalho, C. F. Campelo de Albuquerque, T. de Oliveira, E. C. Holmes, J. Lourenco, L. C. Junior Alcantara, M. Aparecida Assuncao Oliveira, The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. medRxiv 2020.2005.2005.20091611 [Preprint] (2020); https://doi.org/10.1101/2020.05.05.20091611.

28. V. A. Nascimento, A. L. G. Corado, F. O. Nascimento, A. K. A. Costa, D. C. G. Duarte, M. S. Jesus, S. L. B. Luz, L. M. F. Goncalves, C. F. Costa, E. Delatorre, F. G. Naveca, Genomic and phylogenetic characterization of an imported case of SARS-CoV-2 in Amazonas State, Brazil. *Memoirs of the Oswaldo Cruz Institute* 10.1590/0074-02760200310 (2020).

29. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro. Surveill.* **22**, 30494 (2017) doi:10.2807/1560-7917.ES.2017.22.13.30494 Medline

30. M. Cotten, S. J. Watson, P. Kellam, A. A. Al-Rabeeah, H. Q. Makhdoom, A. Assiri, J. A. Al-Tawfiq, R. F. Alhakeem, H. Madani, F. A. AlRabiah, S. Al Hajjar, W. N. Al-nassir, A. Albarrak, H. Flemban, H. H. Balkhy, S. Alsubaie, A. L. Palser, A. Gall, R. Bashford-Rogers, A. Rambaut, A. I. Zumla, Z. A. Memish, Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: A descriptive genomic study. *Lancet* **382**, 1993–2002 (2013). doi:10.1016/S0140-6736(13)61887-5 Medline

31. M. Cotten, S. J. Watson, A. I. Zumla, H. Q. Makhdoom, A. L. Palser, S. H. Ong, A. A. Al Rabeeah, R. F. Alhakeem, A. Assiri, J. A. Al-Tawfiq, A. Albarrak, M. Barry, A. Shibl, F. A. Alrabiah, S. Hajjar, H. H. Balkhy, H. Flemban, A. Rambaut, P. Kellam, Z. A. Memish, Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *mBio* **5**, e01062-13 (2014). doi:10.1128/mBio.01062-13 Medline

32. G. Dudas, L. M. Carvalho, A. Rambaut, T. Bedford, MERS-CoV spillover at the camel-human interface. *eLife* **7**, e31257 (2018). doi:10.7554/eLife.31257 Medline

33. Z. Zhao, H. Li, X. Wu, Y. Zhong, K. Zhang, Y.-P. Zhang, E. Boerwinkle, Y.-X. Fu, Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol. Biol.* **4**, 21 (2004). doi:10.1186/1471-2148-4-21 Medline

34. S. Duchene, L. Featherstone, M. Haritopoulou-Sinanidou, A. Rambaut, P. Lemey, G. Baele, Temporal signal and the phylodynamic threshold of SARS-CoV-2. bioRxiv 2020.05.04.077735 [Preprint] (2020); https://doi.org/10.1101/2020.05.04.077735.

35. J. Lu, L. du Plessis, Z. Liu, V. Hill, M. Kang, H. Lin, J. Sun, S. François, M. U. G. Kraemer, N. R. Faria, J. T. McCrone, J. Peng, Q. Xiong, R. Yuan, L. Zeng, P. Zhou, C. Liang, L. Yi, J. Liu, J. Xiao, J. Hu, T. Liu, W. Ma, W. Li, J. Su, H. Zheng, B. Peng, S. Fang, W. Su, K. Li, R. Sun, R. Bai, X. Tang, M. Liang, J. Quick, T. Song, A. Rambaut, N. Loman, J. Raghwani, O. G. Pybus, C. Ke, Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997–1003.e9 (2020). doi:10.1016/j.cell.2020.04.023 Medline

36. D. D. S. Candido, A. Watts, L. Abade, M. U. G. Kraemer, O. G. Pybus, J. Croda, W. de Oliveira, K. Khan, E. C. Sabino, N. R. Faria, Routes for COVID-19 importation in Brazil. *J. Travel Med.* **27**, taaa042 (2020). doi:10.1093/jtm/taaa042 Medline

37. S. Dellicour, K. Durkin, S. L. Hong, B. Vanmechelen, J. Martí-Carreras, M. S. Gill, C. Meex, S. Bontems, E. André, M. Gilbert, C. Walker, N. De Maio, N. R. Faria, J. Hadfield, M.-P. Hayette, V. Bours, T. Wawina-Bokalanga, M. Artesi, G. Baele, P. Maes, A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. bioRxiv 2020.05.05.078758 [Preprint] (2020); https://doi.org/10.1101/2020.05.05.078758.

38. World Health Organization, Coronavirus disease 2019 (COVID-19): Situation report –72 (WHO, 2020); https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200401-sitrep-72-covid-19.pdf?sfvrsn = 3dd8971b_2 .

39. Centre for Genomic Pathogen Surveillance, Imperial College London, Report of 427 novel genomes from Brazil and the associated metadata, Microreact (2020); https://microreact.org/project/rKjKLMrjdPVHKR1erUzKyi.

40. Data and code for: D. S. Candido, I. M. Claro, J. G. de Jesus, W. M. Souza, F. R. R. Moreira, S. Dellicour, T. A. Mellan, L. du Plessis, R. H. M. Pereira, F. C. S. Sales, E. R. Manuli, J. Thézé, L. Almeida, M. T. Menezes, C. M. Voloch, M. J. Fumagalli, T. M. Coletti, C. A. M. da Silva, M. S. Ramundo, M. R. Amorim, H. Hoeltgebaum, S. Mishra, M. S. Gill, L. M. Carvalho, L. F. Buss, C. A. Prete Jr., J. Ashworth, H. I. Nakaya, P. S. Peixoto, O. J. Brady, S. M. Nicholls, A. Tanuri, Á. D. Rossi, C. K. V. Braga, A. L. Gerber, A. P. de C. Guimarães, N. Gaburo Jr., C. Salete Alencar, A. C. S. Ferreira, C. X. Lima, J. E. Levi, C. Granato, G. M. Ferreira, R. S. Francisco Jr., F. Granja, M. T. Garcia, M. L. Moretti, M. W. Perroud Jr., T. M. P. P. Castiñeiras, C. S. Lazari, S. C. Hill, A. A. de Souza Santos, C. L. Simeoni, J. Forato, A. C. Sposito, A. Z. Schreiber, M. N. N. Santos, C. Zolini de Sá, R. P. Souza, L. C. Resende-Moreira, M. M. Teixeira, J. Hubner, P. A. F. Leme, R. G. Moreira, M. L. Nogueira, CADDE-Genomic-Network, N. M. Ferguson, S. F. Costa, J. L. Proenca-Modena, A. T. R. Vasconcelos, S. Bhatt, P. Lemey, C.-H. Wu, A. Rambaut, N.

J. Loman, R. S. Aguiar, O. G. Pybus, E. C. Sabino, N. Rodrigues Faria, Evolution and epidemic spread of SARS-CoV-2 in Brazil, Dryad (2020); https://doi.org/10.5061/dryad.rxwdbrv5z.

41. A. Aktay, S. Bavadekar, G. Cossoul, J. Davis, D. Desfontaines, A. Fabrikant, E. Gabrilovich, K. Gadepalli, B. Gipson, M. Guevara, C. Kamath, M. Kansal, A. Lange, C. Mandayam, A. Oplinger, C. Pluntke, T. Roessler, A. Schlosberg, T. Shekel, S. Vispute, M. Vu, G. Wellenius, B. Williams, R. J. Wilson, Google COVID-19 community mobility reports: Anonymization process description (version 1.0). arXiv:2004.04145 [cs.CR] (8 April 2020).

42. inloco, Mapa Brasileiro da COVID-19 (2020); https://mapabrasileirodacovid.inloco.com.br/pt/.

43. P. S. Peixoto, D. Marcondes, C. Peixoto, S. M. Oliva, Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil. *PLoS One* **15**, e0235732 (2020). doi:10.1371/journal.pone.0235732 Medline

44. Y. Liu, A. A. Gayle, A. Wilder-Smith, J. Rocklöv, The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **27**, taaa021 (2020). doi:10.1093/jtm/taaa021 Medline

45. J. J. Waggoner, V. Stittleburg, R. Pond, Y. Saklawi, M. K. Sahoo, A. Babiker, L. Hussaini, C. S. Kraft, B. A. Pinsky, E. J. Anderson, N. Rouphael, Triplex real-time RT-PCR for severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.* **26**, 1633–1635 (2020). doi:10.3201/eid2607.201285 Medline

46. J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. de Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara Jr, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, N. J. Loman, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017). doi:10.1038/nprot.2017.066 Medline

47. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

48. I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, D. Marshall, Tablet: Next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010). doi:10.1093/bioinformatics/btp666 Medline

49. K. Katoh, D. M. Standley, MAFFT: Iterative refinement and additional methods. *Methods Mol. Biol.* **1079**, 131–146 (2014). doi:10.1007/978-1-62703-646-7_8 Medline

50. M. Hasegawa, H. Kishino, T. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985). doi:10.1007/BF02101694 Medline

51. Z. Yang, Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994). doi:10.1007/BF00160154 Medline

52. 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**, 1530–1534 (2020). doi:10.1093/molbev/msaa015 Medline

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

54. J. Singer, R. Gifford, M. Cotten, D. Robertson, CoV-GLUE: A web application for tracking SARS-CoV-2 genomic variation (2020); https://doi.org/10.20944/preprints202006.0225.v1.

55. J. G. Jesus, C. Sacchi, D. D. S. Candido, I. M. Claro, F. C. S. Sales, E. R. Manuli, D. B. B. D. Silva, T. M. Paiva, M. A. B. Pinho, K. C. O. Santos, S. C. Hill, R. S. Aguiar, F. Romero, F. C. P. D. Santos, C. R. Gonçalves, M. D. C. Timenetsky, J. Quick, J. H. R. Croda, W. Oliveira, A. Rambaut, O. G. Pybus, N. J. Loman, E. C. Sabino, N. R. Faria, Importation and early local transmission of COVID-19 in Brazil, 2020. *Rev. Inst. Med. Trop. São Paulo* **62**, e30 (2020). doi:10.1590/s1678-9946202062030 Medline

56. Centers for Disease Control and Prevention, *Research Use Only 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Primers and Probes* (2020); https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html.

57. J. Northill, I. Mackay, Wuhan coronavirus (2019-nCoV) real-time RT-PCR N gene 2020 (Wuhan-N) V.1 (2020); https://www.protocols.io/view/wuhan-coronavirus-2019-ncov-real-time-rt-pcr-n-gen-ba86ihze?version_warning = no.

58. N. Nao, Shirato, K., Katano, H., Matsuyama, S., Takeda, M., Detection of second case of 2019-nCoV infection in Japan (corrected version) (2020); https://www.niid.go.jp/niid/images/vir3/nCoV/method-niid-20200123-2_erratum.pdf.

59. Thailand Ministry of Public Health, Diagnostic detection of novel coronavirus 2019 by real time RT-PCR (2020); https://www.who.int/docs/default-source/coronaviruse/conventional-rt-pcr-followed-by-sequencing-for-detection-of-ncov-rirl-nat-inst-health-t.pdf?sfvrsn = 42271c6d_4.

60. Chinese National Institute for Viral Disease Control and Prevention, Specific primers and probes for detection 2019 novel coronavirus (2020); http://ivdc.chinacdc.cn/kyjz/202001/t20200121_211337.html.

61. HKU Med, LKS Faculty of Medicine, School of Public Health, Detection of, 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR (2020); https://www.who.int/docs/default-source/coronaviruse/peiris-protocol-16-1-20.pdf?sfvrsn = af1aac73_4.

62. T. C. Bruen, H. Philippe, D. Bryant, A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006). doi:10.1534/genetics.105.048975 Medline

63. D. H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006). doi:10.1093/molbev/msj030 Medline

64. D. P. Martin, B. Murrell, M. Golden, A. Khoosal, B. Muhire, RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015). doi:10.1093/ve/vev003 Medline

65. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018). doi:10.1093/ve/vey016 Medline

66. D. L. Ayres, A. Darling, D. J. Zwickl, P. Beerli, M. T. Holder, P. O. Lewis, J. P. Huelsenbeck, F. Ronquist, D. L. Swofford, M. P. Cummings, A. Rambaut, M. A. Suchard, BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012). doi:10.1093/sysbio/syr100 Medline

67. M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, M. A. Suchard, Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013). doi:10.1093/molbev/mss265 Medline

68. M. A. R. Ferreira, M. A. Suchard, Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.* **36**, 355–368 (2008). doi:10.1002/cjs.5550360302

69. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018). doi:10.1093/sysbio/syy032 Medline

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

71. N. R. Faria, A. Rambaut, M. A. Suchard, G. Baele, T. Bedford, M. J. Ward, A. J. Tatem, J. D. Sousa, N. Arinaminpathy, J. Pépin, D. Posada, M. Peeters, O. G. Pybus, P. Lemey, HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56–61 (2014). doi:10.1126/science.1256739 Medline

72. J. D. O'Brien, V. N. Minin, M. A. Suchard, Learning to count: Robust estimates for labeled distances between molecular sequences. *Mol. Biol. Evol.* **26**, 801–814 (2009). doi:10.1093/molbev/msp003 Medline

73. V. N. Minin, M. A. Suchard, Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3985–3995 (2008). doi:10.1098/rstb.2008.0176 Medline

74. V. N. Minin, M. A. Suchard, Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56**, 391–412 (2008). doi:10.1007/s00285-007-0120-8 Medline

75. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010). doi:10.1093/molbev/msq067 Medline

76. S. Dellicour, G. Baele, G. Dudas, N. R. Faria, O. G. Pybus, M. A. Suchard, A. Rambaut, P. Lemey, Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nat. Commun.* **9**, 2222 (2018). doi:10.1038/s41467-018-03763-2 Medline

77. S. Dellicour, R. Rose, N. R. Faria, P. Lemey, O. G. Pybus, SERAPHIM: Studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016). doi:10.1093/bioinformatics/btw384 Medline

science.sciencemag.org/cgi/content/full/science.abh2644/DC1

Supplementary Material for

Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil

Nuno R. Faria*, Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan da S. Candido, Swapnil Mishra, Myuki A. E. Crispim, Flavia C. S. Sales, Iwona Hawryluk, John T. McCrone, Ruben J. G. Hulswit, Lucas A. M. Franco, Mariana S. Ramundo, Jaqueline G. de Jesus, Pamela S. Andrade, Thais M. Coletti, Giulia M. Ferreira, Camila A. M. Silva, Erika R. Manuli, Rafael H. M. Pereira, Pedro S. Peixoto, Moritz U. G. Kraemer, Nelson Gaburo Jr, Cecilia da C. Camilo, Henrique Hoeltgebaum, William M. Souza, Esmenia C. Rocha, Leandro M. de Souza, Mariana C. de Pinho, Leonardo J. T Araujo, Frederico S. V. Malta, Aline B. de Lima, Joice do P. Silva, Danielle A. G. Zauli, Alessandro C. de S. Ferreira, Ricardo P Schnekenberg, Daniel J. Laydon, Patrick G. T. Walker, Hannah M. Schlüter, Ana L. P. dos Santos, Maria S. Vidal, Valentina S. Del Caro, Rosinaldo M. F. Filho, Helem M. dos Santos, Renato S. Aguiar, José L. Proença-Modena, Bruce Nelson, James A. Hay, Mélodie Monod, Xenia Miscouridou, Helen Coupland, Raphael Sonabend, Michaela Vollmer, Axel Gandy, Carlos A. Prete Jr., Vitor H. Nascimento, Marc A. Suchard, Thomas A. Bowden, Sergei L. K. Pond, Chieh-Hsi Wu, Oliver Ratmann, Neil M. Ferguson, Christopher Dye, Nick J. Loman, Philippe Lemey, Andrew Rambaut, Nelson A. Fraiji, Maria do P. S. S. Carvalho, Oliver G. Pybus, Seth Flaxman, Samir Bhatt*, Ester C. Sabino*

*Corresponding author. Email: n.faria@imperial.ac.uk (N.R.F.); samir.bhatt@sund.ku.dk (S.B.); sabinoec@usp.br (E.S.C.)

Published 14 April 2021 as *Science* First Release
DOI: 10.1126/science.abh2644

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/science.abh2644/DC1)

MDAR Reproducibility Checklist
Data Files S1 to S6 as separate .csv files

Materials and Methods

Ethics

Residual oropharyngeal and nasal swab collections from Manaus residents testing positive for SARS-CoV-2 RT-qPCR between 1 November 2020 and 9 January 2021 were obtained from two private clinical laboratories in Manaus. Metadata associated with positive SARS-CoV-2 RT-qPCR results in Manaus residents testing between 1 July 2020 and 15 January 2021 were obtained from a third private clinical laboratory in Manaus. All samples were de-identified before receipt by the researchers. Ethical approval for this study was confirmed by the national ethical review board (Comissão Nacional de Ética em Pesquisa), protocol number CAAE 30127020.0.0000.0068.

Sampling and Metadata Collection

A total of 436 SARS-CoV-2 samples RT-qPCR confirmed or suspected were collected for genomic sequencing between 1 November 2020 and 9 January 2021. Samples were provided for confirmatory testing and genome sequencing. For clinical laboratory A (*n*=37 RT-qPCR positive cases with sampling dates from 15 to 23 December 2020), SARS-CoV-2 diagnosis was performed using the Allplex 2019-nCoV Assay (Seegene, South Korea) assay that detects the RNA-dependent RNA polymerase (RdRP), nucleocapsid (N) specific genes for SARS-CoV-2 and the E gene for all Sarbecovirus subgenus, including SARS-CoV-2 (*70, 71*). For clinical laboratory B, 399 RT-qPCR positive or suspected cases [representing 73% of all 548 samples with a RT-qPCR positive (*n*=545) or inconclusive (*n*=3) results between 2 November and 9 January 2021] were processed for genome sequencing. In this case, RT-qPCR was determined using the Xpert Xpress SARS-CoV-2 platform (GeneXpert) that detects the N viral target specific for SARS-CoV-2 and the E viral target of *Sarbecovirus* subgenus including SARS-CoV-2 (Cepheid, USA). Samples were shipped in dry ice to the Institute of Tropical Medicine, University of São Paulo, Brazil, for genome sequencing. RT-qPCR cycle threshold values and associated metadata (patient age and sex, date of onset symptom, date of RT-qPCR test, data of sample collection when available, cycle threshold, Ct, values for E and N viral targets) were recorded for 1,084 RT-qPCR positive and 16 inconclusive results from laboratory B in Manaus between 18 May 2020 and 27 January 2021 (**Data S1**). Line-list metadata (patient age and sex, date sampling collection, cycle threshold values) from a third clinical in Manaus (laboratory C) was obtained for RT-qPCR positive samples tested using the TaqPath COVD-19 Combo kit (ThermoFisher-Applied Biosystems, United Kingdom) that detects the N, S, and ORF1ab viral targets (**Data S3**).

PCR Amplification and Virus Nanopore Sequencing

Viral RNA was isolated from 200-µl SARS-CoV-2-suspected samples using the QIAamp Viral RNA Mini kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. Virus genome sequencing was carried out on all positive samples regardless of laboratory reported RT-qPCR cycle threshold values using a combination of targeted multiplex-PCR amplification and portable nanopore sequencing MinION platform (Oxford Nanopore Technologies, ONT, UK).

cDNA synthesis was performed from the extracted RNA using random hexamers, and the Protoscript II First Strand cDNA synthesis Kit (New England Biolabs, UK). Subsequently, the ARTIC network SARS-CoV-2 V3 primer scheme and Q5 High-Fidelity DNA polymerase (New England Biolabs, UK) were used for SARS-CoV-2 whole-genome multiplex-PCR amplification (*24*). AmpureXP beads (Beckman Coulter, United Kingdom) were used for PCR product purification and fluorimetry-based quantification was carried out using the Qubit dsDNA High Sensitivity assay on the Qubit 3.0 (Life Technologies, USA).

To ensure uniform sequencing of samples, equimolar normalisation of 10 ng per sample was performed followed by barcoding using the EXP-NBD104 (1–12) and EXP-NBD114 (13–24) Native Barcoding Kits (Oxford Nanopore Technologies, UK). Finally, barcoded samples were pooled followed by library preparation using the SQK-LSK109 Kit (Oxford Nanopore Technologies, UK).

Nanopore sequencing libraries were loaded onto an R9.4.1 flow-cell (Oxford Nanopore Technologies, UK) and sequenced using MinKNOW version 20.10.3 (Oxford Nanopore Technologies, UK). FAST5 files containing the raw signal data were basecalled, demultiplexed, and trimmed using Guppy v4.4.1 (Oxford Nanopore Technologies, UK). The process reads were aligned against the reference genome Wuhan-Hu-1 (GenBank: MN908947.3) using minimap2 v2.17.r941 and converted to a sorted BAM file using SAMtools (*72*). Length filtering, quality test and primmer trimming was performed for each barcode using artic guppyplex and variant calling and consensus sequences using artic minion with Nanopolish and Medaka versions from ARTIC bioinformatics pipeline (https://github.com/artic-network/fieldbioinformatics). Genome regions with a depth of <20-fold were represented with N characters. The genome statistics were obtained from SAMtools and the Tablet viewer (*73*). Any runs suspected to have any level of contamination were discarded. We analysed intra-host sequencing data by reference to variant allele frequency measurements at P.1 lineage-defining positions of the genome by reference to the underlying sequence read alignment files. Lineage-defining mutations (**table S1**) were highly stable across the genomes (**fig. S15**). Limited evidence of mixed infections was observed, with only one genome demonstrating coverage patterns suggestive of mixed infection (sample CD1721). Recombination is unlikely but difficult to formally exclude with existing datasets. Individual nanopore sequencing statistics for each sequence generated in this study can be found in **Data S2**.

Genome Datasets

Genome coverage of 184 generated sequences obtained from clinical samples varied from 27 to 99% of the virus SARS-CoV-2 genome. Of these, 35 sequences (average Ct of 21.78, range 14.5–29.2) had a virus genome coverage between 25–75%. However, even partial sequences can provide important information about changes in SARS-CoV-2 lineage structure (*25*).

We compiled three genome datasets from Manaus from data generated in this study: *dataset A* included 184 sequences with >25% virus genome coverage (37 from laboratory A and 147 from laboratory B) and was used to estimate virus lineage frequency in Manaus over time; *dataset B* included 143 near-complete genome sequences with >75% of the virus genome coverage (31 from laboratory A, 112 from laboratory B); and *dataset C* included 48 sequences with >95% of the virus genome complete (*n*=17 from laboratory A, *n*=31 from laboratory B).

For *datasets A, B and C*, a reference genome sequence Wuhan-Hu-1 (GenBank: MN908947.3) was appended before multiple sequence alignment using MAFFTv.7 (*74*). For *dataset A*, lineage classification was conducted using manual phylogenetic analysis. Sequences with genome coverage between 25% and 75% were appended to *dataset C* and assigned to B.1.1.28, P.2 and P.1 lineages based on monophyletic clustering of each sequence within each of these lineages (**fig. S2**). Manual phylogenetic subtyping and PANGO lineage classification using the latest pangolin version (v.2.2.1, 6 February 2021) (*26;* http://pangolin.cog-uk.io/) was conducted for *dataset B* and *dataset C* **(figs. S3-S5)**. Date of sample collection, age, sex, RT-qPCR CT values, lineage assignment and sequencing statistics for 184 sequences generated in this study from Manaus can be found in **Data S2**.

We downloaded all sequences publicly available in GISAID (up to 14-01-2021) and selected for analysis those that were published in *PubMed*, *MedRxiv*, *BioRxiv* or *Preprint* repositories. Specifically, *dataset B* and *dataset C* from Manaus were appended to (*i*) B1.1.28 genome sequences with >95% virus genome coverage from the Brazilian Amazon (*75*), Minas Gerais (*76*), Pernambuco (*77*), Rio de Janeiro (*78,79*), Rio Grande do Sul (*80*) and to data from an early country-wide study of SARS-CoV-2 diversity in Brazil (*19*). *Datasets B* and *C* were also appended to (*ii*) P.2 genome sequences with >95% virus genome coverage from Rio de Janeiro (*79*), Rio Grande do Sul (*80, 81*). Exceptionally, written permission was obtained from the reference laboratory in São Paulo, Institute Adolfo Lutz, to use P.1 complete genome sequences shared in GISAID (up to 19-01-2021) as part of their surveillance activities. Duplicate sequences were removed from the alignments, 5' and 3' untranslated regions from each genome were discarded. Sequence CD1721 (EPI_ISL_1060918) was identified as potential mixed infection (**Fig. S15**) and removed from phylogenetic analysis. A table describing GISAID IDs, authors, originating and submitting laboratory for all publicly available data used in *dataset B'* and *dataset C'* (with data from this study and publicly available data) can be found in **Data S4**.

Maximum Likelihood Tree Reconstruction

Fast and efficient maximum likelihood (ML) phylogenetic trees were reconstructed using IQTREE 2 (*82*) for *dataset A'* (*n*=988) (used for lineage classification of >25 and <75% virus genome coverage samples from Manaus, **Data S2**), and for the B.1.28, P.1 and P.2-specific *dataset B'* and *dataset C'* (ML trees can be found in **figs. S2-S4**). Briefly, a Jukes Cantor (*83*) DNA substitution model assuming equal substitution rates and equal base frequencies. Near zero branches were collapsed so that the final tree could be multifurcating to account for the many polytomies observed in SARS-CoV-2 phylogenetic trees. To explore temporal structure of *dataset B'* and *dataset C'*, root-to-tip genetic distances (*d*) were regressed against sampling dates (yyyy-mm-dd) using TempEst v.1.5.3 (*29*). Two sequences showed incongruent genetic diversity compared with its sampling date (SP-322 EPI_ISL_693197 and AM-1061 EPI_ISL_940616) and were discarded from subsequent analyses. Strong and identical correlations between *d* and sequence sampling dates were observed for *dataset B'* (*n*=962, r^2=0.82) (**fig. S6)** and *dataset C'* (*n*=871, r^2=0.81) (**fig. S7**). Therefore, we used *dataset B'* for subsequent phylogenetic analyses. Regressions between root-to-tip distances and sampling dates in *dataset B'* were also fit separately for B.1.1.28, P.2, and P.1 lineages in R v.3.6.2 (*84*). These showed no obvious difference in evolutionary rates and suggested that P.2 and P.1 show a tendency to fall over the regression line of B.1.1.28 (**fig. S8**). This supports an increased evolutionary rate on the ancestral branches leading to P.1 and P.2 compared to B.1.1.28 evolutionary rate, an evolutionary scenario that seems to be characteristic of SARS-CoV-2 lineages of concern (*13*).

Bayesian Coalescent Inference

Next, we used a fully probabilistic Bayesian framework to reconstruct molecular clock phylogenies and estimate growth rates directly from time-stamped genome sequence data. Substitution rates were modelled according to a HKY with 4 gamma categories to account for among site rate variation (*84, 86*). The B.1.1.28 lineage has been circulating in Brazil since late February-early March (*19*). P.1 and P.2 lineages are phylogenetically nested within the more diverse and older B.1.1.28 strains and form separate monophyletic clades (see **figs. S2-S5**).

To account for rate differences in ancestral branches leading to P.1 and P.2, we estimate molecular clock trees using a local clock implementation in a Bayesian framework (*32*) that explicitly allows for distinct evolutionary rates on the ancestral branch leading to P.1 and P.2.

Local clock models allow different rates along distinct lineages of a single phylogeny (*32*), which may be particularly suitable to estimate dates of emergence of P.1 and P.2 because they can take into account higher rates of mutation accumulation over short periods of time that could to be linked to selective pressures associated with the emergence of lineages of concern (*13*). Five independent analyses were performed using a flexible non-parametric skygrid tree prior (*33*) for 200 million MCMC steps using BEAST version 1.10.4 (*87*) with BEAGLE library v3.1.0 for accelerated likelihood evaluation (*88*). Parameters and trees were sampled every 25,000 steps and convergence of MCMC chains was inspected with Tracer v1.7.1 (*89*). Posterior probability distributions for the most recent common ancestor of the P.1 and P.2 clade are shown in **fig. S9**.

We also used a nested coalescent model (*90*) to estimate viral growth rates for B.1.1.28, P.1 and P.2. We fit a constant demographic model to the phylogeny excluding the P.1 and P.2 clades. Two separate logistic growth models were then fitted to P.1 and P.2 clades. For the logistic growth coalescent models, a lognormal prior with mean 1.0 and a standard deviation of 10.0 were used for the population size; for the growth rate, we used a Laplace prior with a mean of 0 and a scale of 10. For the local clock model, we use a normal prior with mean -7.0 and standard deviation of 5.0 on the background rate in log space and a normal prior with mean 0 and standard deviation of 0.5 for the log effect sizes on the branches for which the rates are allowed to deviate from the background rate (*91*). All other priors used for phylogenetic inference were kept at default settings. Posterior probability distributions for the most recent common ancestor of the P.1 and P.2 clade according to the constant-logistic-logistic model and corresponding growth rate and doubling time parameters are show as **fig. S9** and **fig. S10**.

To quantify the support for both the rate differences in the local clock model and the growth rates in the nested coalescent model, we conduct Bayes Factor (BF) tests. For this purpose, we employ posterior indicator functions that allow estimating the posterior probability that a specific substitution rate (the rate on the branch ancestral to P.1 or P.2) is larger than another rate (the background rate for B.1.1.28) and that one growth rate (for P.1) is larger than another growth rate (for P.2). We use these posterior probabilities to calculate the BF as the ratio of the posterior odds over the prior odds that substitution rates or growth rates are different, assuming that the prior probabilities for these differences are 0.5 (in line with our prior specification on these parameters).

Adaptive Evolution of P.1 lineage

We investigated the extent of selective forces acting on P.1 and P.2 SARS-CoV-2 lineages using HyPhy v2.5.27 (*48*). We analyzed a median of 5 unique P.1 haplotypes per gene/peptide in the context of a median of 79 reference sequences, and a median of 9 unique P.2 haplotypes per gene/peptide in the context of a median of 70 reference sequences. The summary of P.1 lineage-defining sites subject to episodic diversifying selection ($p \leq 0.05$) identified using MEME (*92*) is shown in **table S1**. In addition to individual sites under selection, we also recorded instances of putative convergence, i.e., substitutions to the same amino-acid at the same site in both lineages; there were only 2 such events (S/484K and ORF1A/318L). The evolutionary "credibility" of target residues was estimated using the PRIME method (*92*) based on a bat/pangolin Sarbecovirus alignment (*93*) and results can be visualized at https://hackmd.io/7hFvRdJdSVSONv_wW40_Rg.

Structural Analysis of P.1 lineage

Lineage defining mutations were mapped onto a previously reported cryoEM structure of the cleaved trimeric SARS-CoV-2 S ectodomain [PDB: 6ZGI, (*94*)] with PyMOL v 2.4.0

(*95*) (**fig. S14**). Substituted residues are indicated as spheres and coloured by type of selection according to MEME support **table S1**). Similarly, SARS-CoV-2 S RBD-hACE2 contact residues were mapped as observed in the RBD-hACE2 complex crystal structure [PDB: 6MOJ, (*11*)] together with RBD-resident substitutions specific to the P.1 lineage (**fig. S14**). N-linked glycans are omitted for clarity.

Air Travel and Mobile Geolocation Data

To better contextualize the spread of P.1 lineage within Brazil, we investigate two different mobility data sources. First, we analysed monthly air passenger travel data produced by Brazil's Civil Aviation Agency (ANAC) which is publicly available at https://www.anac.gov.br/assuntos/setorregulado/empresas/envio-de-informacoes/base-de-dados-estatisticos-do-transporte-aereo. This includes the number of passengers and connections for international flights to and from Brazil, as well as domestic flights within the country. Using this data, we calculated the total number of passengers who travelled from Manaus between November and December 2020 disaggregated by state of origin and destination (**Fig. 1D, fig. S10**).

State-level mobility where the origin of the trip was the municipality of Manaus were calculated from approximately 5 million trips aggregated from anonymized cell phone data users in the month of November 2020 (*96*) (**fig. S10**). Data was obtained from In Loco (mapabrasileirodacovid.inloco.com.br), a company that provides geolocation services for a broad range of mobile applications and covers ~20% of the mobile devices in the country. Anonymized cell-phone shows a similar pattern to air travel data but shows travel from Manaus to other municipalities in the Amazonas states being even more important than with flights. Numbers of recorded state-level movements from and to Amazonas state, as well as city-level movements from and to Manaus municipality, are available as **Data S5** and **Data S6**, respectively.

Logistic Function Fitting to P.1 Genome Fraction

We fitted a logistic function to the time-varying fraction of sequenced genomes belonging to P.1 from a single laboratory (laboratory B), binned according to the week sampling had occurred in. The form of the logistic function is as follows:

$$f(t) = \frac{L}{1 + e^{-k(t-t2)+\sigma_i}} \quad (1)$$

where L = maximum value of the logistic function, k = logistic growth rate, t = time since P.1's emergence,$t2$ is the (inferred) time at which half of the genomes in sequenced cases belong to P.1 and σ_i is independently and identically distributed gaussian noise added to account for overdispersion. Model fitting was carried out using a Bayesian framework, written in the probabilistic programming language STAN and implemented in the statistical software R (Version 4.0.2) using the package *rStan* (Version 2.19.03). Three chains of 5000 iterations each were run, with the 1st 2500 samples from each discarded as burn-in and the remaining 2500 (for each chain, 7500 samples total) retained for inference.

We note that the proportion of P.1 cases used for epidemiological modelling was derived from clinical samples obtained from a single laboratory which used the same RT-PCR assay over the course of the pandemic, as well as consistent methods of specimen collection, handling of sample and test interpretation. We also note that sequencing was attempted in all clinical samples regardless of cycle threshold values to ensure further minimal impact in selection biases when assessing proportion of P.1 cases during the study period. Nevertheless, we cannot exclude the possibility that the representativity of samples used here might have

changed through time as no random population surveillance was being attempted during the study period under investigation.

Description of the Epidemiological Model

We utilised a Bayesian semi-mechanistic model of SARS-CoV-2 spread and mortality based on a renewal-process equation (*40, 41, 97*) and extended to include i) multiple SARS-CoV-2 lineages introduced at different points in time; ii) the possibility for these different lineages to possess distinct epidemiological characteristics (such as severity, transmissibility and immune evasion); and iii) waning of natural immunity due to prior infection - parameterised from the results of the recent Public Health England SIREN study - a longitudinal cohort study tracking (re)infection in healthcare workers in the United Kingdom (*43*).

Here, we modeled two different SARS-CoV-2 lineages, hereafter referred to as P.1 and non-P.1, with the timing of P.1's emergence based on the phylogenetic analyses described previously. A full mathematical description of the model and its associated parameters are available in **Supplementary Text.** We ran a number of model scenarios in order to evaluate support for P.1 possessing distinct epidemiological characteristics, specifically running multiple models that varied in their assumption surrounding timing of P.1 emergence.

Estimates of tMRCA are built from genetic sequence data that is not population representative, but rather are estimated based on a limited subset of sequences and therefore are expected to suffer from systematic biases. For this reason, it is important to note that tMRCA and the date of first infection are, in general, not necessarily expected to be exactly the same and we would expect the date of first infection to be before tMRCA. For the results of the epidemiological modelling presented in the main text we assume the mean estimate of tMRCA (9th November 2020) and the date of first infection are the same. We acknowledge that this assumption is likely to change given more representative data. To address these limitations, we have conducted a comprehensive sensitivity analysis (summarised in **table S5**) where we vary tMRCA to cover the entire 95% BCI of the estimated tMRCA distribution. This sensitivity analysis showed that our main conclusions around altered epidemiological characteristics were robust to assumptions regarding the timing of P.1 emergence.

We also varied our assumptions surrounding the duration of protective immunity following infection (i.e. the rate at which natural immunity elicited by prior SARS-CoV-2 infection declines).

The model was fitted to two sources of data. Mortality data from the SIVEP-Gripe (*Sistema de Informação de Vigilância Epidemiológica da Gripe*) SARI (severe acute respiratory infections) hospitalisation database (*67*, *69*), including both class 4 and 5 death records (corresponding to confirmed and suspected COVID-19 deaths), consistent with earlier analyses (*2*) and corrected for known delays in mortality reporting using a Gaussian process nowcasting based framework (*44, 45*). In addition to COVID-19 mortality data, we also integrate genomic data from the sequenced samples, fitting the model to the fraction of sequenced genomes each week that belong to P.1 described in **table S4**.

Model fitting was carried out using a Bayesian framework, written in the probabilistic programming language STAN and implemented in the statistical software R (Version 4.0.2) using the package *rStan* (Version 2.19.03). Hamiltonian Monte Carlo with 3 chains of 1000 iterations each were ran, with half the samples discarded as burnin and the remaining retained for inference. In every instance chains mixing was satisfactory, with traditional rhat statistics (for assessing convergence) less than 1.02. All code used for inference and plotting is available at https://github.com/CADDE-CENTRE.

Data sharing and code availability

Preliminary genome sequences generated from samples obtained from laboratory A were shared on GISAID on 12 January 2021. Findings were shared with representatives from the World Health Organization, Pan American Health Organization, Secretary of Health Amazonas, and FioCruz Manaus on 11 January 2021. Preliminary report describing first P.1 genomes from Manaus was shared on 12 January 2021 (*27*). Epidemiological data and epidemiological model code, together with BEAST XML files, tree files, log files are archived at https://github.com/CADDE-CENTRE and Zenodo (DOI: https://zenodo.org/record/4676853). GISAID IDs for the SARS-CoV-2 Manaus sequences (>50% virus genome coverage) can be found in **Data S1**. All consensus sequences generated by this study can be found at https://github.com/CADDE-CENTRE.

Supplementary Text

Epidemiological Model

This work builds on a previously published mathematical model of SARS-CoV-2 transmission introduced in Flaxman et al, 2020 (*41*). Specifically, we extend this semi-mechanistic Bayesian model to include multiple SARS-CoV-2 strains and the possibility for these strains to possess distinct, strain-specific epidemiological characteristics (such as transmissibility, ability to evade prior immunity, and severity of COVID-19 disease elicited).

Although any number of strains are possible within a framework of this type, we consider only two strains here, defined as $s \in \{1,2\}$. For strain 1, the population-unadjusted reproduction number is defined as follows:

$$R_{s=1,t} = \mu_0 2\sigma(X_t) \quad (2)$$

where μ_0 is a scale parameter (3.3), σ is a logistic function, and X_t is a second-order autoregressive process with weekly time innovations, as specified in earlier work (*40*). The population-unadjusted reproduction number of the second strain is modelled as:

$$R_{s=2,t} = \rho \mathbf{1}_{[t_2,\infty)} R_{1,t} \quad (3)$$

$$\rho \sim \text{Normal}\ (1,1) \in [0,\infty) \quad (4)$$

where ρ is a parameter defining the relative transmissibility of strain 2 compared to strain 1, and $\mathbf{1}_{[t_2,\infty)}$ is an indicator function taking the value of 0 prior to t_2, and 1 thereafter, highlighting that strain 2 does not contribute to the observed evolution of the epidemic before its emergence. Introduction of the second strain at time t_2 is informed through our local molecular clock analysis (see **Fig. 2**, **fig. S9**). We note that the reproduction number estimates take into account the effect of population-level immunity and behavioural changes (modelled using a latent stochastic process). For the purposes of the primary results presented in the main text, it is assumed t_2= 9 Nov 2020, though four additional scenarios are presented varying the assumed date of P.1 emergence (see **table S5**). As in earlier models (*41*), we make the assumption of a homogeneously mixed population, and therefore ignore heterogeneites in transmission. This is an important area for future research.

Infections arise for each strain according to a discrete renewal process (*98, 99*):

$$i_{s,t} = \left(1 - \frac{n_{s,t}}{N}\right) R_{s,t} \sum_{\tau<t} i_{s,\tau} g_{t-\tau} \quad (5)$$

where N is the total population size, $n_{s,t}$ is the total extent of population immunity to strain s present at time t (accounting for the cumulative number of infections with strain s, the extent to which immunity from these infections has waned, and the degree of cross-protection infections with other strains provide, all of which are described in more detail below). The generation of infections is then determined by the fraction of the population susceptible and available to be infected, as well as the time-varying reproduction numbers of each strain $R_{s,t}$ and generation time distribution g from ref. (*41*).

For the original strain, infections are seeded for six days as:

$$i_{1,t_{1..6}} \sim \text{Exponential}(1/\tau) \quad (6)$$

$$\tau \sim \text{Exponential}(0.03) \quad (7)$$

and the second strain for 1 day (t_2) as:

$$i_{2,t_2} \sim \text{Normal}(1,1) \in [1,\infty) \quad (8)$$

The susceptible depletion term for strain s is modelled as:

$$n_{s,t} = \sum_{\tau<t} i_{s,\tau} W_{t-\tau} + \beta_s \left(1 - \alpha_{s,t}\right) \sum_{\tau<t} i_{\setminus s,\tau} W_{t-\tau} \quad (9)$$

The first term describes the contribution of prior infections with strain s to population-level immunity for s. The second term describes the contribution of prior infections with not strain s (i. e. $\setminus s$), which is a function of the assumed cross-immunity, $\beta_s \in [0,1]$. With this formulation, $\beta = 0$ indicates no cross-protection between infections caused by different strains and $\beta = 1$ indicates complete cross-protection between infections caused by different strains. In practice we only consider symmetric cross-immunity $\beta_s = \beta$, which is given the prior:

$$\beta \sim \text{Beta}(2,1) \quad (10)$$

This choice of prior reflects the need to maintain a null hypothesis of no change in cross-immunity while also capturing our uncertainty in the plausible range of immunity conferred by prior infections against the P.1 variant. A range of other choices from $\text{Beta}(1,1)$ to $\text{Beta}(4,1)$ are shown in **fig. S16**, where in each instance the posterior contracts toward partial cross-immunity greater than one half but less than complete cross-immunity. Additional prior sensitivity analyses assessing the choice of cross-immunity prior on other inferred quantities is presented in **table S6**.

We emphasize that the representation of cross-immunity included in the model does not distinguish between protection against severe disease and protection against infection. It is however possible that the extent of protection from severe disease and protection from infection may be different. This is an important distinction that should be explored in more sophisticated models in future research.

We also define $W_{t-\tau}$ as the time-dependent waning of immunity elicited by previous infection, which is modelled as a Rayleigh survival-type function. Recent results (*16, 43*) suggest that immunity is robust to waning over 8 months, and so for the results presented in the main text we use a Rayleigh parameter of $\sigma = 310$, which produces 50% of individuals still immune after 1 year. Estimates of the duration of protection elicited by prior infection remain uncertain however, and so we consider a range of different scenarios that vary the rate at which immunity wanes, the results of which are presented in **table S7**.

The cross-immunity susceptible term $\alpha_{s,t}$ is then modelled as:

$$\alpha_{s,t} = \frac{(1-\beta_s)}{N} \sum_{\tau<t} i_{s,\tau} W_{t-\tau} \quad (11)$$

And describes the proportion of infections with variant $\setminus s$ expected to occur in individuals who have been previously infected with variant s – itself a function of both cross-immunity (β) and the proportion of the population previously infected with s.

Infections in the model generate deaths via the following mechanistic relationship:

(12)

$$d_t = \sum_{s\in\{1,2\}} \text{ifr}_s \sum_{\tau<t} i_{s,\tau}\pi_{t-\tau}$$

with infection fatality ratio priors:

$$\text{ifr}_1 \sim \text{Normal}(0.32, 0.1^2) \in [0{,}100], \quad (13)$$

based on results from (*28*), adjusted for the age structure of the population of Manaus and:

$$\text{ifr}_2 \sim \text{RR} \cdot \text{ifr}_1, \quad (14)$$
$$\text{RR} \sim \text{LogNormal}(0{,}0.5) \in [0, \infty), \quad (15)$$

where RR denotes relative risk. We note that other work has shown that the degree of transmission advantage exhibited by the B.1.1.7 can vary over time as a function of control interventions and behavior (*17*).

The infection-to-death distribution π is composed of infection-to-onset and onset-to-death contributions as in previous work (*97*), with adaptations to take into account the most likely onset-to-death distribution in Amazonas state based on hospitalization distributions obtained by ref. (*100*).

The observation model uses two sources of data. In the first likelihood, the expected deaths D_t are modelled as negative-binomially distributed:

$$D_t \sim \text{NegativeBinomial}\left(d_t, d_t + \frac{d_t^2}{\phi}\right) \quad (16)$$

with mortality data d_t and dispersion prior:

$$\phi \sim \text{Normal}(0{,}5^2) \in [0, \infty). \quad (17)$$

The deaths data source is based on class 4 and 5 SARI COVID-19 deaths (*67*, *69*), from the SIVEP-Gripe database for Manaus city, that have been amended with Gaussian Process nowcasting to correct for known delays between deaths occurring and being recorded in the dataset (*43*, *45*).

The second likelihood is based on genomic data from symptomatic individuals presenting for testing and who had both a positive PCR diagnosis and the infecting SARS-CoV-2 genome sequenced. Specifically, the proportion of sequenced genomes identified as P.1 lineage at time t are modelled with a binomial likelihood:

$$G_t^+ \sim \text{Binomial}(G_t^+ + G_t^-, \theta_t) \quad (18)$$

with positive counts for P.1 denoted G_t^+ and counts for lineages not belonging to P.1 recorded as G_t^-. The success probability for P.1 positivity is modelled as the infection ratio:

$$\theta_t = \frac{\tilde{\imath}_{2,t}}{\tilde{\imath}_{1,t} + \tilde{\imath}_{2,t}} \quad (19)$$

where $\tilde{\imath}_{s,t}$ is given by:

(20)

$$\tilde{\imath}_{s,t} = \sum_{\tau \leq t} i_{s,\tau} \kappa_{t-\tau}$$

to account for the time varying PCR positivity displayed over the natural course of a COVID-19 infection. The distribution κ describes the probability of being PCR positive over time following infection, and is based on ref. (*101*).

Serological data is *not* explicitly used in our modelling framework but rather is used for external validation of the model outputs. For purposes of comparison with previously published, and independent, serological data, we also calculate:

$$\sum_{\tau \leq t} i_{s,\tau} C_{t-\tau} \tag{21}$$

where $C_{t-\tau}$ is the cumulative probability of an individual infected on day τ having seroconverted by time t. This distribution is empirical and based on ref. (*32*).

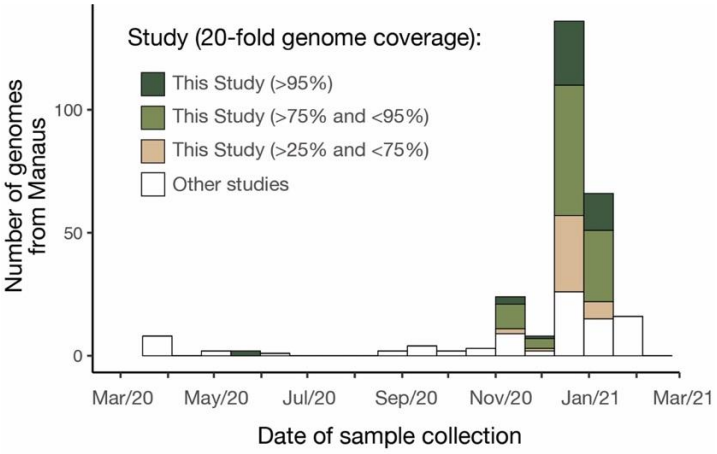


Figure S1.
Number of genome sequences from Manaus grouped by sequence coverage and study (see also Materials and Methods).

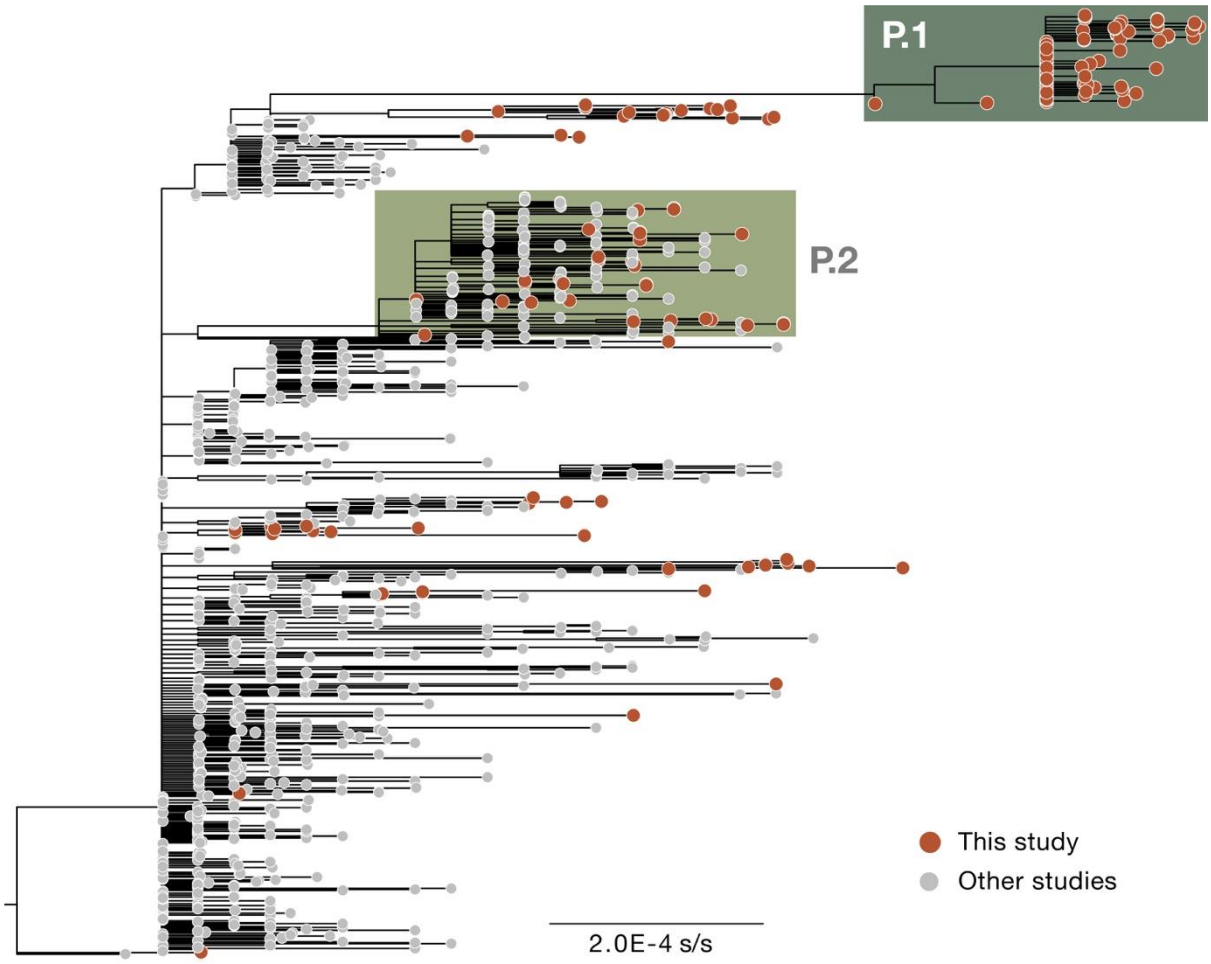


Figure S2.
Maximum likelihood tree estimated for *dataset A'* (*n*=988). This dataset was used to confirm lineage assignment for all sequences generated in this study regardless of genome coverage (see also **fig. S1**). s/s=nucleotide substitutions per site.

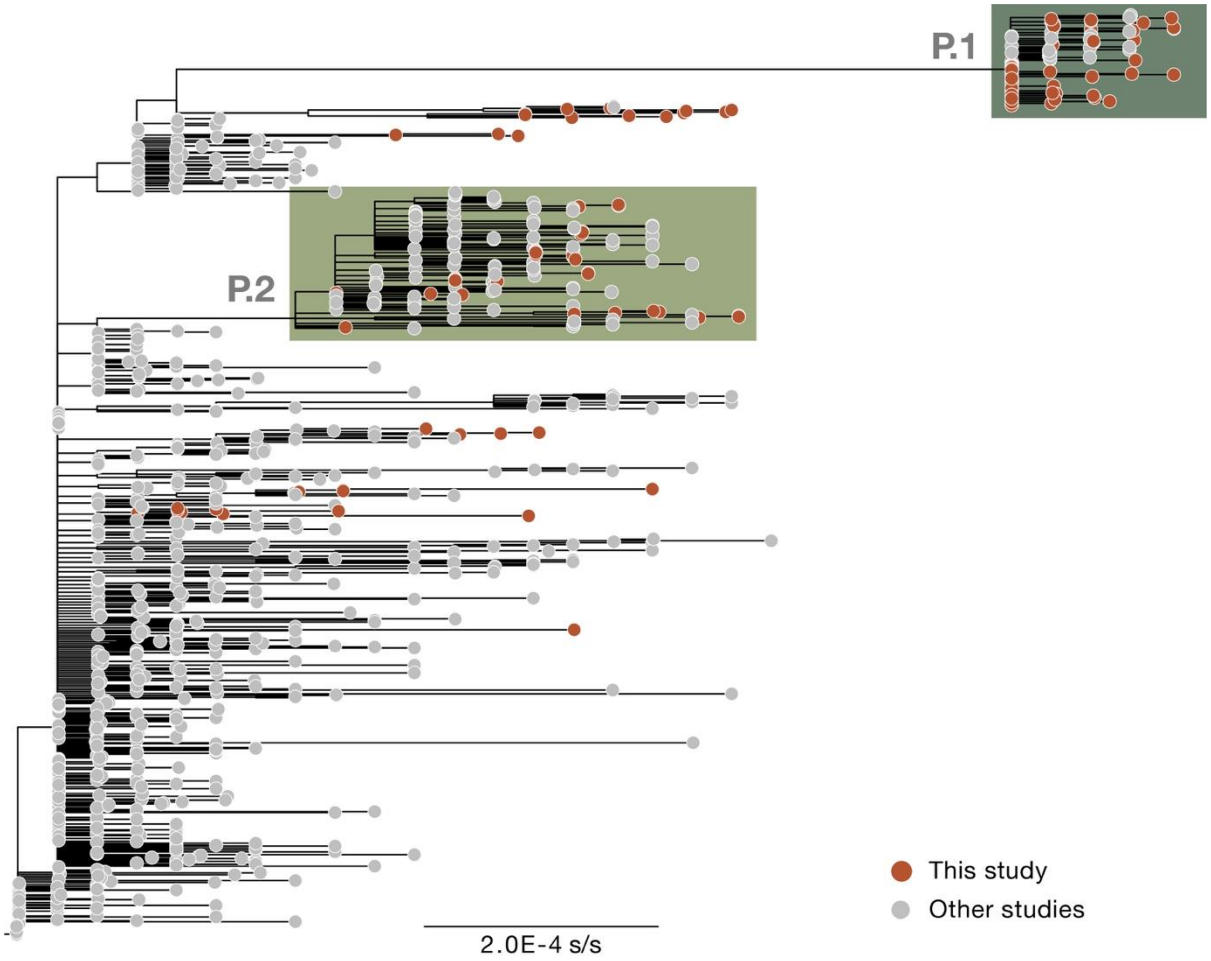


Figure S3.
Maximum likelihood tree estimated for *dataset B'* (*n*=962). s/s=nucleotide substitutions per site. This phylogeny includes only publicly available and published sequences classified as B.1.1.28, P.1 and P.2 lineages.

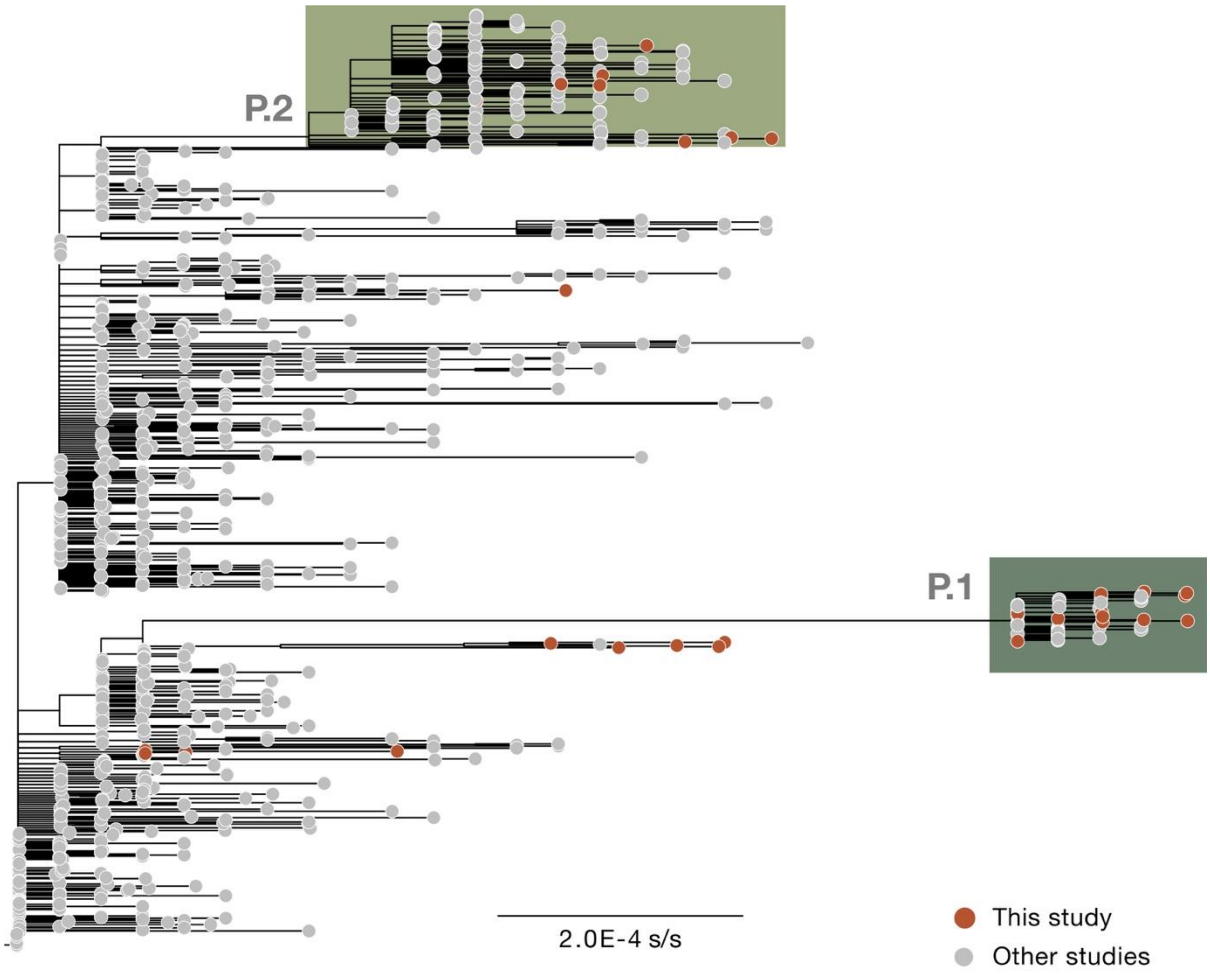


Figure S4.
Maximum likelihood tree estimated for *dataset C'* (*n*=871). s/s=nucleotide substitutions per site. This phylogeny includes only publicly available and published sequences classified as B.1.1.28, P.1 and P.2 lineages.

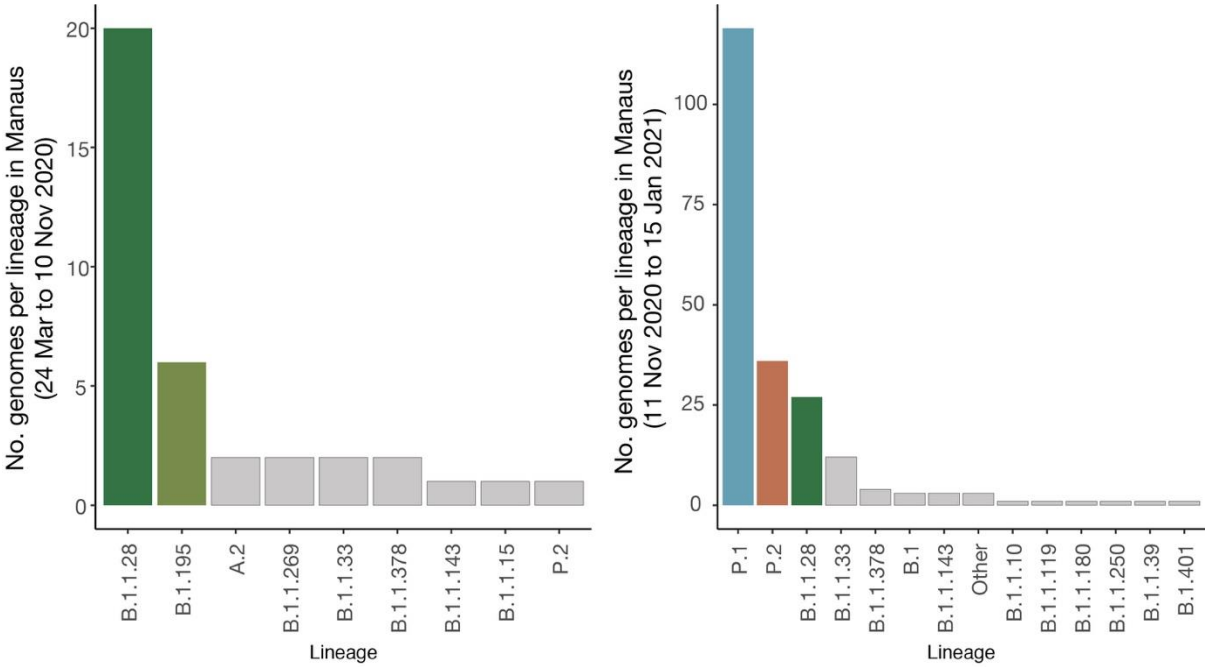


Figure S5.
Pango lineages identified in Manaus among our 184 sequences samples and publicly available genomes in GISAID between 24 March and 10 November 2020 (left panel) and between 11 November and 15 January 2020 (right panel). Duplicate sequences and sequences with no date or location of sample collection were removed. Coloured bars correspond to lineages that represent >10% of sequenced samples during each time period.

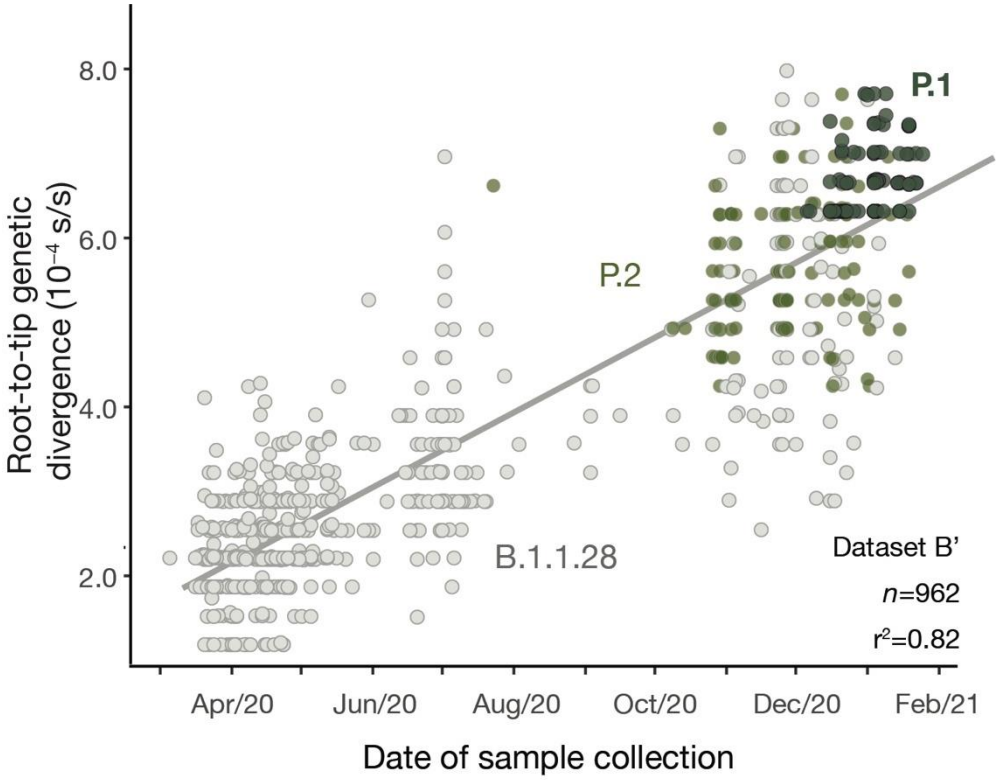


Figure S6.
Regression of root-to-tip genetic distances and sampling dates for *dataset B'* estimated using TempEst v.1.5.3 (*29*). Circles corresponds to the tips of the maximum likelihood phylogenetic tree show in **fig. S3**.

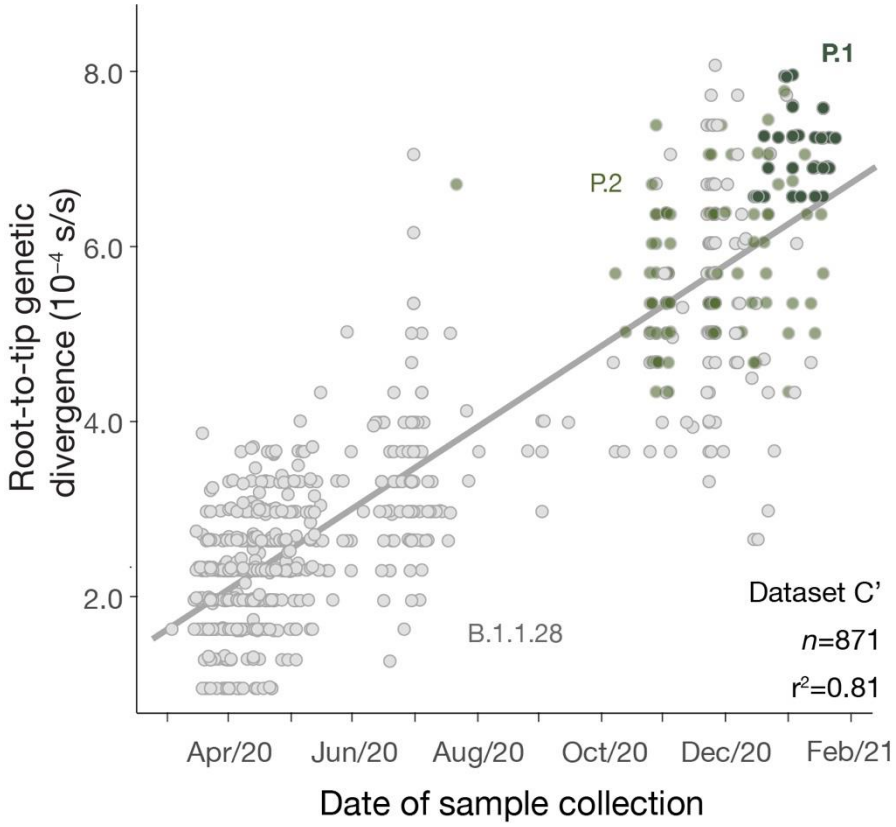


Figure S7.
Regression of root-to-tip genetic distances and sampling dates for *dataset C'* estimated using TempEst v.1.5.3 (*29*). Circles corresponds to the tips of the maximum likelihood phylogenetic tree show in **fig. S4**.

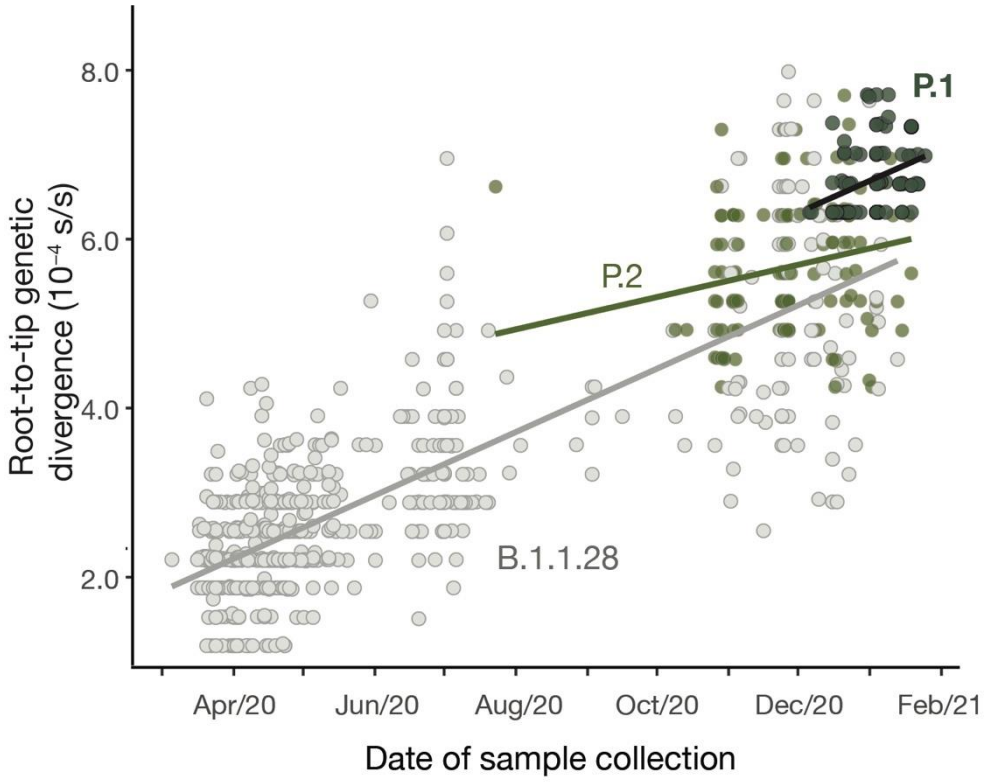


Figure S8.
Regression of root-to-tip genetic distances and sampling dates for *dataset B'* estimated using TempEst v.1.5.3 (*29*), with separate regression lines for B.1.1.28 and P.1 lineages computed in R v 3.6.2 (*84*). Circles corresponds to the tips of the maximum likelihood phylogenetic tree show in **fig. S3**.

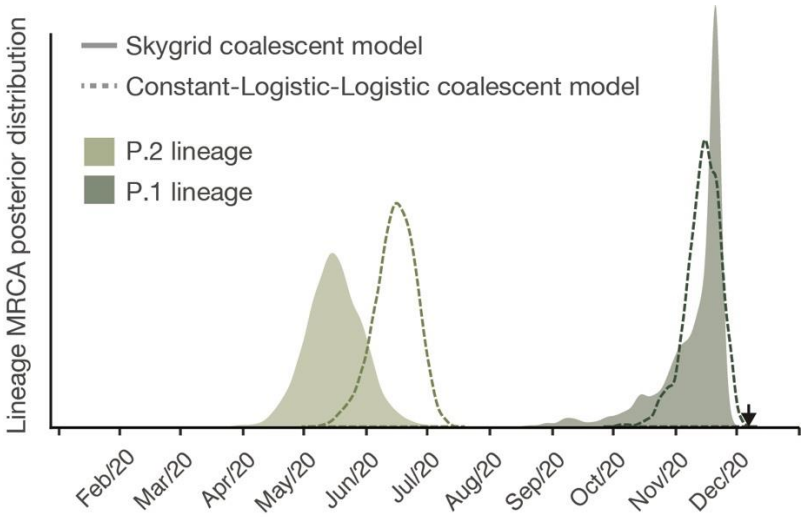


Figure S9.
Posterior estimates of the time of the most recent common ancestor of P.1 (dark green) and P.2 (light green) lineages estimated using a flexible non-parametric skygrid coalescent model (*33*). Dashed lines show the posterior estimates for the same evolutionary parameters but estimated using a constant-logistic-logistic model (see Materials and Methods for details). Both coalescent models are implemented in BEAST v.1.10 (*87*).

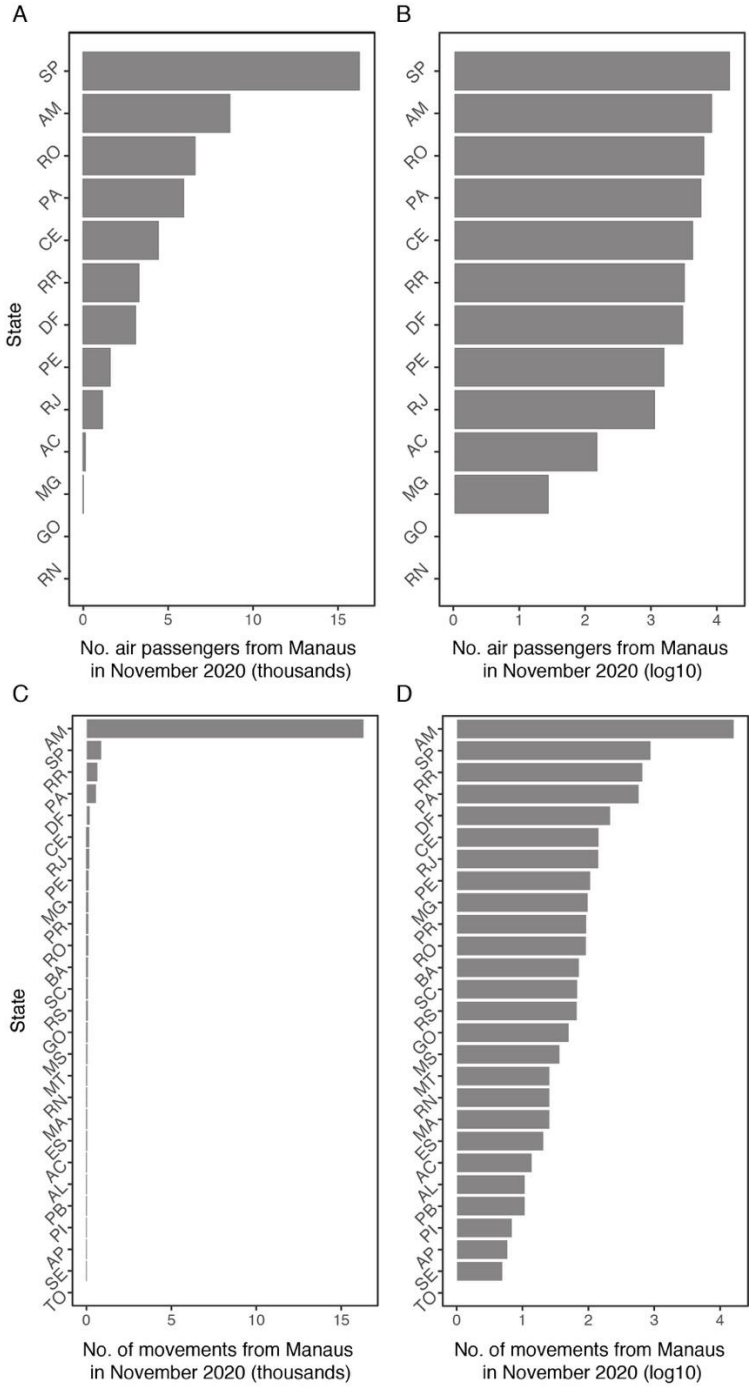


Figure S10.
Number of movements from Manaus to federal units in Brazil obtained from ANAC flight data (A and B) and from anonymized cell phone data (C and D). X-axis of panels B and D are shown in $\log_{10}$ units. The ISO 3166-2:BR codes of the states AC–Acre, AL–Alagoas, AP–Amapá, AM–Amazonas, BA–Bahia, CE–Ceará, DF–Distrito Federal, ES–Espírito Santo, GO–Goiás, MA–Maranhão, MT–Mato Grosso, MS–Mato Grosso do Sul, MG–Minas Gerais, PA–Pará, PB–Paraíba, PR–Paraná, PE–Pernambuco, PI–Piauí, RJ–Rio de Janeiro, RN–Rio Grande do Norte, RS–Rio Grande do Sul, RO–Rondônia, RR–Roraima, SC–Santa Catarina, SP–São Paulo, SE–Sergipe, TO–Tocantins.

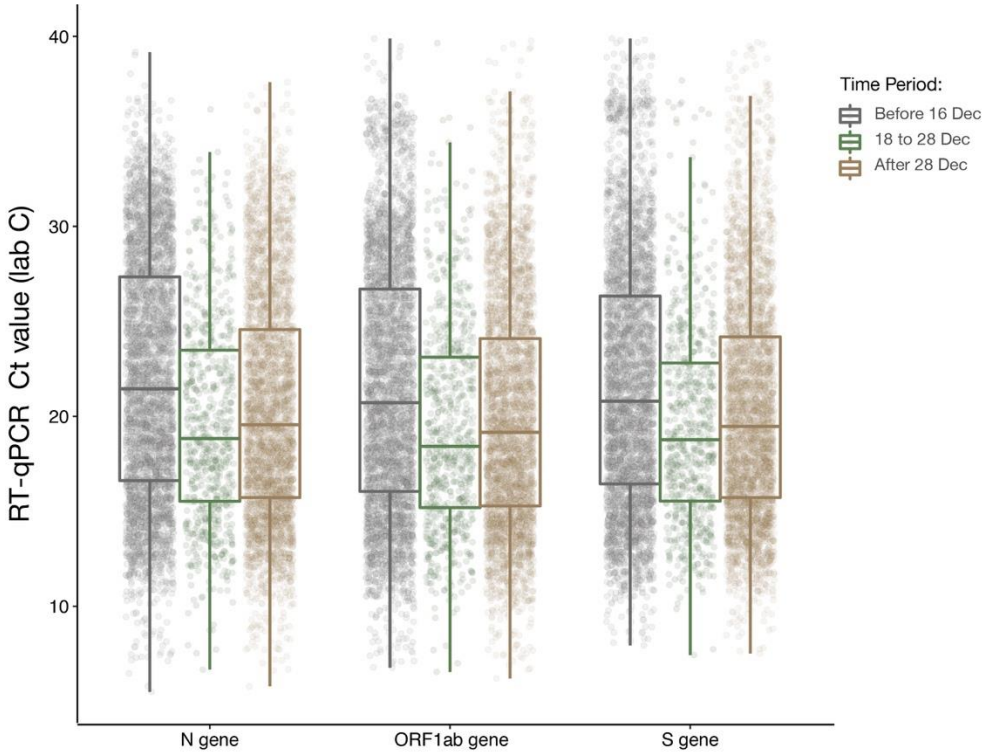


Figure S11.
Trends in RT-qPCR Ct values for COVID-19 infections in Manaus (laboratory C). Ct values for genes *N*, *ORF1ab*, and *S* in a sample of symptomatic cases presenting for testing at a healthcare facility in Manaus, stratified according to the period defined in **Fig. 2** (see main text). Line-list data can be found in **Data S3**.

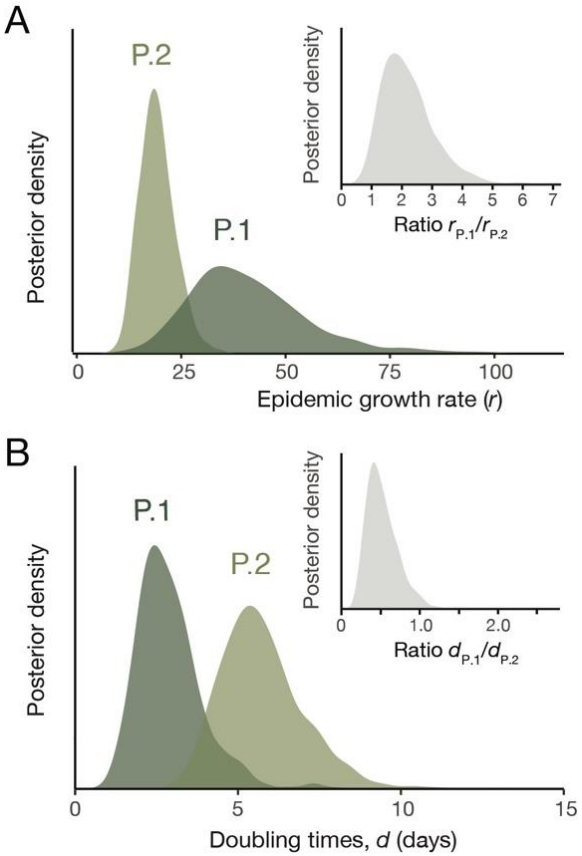


Figure S12.
Coalescent growth rates for P.1 and P.2 lineages estimated using a constant-logistic-logistic approach implemented in BEAST v.1.10 (*87*). (**A**) Light and dark green posterior probability distributions show virus lineage population growth (*r*) for P.2 and P.1, respectively. Inset shows posterior probability estimates for the ratio of epidemic growth rates between P.1 and P.2 (**B**) Light and dark green posterior probability distributions of the estimated doubling times for P.2 and P.1 lineage, respectively. Inset shows posterior probability estimates for the ratio of doubling times between P.1 and P.2.

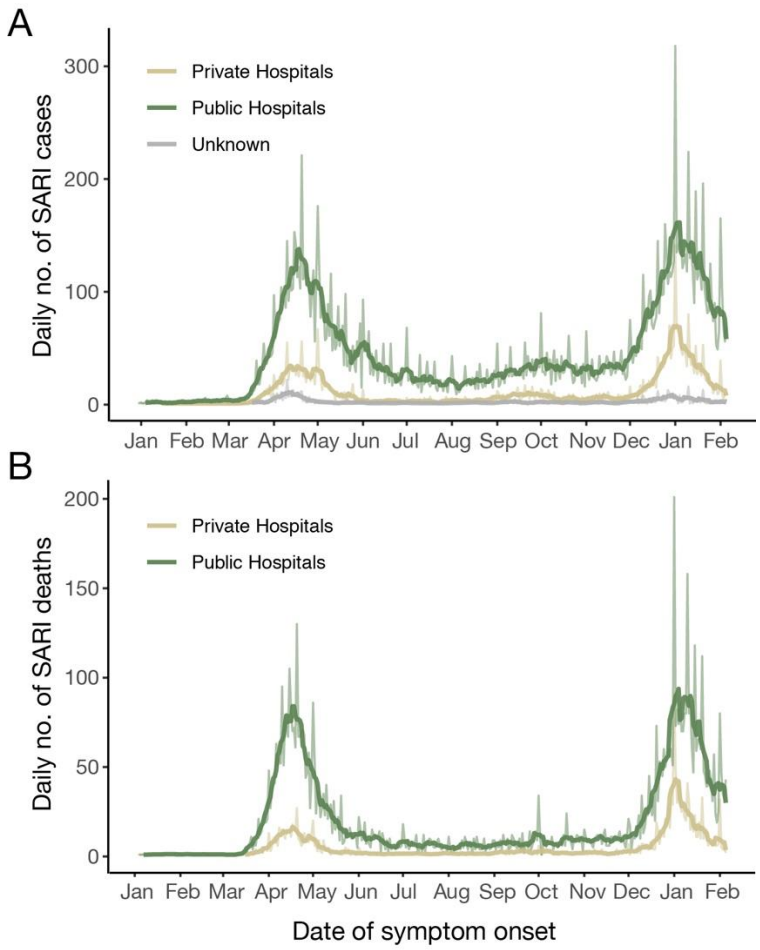


Figure S13.
Daily number of cases (A) and deaths (B) attending public or private hospitals in Manaus. Dark solid lines show the 7-day rolling average. Data was obtained from the SIVEP-Gripe dataset described in Materials and Methods. SARI = severe acute respiratory infections.

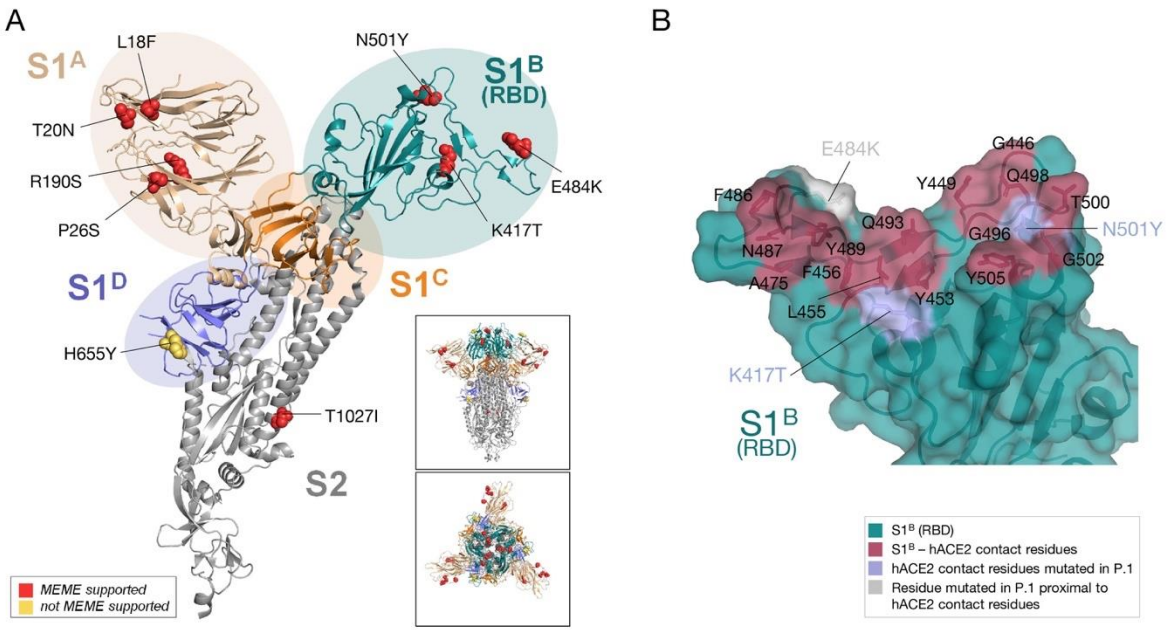


Figure S14.
Mapping of adaptive substitutions onto the structure of SARS-COV-2 S. (A) Lineage defining mutations within the S protein of the P.1 lineage are mapped onto the spike glycoprotein structure of SARS-CoV-2 (*93*) (*NSMB;* PDB: 6ZGI). Cartoon representation of one protomer of the trimeric S ectodomain structure with the different domains and subunits indicated by color: $S1^A$ (wheat), $S1^B$ (RBD, teal), $S1^C$ (orange), $S1^D$ (blue), S2 subunit (grey). Residues under selection are shown as spheres with associated mutations indicated and colored according to the respective type of selection as analyzed under MEME (*92*) and supported for at least one branch (red) or not supported (yellow). The positively selected residue V1176F was not resolved in the cryoEM map of the SARS-CoV-2 S ectodomain used here. The inset panels reflect the same presentation in the trimeric context of the S protein in a side (*upper panel*) and top-down view (*lower panel*). N-linked glycans are omitted for clarity. (**B**) Surface representation of the SARS-CoV-2 S RBD-hACE2 contact interface with residues mutated in the RBD of P.1 highlighted. Contact residues as observed in the SARS-CoV-2 S RBD (deep teal) in complex with hACE2 (*11*) (PDB: 6MOJ) are colored red with side chains shown as sticks. Residues mutated in lineage P.1 that are part of the contact interface (K417 and N501) are colored light blue. A nearby residue that is observed to be mutated in P.1 that is not part of the direct contact interface (E484) is colored grey with side chains shown as sticks.

Figure S15.
Intra-host sequencing data by reference to variant allele frequency measurements at P.1 lineage-defining positions. The analysis shows that lineage-defining mutations are highly stable across genomes taking into account the underlying sequence read error rate at Oxford Nanopore sequencing, which limits the level of detection of mixed species to a minimum relative abundance of 5-10%. One genome demonstrates coverage patterns suggestive of mixed infection (CD1721) and excluded from phylodynamic analysis.

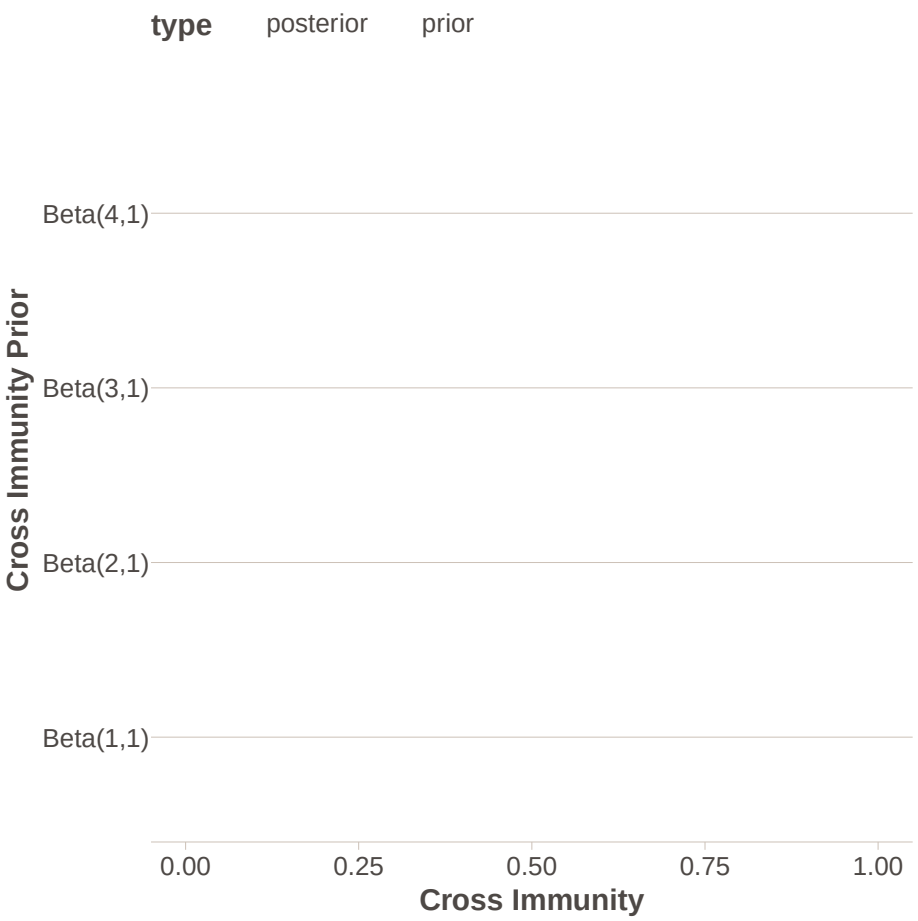


Figure S16. Cross-immunity prior and posterior distributions for a range of Beta priors. The red bars represent the prior, and the blue bars the resultant posterior. Contraction of the posterior toward partial cross-immunity greater than 0.5 but less than complete cross-immunity is consistently observed.

Table S1.

P.1 lineage-defining mutations. These have been defined based on the genomic datasets analysed in this study, which included 95 P.1 isolates. None of the mutations was observed in other isolates in the lineage B.1.1.28 analysed in this study. [1]One isolate (ID: CD2241) did not present the 12778C>T; [2]One isolate (ID: CD2293) did not present the 21614C>T; [3]One isolate (ID: CD2293) did not present the 21638C>T; [4]One isolate (ID: CD1721) did not present the 22132G>T; [5]Fourty-five percent of P.1 isolates sequenced here did not acquire this insertion, which is located in the intergenic region between ORF8 and N genes. The last column indicates positively selected sites in P.1 lineage-defining mutations with statistical support (MEME p ≤0.05).

Gene	Amino acid	Nucleotide change	dN/dS>1
ORF1ab	-	733T>C	
	-	2749C>T	
	S1188L	3828C>T	Yes
	K1795Q	5648A>C	Yes
	-	del11288-11296 (3675-3677)	
	-	12778C>T[1]	
	-	13860C>T	
	E5662D	17259G>T	
Spike	L18F	21614C>T[2]	Yes
	T20N	21621C>A	Yes
	P26S	21638C>T[3]	Yes
	D138Y	21974G>T	
	R190S	22132G>T[4]	Yes
	K417T	22812A>C	Yes
	E484K	23012G>A	Yes
	N501Y	23063A>T	Yes
	H655Y	23525C>T	
	T1027I	24642C>T	Yes
ORF8	E92K	28167G>A	
	-	28263insAACA[5]	
N	P80R	28512C>G	Yes
	-	28877A>T	
	-	28878G>C	

Table S2.
Overview of the P.1 sequences used in this study. N = Number; Ct = cycle threshold (RT-PCR); LDM = lineage defining mutations (see definition in **table S1**); no. = number. Line-list information for the sequence data generated by this study can be found in **Data S2**.

Genome coverage	N	Primary use in this study	Lab source	Ct values	Mean no. P.1 LDM (range)
>25 to <75%	25	Epidemiological (Epi) Modelling	Lab A	E: 20.2 (13.3–28.8) N: 23.6 (15.2–41)	11 (5–18)
>75 to <95%	47	Phylodynamic and Epi Modelling	Lab A, B	E: 18.9 (12.9–30.6) N: 21.1 (7–43)	19 (14–23)
>95%	23	Phylodynamic and Epi Modelling	Lab A, B	E: 17.6 (13.6–23.4) N: 18.4 (9–30)	22 (18–23)

Table S3.
Epidemiological information regarding P.1 sequences from GISAID used for phylodynamics analyses. *Federal unit corresponding to municipality of sampling is São Paulo state, except for Teresina (Piauí state). N. A. = Not available.

GISAI ID	Collection date	Age	Sex	Municipality of residence	Municipality of sampling*	Travel history
EPI_ISL_906075	2021-01-1	83	M	Manaus	Sao Paulo	No
EPI_ISL_906069	2021-01-1	45	M	Manaus	Aguas Lindonia	No
EPI_ISL_906076	2021-01-1	52	M	Manaus	Sao Caetano Sul	No
EPI_ISL_906077	2021-01-1	49	F	Manaus	Sao Caetano Sul	No
EPI_ISL_906080	2021-01-2	74	M	Manaus	Sao Paulo	No
EPI_ISL_906081	2021-01-2	69	M	Manaus	Sao Paulo	No
EPI_ISL_940614	2021-01-1	57	F	Manaus	Teresina (PI)	No
EPI_ISL_940615	2021-01-1	38	F	Manaus	Teresina (PI)	No
EPI_ISL_906071	2021-01-1	59	M	Manaus	Teresina (PI)	No
EPI_ISL_940617	2021-01-1	30	F	Manaus	Teresina (PI)	No
EPI_ISL_940618	2021-01-1	46	F	Manaus	Teresina (PI)	No
EPI_ISL_940620	2021-01-1	46	M	Manaus	Sao Paulo	No
EPI_ISL_940623	2021-01-0	64	F	Manaus	Sao Paulo	No
EPI_ISL_940624	2021-01-1	29	M	Manaus	Sao Paulo	No
EPI_ISL_940626	2021-01-2	78	M	Manaus	Sao Caetano Sul	No
EPI_ISL_940627	2021-01-2	64	M	Manaus	Sao Caetano Sul	No
EPI_ISL_833169	2020-12-2	N.A.	N.A.	N.A.	N.A.	N.A.
EPI_ISL_940630	2021-01-2	40	M	Rio Janeiro	Sao Paulo	Roraima
EPI_ISL_875689	2021-01-1	65	F	Sao Paulo	Sao Paulo	Manaus
EPI_ISL_872191	2021-01-1	51	F	Sao Paulo	Sao Paulo	Manaus
EPI_ISL_872192	2021-01-1	49	F	Sao Paulo	Sao Paulo	Manaus
EPI_ISL_940619	2021-01-1	84	F	Sao Paulo	Sao Paulo	Manaus
EPI_ISL_940621	2021-01-1	40	F	Manaus	Sao Paulo	N.A.
EPI_ISL_940622	2021-01-1	48	M	Sao Paulo	Sao Paulo	Manaus
EPI_ISL_940625	2021-01-1	81	F	Sao Paulo	Sao Paulo	Manaus
EPI_ISL_875688	2021-01-0	49	M	Sao Paulo	Sao Paulo	Manaus

Table S4.
Proportion of P.1 cases in Manaus. Note that week commencing on 30 Nov 2020 includes the date of the first P.1 case detected in our study (6 Dec 2020).

Date (Week Commencing)	No. Sequenced	No. P.1	Proportion
2 November 2020	9	0	0
9 November 2020	2	0	0
16 November 2020	4	0	0
23 November 2020	NA	NA	NA
30 November 2020	2	1	50%
7 December 2020	7	1	14%
14 December 2020	24	7	29%
21 December 2020	37	20	54%
28 December 2020	14	8	57%
4 January 2021	46	40	87%

Table S5.
Inferred changes in epidemiological characteristics of P.1, depending on the timing of P.1 emergence assumed. The central estimate (derived from phylogenetic molecular clock analyses) used is the 9th November. Sensitivity results are shown for this date plus or minus 1 week, as well as for the 95% Bayesian Credible Interval of the most recent common ancestor of the P.1 lineage (6th October and 24th November). The results presented are the mean, with the Bayesian 95% quartiles in brackets.

	Epidemiological Characteristic		
Timing of P.1 Emergence	**Degree of Cross-Immunity**	**Transmissibility Increase**	**Relative Risk of Mortality**
6 Oct 2020	0.71 (0.35-0.94)	1.38 (0.82-2.05)	1.88 (0.99-3.11)
2 Nov 2020	0.69 (0.33-0.93)	1.89 (1.13-2.82)	1.72 (0.95-2.86)
9 Nov 2020	0.65 (0.28-0.90)	2.09 (1.26-3.10)	1.59 (0.90-2.51)
16 Nov 2020	0.57 (0.23-0.83)	2.34 (1.50-3.37)	1.34 (0.82-2.10)
24 Nov 2020	0.42 (0.13-0.69)	2.75 (1.94-3.78)	1.12 (0.73-1.77)

Table S6.
Inferred changes in epidemiological characteristics of P.1, depending on cross-immunity prior assumptions. Sensitivity results presented are the mean, with the Bayesian 95% quartiles in brackets.

	Epidemiological Characteristic		
Degree of Cross-Immunity prior	**Degree of Cross-Immunity**	**Transmissibility Increase**	**Relative Risk of Mortality**
Beta(1,1)	0.55 (0.09–0.87)	1.90 (1.10-2.95)	1.42 (0.77-2.38)
Beta(2,1)	0.65 (0.28-0.90)	2.09 (1.26-3.10)	1.59 (0.90-2.51)
Beta(3,1)	0.70 (0.39-0.91)	2.19 (1.41-3.10)	1.67 (0.97-2.62)
Beta(4,1)	0.73 (0.46-0.92)	2.31 (1.49-3.27)	1.74 (1.06-264)
Complete immune-escape	0	1.07 (0.82-1.39)	0.99 (0.65-1.68)
Complete cross-immunity	1	3.89 (3.52-4.26)	2.72 (1.54-3.88)

Table S7.
Inferred changes in epidemiological characteristics of P.1, depending on the rate of natural immunity waning assumed (for an emergence date of 9 Nov 2020). Sensitivity results presented are the mean, with the Bayesian 95% quartiles in brackets.

	Epidemiological Characteristic		
Duration of Immunity Assumed	**Degree of Cross-Immunity**	**Transmissibility Increase**	**Relative Risk of Mortality**
50% waning over 6-month period	0.48 (0.13-0.82)	2.25 (1.60-3.16)	1.35 (0.98-1.89)
50% waning over 8-month period	0.56 (0.19-0.86)	1.92 (1.19-2.91)	1.54 (0.99-2.34)
50% waning over 1-year period	0.65 (0.28-0.90)	2.09 (1.26-3.10)	1.59 (0.90-2.51)
50% waning over 1.5-year period	0.69 (0.35-0.91)	2.17 (1.36-3.11)	1.61 (0.86-2.65)
50% waning over 2-year period	0.71 (0.39-0.92)	2.24 (1.41-3.17)	1.65 (0.84-2.71)

Table S8.
Inferred changes in epidemiological characteristics of P.1, depending on non-P.1 infection fatality ratio (IFR) prior assumptions. Sensitivity results presented are the mean, with the Bayesian 95% quartiles in brackets.

	Epidemiological Characteristic		
Non-P.1 IFR prior	**Degree of Cross-Immunity**	**Transmissibility Increase**	**Relative Risk of Mortality**
Normal^{+}(0.24,0.01)	0.65 (0.28-0.90)	2.06 (1.29-3.03)	1.61 (0.92-2.57)
Normal^{+}(0.32,0.01)	0.65 (0.28-0.90)	2.09 (1.26-3.10)	1.59 (0.90-2.51)
Normal^{+}(0.48,0.01)	0.63 (0.25-0.90)	2.11 (1.28-3.12)	1.46 (0.80-2.36)
Gamma(3.2,10)	0.65 (0.29-0.89)	2.07 (1.25-3.08)	1.58 (0.88-2.50)

Table S9.
Inferred changes in epidemiological characteristics of P.1, depending on relatively risk of mortality prior assumptions. Sensitivity results presented are the mean, with the Bayesian 95% quartiles in brackets.

	Epidemiological Characteristic		
Relative Risk of Mortality prior	**Degree of Cross-Immunity**	**Transmissibility Increase**	**Relative Risk of Mortality**
LogNormal(0,0.25)	0.54 (0.21–0.80)	1.83 (1.16-2.66)	1.23 (0.85-1.75)
LogNormal(0,0.5)	0.65 (0.28-0.90)	2.09 (1.26-3.10)	1.59 (0.90-2.51)
LogNormal(0,0.75)	0.69 (0.34-0.92)	2.21 (1.31-3.27)	1.74 (0.95-2.89)

Table S10.
Inferred changes in epidemiological characteristics of P.1, depending on prior assumptions for the increase in transmissibility of P.1 compared to non-P.1. Sensitivity results presented are the mean, with the Bayesian 95% quartiles in brackets.

	Epidemiological Characteristic		
Transmissibility increase prior	**Degree of Cross-Immunity**	**Transmissibility Increase**	**Relative Risk of Mortality**
Normal^{+}(2,1)	0.75 (0.43–0.95)	2.46 (1.49-3.54)	1.77 (0.99-2.78)
Normal^{+}(1,1)	0.65 (0.28-0.90)	2.09 (1.26-3.10)	1.59 (0.90-2.51)
Gamma(2.2,0.5)	0.62 (0.25-0.89)	2.01 (1.21-3.06)	1.51 (0.86-2.52)

Data S1. (separate file)
Metadata for 1084 SARS-CoV-2 samples from Manaus (laboratory B).

Data S2. (separate file)
Metadata and GISAID IDs for 184 sequenced samples from Manaus.

Data S3. (separate file)
Metadata for 8542 SARS-CoV-2 samples from Manaus (laboratory C).

Data S4. (separate file)
GISAID Acknowledgment table (used for datasets B and C).

Data S5. (separate file)
Flight mobility data from Manaus (state level).

Data S6. (separate file)
Cell-phone derived mobility data from Manaus (municipality level).

References and Notes

1. P. C. Hallal, F. P. Hartwig, B. L. Horta, M. F. Silveira, C. J. Struchiner, L. P. Vidaletti, N. A. Neumann, L. C. Pellanda, O. A. Dellagostin, M. N. Burattini, G. D. Victora, A. M. B. Menezes, F. C. Barros, A. J. D. Barros, C. G. Victora, SARS-CoV-2 antibody prevalence in Brazil: Results from two successive nationwide serological household surveys. *Lancet Glob. Health* **8**, e1390–e1398 (2020). doi:10.1016/S2214-109X(20)30387-9 Medline

2. L. F. Buss, C. A. Prete Jr., C. M. M. Abrahim, A. Mendrone Jr., T. Salomon, C. de Almeida-Neto, R. F. O. França, M. C. Belotti, M. P. S. S. Carvalho, A. G. Costa, M. A. E. Crispim, S. C. Ferreira, N. A. Fraiji, S. Gurzenda, C. Whittaker, L. T. Kamaura, P. L. Takecian, P. da Silva Peixoto, M. K. Oikawa, A. S. Nishiya, V. Rocha, N. A. Salles, A. A. de Souza Santos, M. A. da Silva, B. Custer, K. V. Parag, M. Barral-Netto, M. U. G. Kraemer, R. H. M. Pereira, O. G. Pybus, M. P. Busch, M. C. Castro, C. Dye, V. H. Nascimento, N. R. Faria, E. C. Sabino, Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. *Science* **371**, 288–292 (2021). doi:10.1126/science.abe9728 Medline

3. C. Álvarez-Antonio, G. Meza-Sánchez, C. Calampa, W. Casanova, C. Carey, F. Alava, H. Rodríguez-Ferrucci, A. M. Quispe, Seroprevalence of Anti-SARS-CoV-2 Antibodies in Iquitos, Loreto, Peru. medRxiv 21249913 [Preprint] 20 January 2021. doi:10.1101/2021.01.17.21249913.

4. M. Mercado, M. Ospina, Institutoo Nacional de Salud, "Seroprevalencia de SARS-CoV-2 durante la epidemia en Colombia: estudio país" (2020); www.ins.gov.co/BibliotecaDigital/Seroprevalencia-estudio-colombia.pdf.

5. Fundação de Vigilância em Saúde do Amazonas, "Perfil clínico e demográfico dos casos de Covid-19 no estado do Amazonas: uma análise comparativa entre 2020 e 2021," No. 17 (2021); www.fvs.am.gov.br/media/publicacao/boletim_covid_17.pdf.

6. A. J. Greaney, A. N. Loes, K. H. D. Crawford, T. N. Starr, K. D. Malone, H. Y. Chu, J. D. Bloom, Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463–476.e6 (2021). doi:10.1016/j.chom.2021.02.003 Medline

7. T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. D. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, J. D. Bloom, Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310.e20 (2020). doi:10.1016/j.cell.2020.08.012 Medline

8. M. A. Suchard, R. E. Weiss, J. S. Sinsheimer, Testing a molecular clock without an outgroup: Derivations of induced priors on branch-length restrictions in a Bayesian framework. *Syst. Biol.* **52**, 48–54 (2003). doi:10.1080/10635150390132713 Medline

9. Z. Wang, F. Schmidt, Y. Weisblum, F. Muecksch, C. O. Barnes, S. Finkin, D. Schaefer-Babajew, M. Cipolla, C. Gaebler, J. A. Lieberman, T. Y. Oliveira, Z. Yang, M. E. Abernathy, K. E. Huey-Tubman, A. Hurley, M. Turroja, K. A. West, K. Gordon, K. G. Millard, V. Ramos, J. Da Silva, J. Xu, R. A. Colbert, R. Patel, J. Dizon, C. Unson-O'Brien, I. Shimeliovich, A. Gazumyan, M. Caskey, P. J. Bjorkman, R. Casellas, T.

Hatziioannou, P. D. Bieniasz, M. C. Nussenzweig, mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature* (2021). doi:10.1038/s41586-021-03324-6 Medline

10. Y. Weisblum, F. Schmidt, F. Zhang, J. DaSilva, D. Poston, J. C. C. Lorenzi, F. Muecksch, M. Rutkowska, H.-H. Hoffmann, E. Michailidis, C. Gaebler, M. Agudelo, A. Cho, Z. Wang, A. Gazumyan, M. Cipolla, L. Luchsinger, C. D. Hillyer, M. Caskey, D. F. Robbiani, C. M. Rice, M. C. Nussenzweig, T. Hatziioannou, P. D. Bieniasz, Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9**, e61312 (2020). doi:10.7554/eLife.61312 Medline

11. J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, X. Wang, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020). doi:10.1038/s41586-020-2180-5 Medline

12. H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M. Allam, A. Ismail, T. Mohale, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, F. Petruccione, A. Sigal, D. Hardie, G. Marais, N. Y. Hsiao, S. Korsman, M.-A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O. Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C. K. Wibmer, B. T. Sewell, J. Lourenço, L. C. J. Alcantara, S. L. Kosakovsky Pond, S. Weaver, D. Martin, R. J. Lessells, J. N. Bhiman, C. Williamson, T. de Oliveira, Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* (2021). doi:10.1038/s41586-021-03402-9 Medline

13. A. Rambaut, N. Loman, O. Pybus, W. Barclay, J. Barrett, Carabelli., A., Connor, T., Peacock, T., Robertson, D. L., Volz, E., on behalf of COVID-19 Genomics Consortium UK (CCoG-UK), "Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations" (2020); https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.

14. SAMRC Report on Weekly Deaths in South Africa (available at https://www.samrc.ac.za/reports/report-weekly-deaths-south-africa?bc=254). (2021).

15. N. L. Washington *et al.*, Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. medRxiv 21251159 [Preprint] 7 February 2021). doi:10.1101/2021.02.06.21251159.

16. C. H. Hansen, D. Michlmayr, S. M. Gubbels, K. Mølbak, S. Ethelberg, Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: A population-level observational study. *Lancet* **397**, 1204–1212 (2021). doi:10.1016/S0140-6736(21)00575-4 Medline

17. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. J. Loman, J. T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N. M. Ferguson; COVID-19 Genomics UK (COG-UK) consortium, Assessing

transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* (2021). doi:10.1038/s41586-021-03470-x Medline

18. D. D. S. Candido, A. Watts, L. Abade, M. U. G. Kraemer, O. G. Pybus, J. Croda, W. de Oliveira, K. Khan, E. C. Sabino, N. R. Faria, Routes for COVID-19 importation in Brazil. *J. Travel Med.* **27**, taaa042 (2020). doi:10.1093/jtm/taaa042 Medline

19. D. S. Candido, I. M. Claro, J. G. de Jesus, W. M. Souza, F. R. R. Moreira, S. Dellicour, T. A. Mellan, L. du Plessis, R. H. M. Pereira, F. C. S. Sales, E. R. Manuli, J. Thézé, L. Almeida, M. T. Menezes, C. M. Voloch, M. J. Fumagalli, T. M. Coletti, C. A. M. da Silva, M. S. Ramundo, M. R. Amorim, H. H. Hoeltgebaum, S. Mishra, M. S. Gill, L. M. Carvalho, L. F. Buss, C. A. Prete Jr., J. Ashworth, H. I. Nakaya, P. S. Peixoto, O. J. Brady, S. M. Nicholls, A. Tanuri, Á. D. Rossi, C. K. V. Braga, A. L. Gerber, A. P. de C Guimarães, N. Gaburo Jr., C. S. Alencar, A. C. S. Ferreira, C. X. Lima, J. E. Levi, C. Granato, G. M. Ferreira, R. S. Francisco Jr., F. Granja, M. T. Garcia, M. L. Moretti, M. W. Perroud Jr., T. M. P. P. Castiñeiras, C. S. Lazari, S. C. Hill, A. A. de Souza Santos, C. L. Simeoni, J. Forato, A. C. Sposito, A. Z. Schreiber, M. N. N. Santos, C. Z. de Sá, R. P. Souza, L. C. Resende-Moreira, M. M. Teixeira, J. Hubner, P. A. F. Leme, R. G. Moreira, M. L. Nogueira, N. M. Ferguson, S. F. Costa, J. L. Proenca-Modena, A. T. R. Vasconcelos, S. Bhatt, P. Lemey, C.-H. Wu, A. Rambaut, N. J. Loman, R. S. Aguiar, O. G. Pybus, E. C. Sabino, N. R. Faria; Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic Network, Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260 (2020). doi:10.1126/science.abd2161 Medline

20. W. M. de Souza, L. F. Buss, D. D. S. Candido, J.-P. Carrera, S. Li, A. E. Zarebski, R. H. M. Pereira, C. A. Prete Jr., A. A. de Souza-Santos, K. V. Parag, M. C. T. D. Belotti, M. F. Vincenti-Gonzalez, J. Messina, F. C. da Silva Sales, P. D. S. Andrade, V. H. Nascimento, F. Ghilardi, L. Abade, B. Gutierrez, M. U. G. Kraemer, C. K. V. Braga, R. S. Aguiar, N. Alexander, P. Mayaud, O. J. Brady, I. Marcilio, N. Gouveia, G. Li, A. Tami, S. B. de Oliveira, V. B. G. Porto, F. Ganem, W. A. F. de Almeida, F. F. S. T. Fantinato, E. M. Macário, W. K. de Oliveira, M. L. Nogueira, O. G. Pybus, C.-H. Wu, J. Croda, E. C. Sabino, N. R. Faria, Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. *Nat. Hum. Behav.* **4**, 856–865 (2020). doi:10.1038/s41562-020-0928-4 Medline

21. J. G. Jesus, C. Sacchi, D. D. S. Candido, I. M. Claro, F. C. S. Sales, E. R. Manuli, D. B. B. D. Silva, T. M. Paiva, M. A. B. Pinho, K. C. O. Santos, S. C. Hill, R. S. Aguiar, F. Romero, F. C. P. D. Santos, C. R. Gonçalves, M. D. C. Timenetsky, J. Quick, J. H. R. Croda, W. Oliveira, A. Rambaut, O. G. Pybus, N. J. Loman, E. C. Sabino, N. R. Faria, Importation and early local transmission of COVID-19 in Brazil, 2020. *Rev. Inst. Med. Trop. São Paulo* **62**, e30 (2020). doi:10.1590/s1678-9946202062030 Medline

22. I. M. Claro, F. C. da Silva Sales, M. S. Ramundo, D. S. Candido, C. A. M. Silva, J. G. de Jesus, E. R. Manuli, C. M. de Oliveira, L. Scarpelli, G. Campana, O. G. Pybus, E. C. Sabino, N. R. Faria, J. E. Levi, Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020. *Emerg. Infect. Dis.* **27**, 970–972 (2021). doi:10.3201/eid2703.210038 Medline

23. V. A. D. Nascimento, A. L. G. Corado, F. O. D. Nascimento, Á. K. A. D. Costa, D. C. G. Duarte, S. L. B. Luz, L. M. F. Gonçalves, M. S. Jesus, C. F. D. Costa, E. Delatorre, F. G. Naveca, Genomic and phylogenetic characterisation of an imported case of SARS-CoV-2 in Amazonas State, Brazil. *Mem. Inst. Oswaldo Cruz* **115**, e200310 (2020). doi:10.1590/0074-02760200310 Medline

24. J. R. Tyson, P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J. H. Choi, H. Lapointe, K. Kamelian, A. D. Smith, N. Prystajecky, 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* 2020.09.04.283077 doi:10.1101/2020.09.04.283077 (2020). Medline

25. World Health Organization, (2021). Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health, 8 January 2021. World Health Organization (2021); https://apps.who.int/iris/handle/10665/338480.

26. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020). doi:10.1038/s41564-020-0770-5 Medline

27. N. R. Faria, I. M. Claro, D. Candido, L. A. M. Franco, P. S. Andrade, T. M. Coletti, C. A. M. Silva, F. C. Sales, E. R. Manuli, R. S. Aguiar, N. Gaburo, C. C. Camilo, N. A. Fraiji, M. A. E. Crispim, M. P. S. S. Carvalho, A. Rambaut, N. Loman, O. G. Pybus, E. C. Sabino, on behalf of CADDE Genomic Network, "Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings" (2021); https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586.

28. 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). doi:10.3201/eid2704.210138 Medline

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

30. B. Choi, M. C. Choudhary, J. Regan, J. A. Sparks, R. F. Padera, X. Qiu, I. H. Solomon, H.-H. Kuo, J. Boucau, K. Bowman, U. D. Adhikari, M. L. Winkler, A. A. Mueller, T. Y.-T. Hsu, M. Desjardins, L. R. Baden, B. T. Chan, B. D. Walker, M. Lichterfeld, M. Brigl, D. S. Kwon, S. Kanjilal, E. T. Richardson, A. H. Jonsson, G. Alter, A. K. Barczak, W. P. Hanage, X. G. Yu, G. D. Gaiha, M. S. Seaman, M. Cernadas, J. Z. Li, Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **383**, 2291–2293 (2020). doi:10.1056/NEJMc2031364 Medline

31. V. A. Avanzato, M. J. Matson, S. N. Seifert, R. Pryce, B. N. Williamson, S. L. Anzick, K. Barbian, S. D. Judson, E. R. Fischer, C. Martens, T. A. Bowden, E. de Wit, F. X. Riedo, V. J. Munster, Case study: Prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. *Cell* **183**, 1901–1912.e9 (2020). doi:10.1016/j.cell.2020.10.049 Medline

32. A. J. Drummond, M. A. Suchard, Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* **8**, 114 (2010). doi:10.1186/1741-7007-8-114 Medline

33. M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, M. A. Suchard, Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013). doi:10.1093/molbev/mss265 Medline

34. F. N. Naveca *et al.*, COVID-19 epidemic in the Brazilian state of Amazonas was driven by long-term persistence of endemic SARS-CoV-2 lineages and the recent emergence of the new Variant of Concern P.1. *Nat. Portfolio* 10.21203/rs.3.rs-275494/v1 (2021).

35. M. Marks, P. Millat-Martinez, D. Ouchi, C. H. Roberts, A. Alemany, M. Corbacho-Monné, M. Ubals, A. Tobias, C. Tebé, E. Ballana, Q. Bassat, B. Baro, M. Vall-Mayans, C. G-Beiras, N. Prat, J. Ara, B. Clotet, O. Mitjà, Transmission of COVID-19 in 282 clusters in Catalonia, Spain: A cohort study. *Lancet Infect. Dis.* S1473-3099(20)30985-3 (2021). doi:10.1016/S1473-3099(20)30985-3 Medline

36. J. A. Hay, Kennedy-Shaffer, L., Kanjilal, S., Lipsitch, M., Mina, M. J, Estimating epidemiologic dynamics from single cross-sectional viral load distributions. medRxiv 20204222 [Preprint] 13 February 2021). doi:10.1101/2020.10.08.20204222.

37. M. Kidd, A. Richter, A. Best, N. Cumley, J. Mirza, B. Percival, M. Mayhew, O. Megram, F. Ashford, T. White, E. Moles-Garcia, L. Crawford, A. Bosworth, S. F. Atabani, T. Plant, A. McNally, S-variant SARS-CoV-2 lineage B1.1.7 is associated with significantly higher viral loads in samples tested by ThermoFisher TaqPath RT-qPCR. *J. Infect. Dis.* jiab082 (2021). doi:10.1093/infdis/jiab082 Medline

38. K. Stephen *et al*., Densely sampled viral trajectories suggest longer duration of acute infection with B.1.1.7 variant relative to non-B.1.1.7 SARS-CoV-2 (2021); https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37366884.

39. E. C. Sabino, L. F. Buss, M. P. S. Carvalho, C. A. Prete Jr., 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**, 452–455 (2021). doi:10.1016/S0140-6736(21)00183-5 Medline

40. H. J. T. Unwin, S. Mishra, V. C. Bradley, A. Gandy, T. A. Mellan, H. Coupland, J. Ish-Horowicz, M. A. C. Vollmer, C. Whittaker, S. L. Filippi, X. Xi, M. Monod, O. Ratmann, M. Hutchinson, F. Valka, H. Zhu, I. Hawryluk, P. Milton, K. E. C. Ainslie, M. Baguelin, A. Boonyasiri, N. F. Brazeau, L. Cattarino, Z. Cucunuba, G. Cuomo-Dannenburg, I. Dorigatti, O. D. Eales, J. W. Eaton, S. L. van Elsland, R. G. FitzJohn, K. A. M. Gaythorpe, W. Green, W. Hinsley, B. Jeffrey, E. Knock, D. J. Laydon, J. Lees, G. Nedjati-Gilani, P. Nouvellet, L. Okell, K. V. Parag, I. Siveroni, H. A. Thompson, P. Walker, C. E. Walters, O. J. Watson, L. K. Whittles, A. C. Ghani, N. M. Ferguson, S. Riley, C. A. Donnelly, S. Bhatt, S. Flaxman, State-level tracking of COVID-19 in the United States. *Nat. Commun.* **11**, 6189 (2020). doi:10.1038/s41467-020-19652-6 Medline

41. S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, M. Monod, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. C. Vollmer, N. M. Ferguson, L. C. Okell, S. Bhatt; Imperial College COVID-19

Response Team, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020). doi:10.1038/s41586-020-2405-7 Medline

42. James Scott, Axel Gandy, Swapnil Mishra, Juliette Unwin, Seth Flaxman, Samir Bhatt, Epidemia - Modeling of epidemics using hierarchical Bayesian models; https://imperialcollegelondon.github.io/epidemia/index.html.

43. V. Hall *et al*., Do antibody positive healthcare workers have lower SARS-CoV-2 infection rates than antibody negative healthcare workers? Large multi-centre prospective cohort study (the SIREN study), England: June to November 2020. bioRxiv 21249642 [Preprint] 15 January 2021). doi:10.1101/2021.01.13.21249642.

44. S. F. McGough, M. A. Johansson, M. Lipsitch, N. A. Menzies, Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLOS Comput. Biol.* **16**, e1007735 (2020). doi:10.1371/journal.pcbi.1007735 Medline

45. I. Hawryluk, H. Hoeltgebaum, S. Mishra, X. Miscouriduo, R. P. Schnekenberg, C. Whittaker, M. Vollmer, S. Flaxman, T. A. Mellaan, Gaussian Process Nowcasting: Application to COVID-19 Mortality Reporting. *arXiv* **arXiv:2102.11249**, (2021).

46. Agencia Brasil, Covid-19: Amazonas já transferiu 424 pacientes para outros estados (2021); https://agenciabrasil.ebc.com.br/saude/noticia/2021-02/covid-19-amazonas-ja-transferiu-424-pacientes-para-outros-estados.

47. S. L. Pond, S. D. Frost, S. V. Muse, HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679 (2005). doi:10.1093/bioinformatics/bti079 Medline

48. X. X. Qu, P. Hao, X.-J. Song, S.-M. Jiang, Y.-X. Liu, P.-G. Wang, X. Rao, H.-D. Song, S.-Y. Wang, Y. Zuo, A.-H. Zheng, M. Luo, H.-L. Wang, F. Deng, H.-Z. Wang, Z.-H. Hu, M.-X. Ding, G.-P. Zhao, H.-K. Deng, Identification of two critical amino acid residues of the severe acute respiratory syndrome coronavirus spike protein for its variation in zoonotic tropism transition via a double substitution strategy. *J. Biol. Chem.* **280**, 29588–29595 (2005). doi:10.1074/jbc.M500662200 Medline

49. H. D. Song, C.-C. Tu, G.-W. Zhang, S.-Y. Wang, K. Zheng, L.-C. Lei, Q.-X. Chen, Y.-W. Gao, H.-Q. Zhou, H. Xiang, H.-J. Zheng, S.-W. W. Chern, F. Cheng, C.-M. Pan, H. Xuan, S.-J. Chen, H.-M. Luo, D.-H. Zhou, Y.-F. Liu, J.-F. He, P.-Z. Qin, L.-H. Li, Y.-Q. Ren, W.-J. Liang, Y.-D. Yu, L. Anderson, M. Wang, R.-H. Xu, X.-W. Wu, H.-Y. Zheng, J.-D. Chen, G. Liang, Y. Gao, M. Liao, L. Fang, L.-Y. Jiang, H. Li, F. Chen, B. Di, L.-J. He, J.-Y. Lin, S. Tong, X. Kong, L. Du, P. Hao, H. Tang, A. Bernini, X.-J. Yu, O. Spiga, Z.-M. Guo, H.-Y. Pan, W.-Z. He, J.-C. Manuguerra, A. Fontanet, A. Danchin, N. Niccolai, Y.-X. Li, C.-I. Wu, G.-P. Zhao, Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2430–2435 (2005). doi:10.1073/pnas.0409608102 Medline

50. W. Li, C. Zhang, J. Sui, J. H. Kuhn, M. J. Moore, S. Luo, S.-K. Wong, I.-C. Huang, K. Xu, N. Vasilieva, A. Murakami, Y. He, W. A. Marasco, Y. Guan, H. Choe, M. Farzan, Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J.* **24**, 1634–1643 (2005). doi:10.1038/sj.emboj.7600640 Medline

51. D. Zhou, W. Dejnirattisai, P. Supasa, C. Liu, A. J. Mentzer, H. M. Ginn, Y. Zhao, H. M. E. Duyvesteyn, A. Tuekprakhon, R. Nutalai, B. Wang, G. C. Paesen, C. Lopez-Camacho, J.

Slon-Campos, B. Hallis, N. Coombes, K. Bewley, S. Charlton, T. S. Walter, D. Skelly, S. F. Lumley, C. Dold, R. Levin, T. Dong, A. J. Pollard, J. C. Knight, D. Crook, T. Lambe, E. Clutterbuck, S. Bibi, A. Flaxman, M. Bittaye, S. Belij-Rammerstorfer, S. Gilbert, W. James, M. W. Carroll, P. Klenerman, E. Barnes, S. J. Dunachie, E. E. Fry, J. Mongkolsapaya, J. Ren, D. I. Stuart, G. R. Screaton, Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* S0092-8674(21)00226-9 (2021). doi:10.1016/j.cell.2021.02.037 Medline

52. C. K. Wibmer, F. Ayres, T. Hermanus, M. Madzivhandila, P. Kgagudi, B. Oosthuysen, B. E. Lambson, T. de Oliveira, M. Vermeulen, K. van der Berg, T. Rossouw, M. Boswell, V. Ueckermann, S. Meiring, A. von Gottberg, C. Cohen, L. Morris, J. N. Bhiman, P. L. Moore, SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* (2021). doi:10.1038/s41591-021-01285-x Medline

53. 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. S. Moosa, R. J. Lessells, T. de Oliveira, A. Sigal; NGS-SA; COMMIT-KZN Team, Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature* (2021). doi:10.1038/s41586-021-03471-w Medline

54. L. Piccoli, Y.-J. Park, M. A. Tortorici, N. Czudnochowski, A. C. Walls, M. Beltramello, C. Silacci-Fregni, D. Pinto, L. E. Rosen, J. E. Bowen, O. J. Acton, S. Jaconi, B. Guarino, A. Minola, F. Zatta, N. Sprugasci, J. Bassi, A. Peter, A. De Marco, J. C. Nix, F. Mele, S. Jovic, B. F. Rodriguez, S. V. Gupta, F. Jin, G. Piumatti, G. Lo Presti, A. F. Pellanda, M. Biggiogero, M. Tarkowski, M. S. Pizzuto, E. Cameroni, C. Havenar-Daughton, M. Smithey, D. Hong, V. Lepori, E. Albanese, A. Ceschi, E. Bernasconi, L. Elzi, P. Ferrari, C. Garzoni, A. Riva, G. Snell, F. Sallusto, K. Fink, H. W. Virgin, A. Lanzavecchia, D. Corti, D. Veesler, Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183**, 1024–1042.e21 (2020). doi:10.1016/j.cell.2020.09.037 Medline

55. A. F. Z. Martins, A. P. Zavascki, P. L. Wink, F. C. Z. Volpato, F. L. Monteiro, C. Rosset, F. De-Paris, Á. K. Ramos, A. L. Barth, Detection of SARS-CoV-2 lineage P.1 in patients from a region with exponentially increasing hospitalisation rate, February 2021, Rio Grande do Sul, Southern Brazil. *Euro Surveill.* **26**, 2100276 (2021). doi:10.2807/1560-7917.ES.2021.26.12.2100276 Medline

56. A. O'Toole, V. Hill, O. G. Pybus, A. Watts, I. I. Bogoch, K. Khan, J. P. Messina, The COVID-19 Genomics UK (COG-UK) consortium, et al., "Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2" (2021) ; https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592.

57. TESSy. The European Surveillance System (TESSy) (2015); www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy.

58. COG-UK, COVID-19 Genomics UK Consortium (2020); www.cogconsortium.uk.

59. SPHERES, SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance; (2020); www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html.

60. L. du Plessis, J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, N. R. Faria, B. Jackson, N. J. Loman, Á. O'Toole, S. M. Nicholls, K. V. Parag, E. Scher, T. I. Vasylyeva, E. M. Volz, A. Watts, I. I. Bogoch, K. Khan, D. M. Aanensen, M. U. G. Kraemer, A. Rambaut, O. G. Pybus; COVID-19 Genomics UK (COG-UK) Consortium, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021). doi:10.1126/science.abf2946 Medline

61. E. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, Á. O'Toole, J. Southgate, R. Johnson, B. Jackson, F. F. Nascimento, S. M. Rey, S. M. Nicholls, R. M. Colquhoun, A. da Silva Filipe, J. Shepherd, D. J. Pascall, R. Shah, N. Jesudason, K. Li, R. Jarrett, N. Pacchiarini, M. Bull, L. Geidelberg, I. Siveroni, I. Goodfellow, N. J. Loman, O. G. Pybus, D. L. Robertson, E. C. Thomson, A. Rambaut, T. R. Connor, Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64–75.e11 (2021). doi:10.1016/j.cell.2020.11.020 Medline

62. B. Xu, M. U. G. Kraemer, Open COVID-19 Data Curation Group, Open access epidemiological data from the COVID-19 outbreak. *Lancet Infect. Dis.* **20**, 534 (2020). doi:10.1016/S1473-3099(20)30119-5 Medline

63. Global.health (2021); https://global.health.

64. GitHub Repository; https://github.com/CADDE-CENTRE.

65. S., CADDE CENTRE, C. Whittaker, CADDE-CENTRE/Novel-SARS-CoV-2-P1-Lineage-in-Brazil: Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil (peer-review version out soon). Zenodo (2021); doi:10.5281/zenodo.4676853.

66. Fundação de Vigilância em Saúde do Amazonas, Amazonas, "Dados epidemiológicos e financeiros das ações de combate à COVID-19. Publicações (2021); www.fvs.am.gov.br/publicacoes.

67. SRAG 2020 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19 - Open Data; https://opendatasus.saude.gov.br/dataset/bd-srag-2020.

68. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017). doi:10.2807/1560-7917.ES.2017.22.13.30494 Medline

69. SRAG 2021 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19 - Open Data; https://opendatasus.saude.gov.br/dataset/bd-srag-2021.

70. E. Farfour, P. Lesprit, B. Visseaux, T. Pascreau, E. Jolly, N. Houhou, L. Mazaux, M. Asso-Bonnet, M. Vasse; SARS-CoV-2 Foch Hospital study group, The Allplex 2019-nCoV (Seegene) assay: Which performances are for SARS-CoV-2 infection diagnosis? *Eur. J. Clin. Microbiol. Infect. Dis.* **39**, 1997–2000 (2020). doi:10.1007/s10096-020-03930-8 Medline

71. F. M. Liotti, G. Menchinelli, S. Marchetti, G. A. Morandotti, M. Sanguinetti, B. Posteraro, P. Cattani, Evaluation of three commercial assays for SARS-CoV-2 molecular detection in upper respiratory tract samples. *Eur. J. Clin. Microbiol. Infect. Dis.* **40**, 269–277 (2021). doi:10.1007/s10096-020-04025-0 Medline

72. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

73. I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, D. Marshall, Tablet—Next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010). doi:10.1093/bioinformatics/btp666 Medline

74. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

75. M. C. dos Santos, E. C. Sousa Jr., J. A. Ferreira, S. P. Silva, M. P. C. Souza, J. F. Cardoso, A. M. Silva, L. S. Barbagelata, W. D. Chagas Jr., J. L. Ferreira, E. M. A. Souza, P. L. A. Vilaca, J. C. S. Alves, M. C. Abreu, P. S. Lobo, F. S. Santos, A. A. P. Lima, C. M. Bragagnolo, L. S. Soares, P. S. M. Almeida, D. S. Oliveira, C. K. N. Amorim, I. B. Costa, D. M. Teixeira, E. T. Penha Jr., D. A. M. Bezerra, J. A. M. Siqueira, F. N. Tavares, F. B. Freitas, J. T. N. Rodrigues, J. Mazaro, A. S. Costa, M. S. P. Cavalcante, M. S. Silva, I. A. Silva, G. A. L. Borges, L. G. Lima, H. L. S. Ferreira, M. T. Livorati, A. L. Abreu, A. C. Medeiros, H. R. Resque, R. C. M. Sousa, G. M. R. Viana, Molecular epidemiology to understand the SARS-CoV-2 emergence in the Brazilian Amazon. bioRxiv 20184523 [Preprint] 7 September 2020). doi:10.1101/2020.09.04.20184523.

76. J. Xavier, M. Giovanetti, T. Adelino, V. Fonseca, A. V. Barbosa da Costa, A. A. Ribeiro, K. N. Felicio, C. G. Duarte, M. V. Ferreira Silva, Á. Salgado, M. T. Lima, R. de Jesus, A. Fabri, C. F. Soares Zoboli, T. G. Souza Santos, F. Iani, M. Ciccozzi, A. M. Bispo de Filippis, M. A. M. Teixeira de Siqueira, A. L. de Abreu, V. de Azevedo, D. B. Ramalho, C. F. Campelo de Albuquerque, T. de Oliveira, E. C. Holmes, J. Lourenço, L. C. Junior Alcantara, M. A. Assunção Oliveira, The ongoing COVID-19 epidemic in Minas Gerais, Brazil: Insights from epidemiological data and SARS-CoV-2 whole genome sequencing. *Emerg. Microbes Infect.* **9**, 1824–1834 (2020). doi:10.1080/22221751.2020.1803146 Medline

77. M. H. S. Paiva, D. R. D. Guedes, C. Docena, M. F. Bezerra, F. Z. Dezordi, L. C. Machado, L. Krokovsky, E. Helvecio, A. F. da Silva, L. R. S. Vasconcelos, A. M. Rezende, S. J. R. da Silva, K. G. D. S. Sales, B. S. L. F. de Sá, D. L. da Cruz, C. E. Cavalcanti, A. M. Neto, C. T. A. da Silva, R. P. G. Mendes, M. A. L. da Silva, T. Gräf, P. C. Resende, G. Bello, M. D. S. Barros, W. R. C. do Nascimento, R. M. L. Arcoverde, L. C. A. Bezerra, S. P. B. Filho, C. F. J. Ayres, G. L. Wallau, Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil. *Viruses* **12**, E1414 (2020). 10.3390/v12121414 Medline

78. J. D. Siqueira, L. R. Goes, B. M. Alves, P. S. de Carvalho, C. Cicala, J. Arthos, J. P. B. Viola, A. C. de Melo, M. A. Soares, SARS-CoV-2 genomic and quasispecies analyses in cancer patients reveal relaxed intrahost virus evolution. *bioRxiv* 2020.08.26.267831 doi:10.1101/2020.08.26.267831 (2020). Medline

79. C. M. Voloch, R. da Silva F Jr, L. G. P. de Almeida, C. C. Cardoso, O. J. Brustolini, A. L. Gerber, A. P. de C. Guimarães, D. Mariani, R. M. da Costa, O. C. Ferreira Jr, A. C.

Cavalcanti, T. S. Frauches, C. M. B. de Mello, R. M. Galliez, D. S. Faffe, T. M. P. P. Castiñeiras, A. Tanuri, A. T. R. de Vasconcelos, C.-U. Workgroup, LNCC-Workgroup, Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J. Virol.* 10.1128/JVI.00119-21 (2020).

80. R. da Silva Francisco, L. Felipe Benites, A. P. Lamarca, L. G. P. de Almeida, A. W. Hansen, J. S. Gularte, M. Demoliner, A. L. Gerber, A. P. de C Guimarães, A. K. E. Antunes, F. H. Heldt, L. Mallmann, B. Hermann, A. L. Ziulkoski, V. Goes, K. Schallenberger, M. Fillipi, F. Pereira, M. N. Weber, P. R. de Almeida, J. D. Fleck, A. T. R. Vasconcelos, F. R. Spilki, Pervasive transmission of E484K and emergence of VUI-NP13L with evidence of SARS-CoV-2 co-infection events by two different lineages in Rio Grande do Sul, Brazil. *Virus Res.* **296**, 198345 (2021).

81. V. B. Franceschi, G. D. Caldana, A. de Menezes Mayer, G. B. Cybis, C. A. M. Neves, P. A. G. Ferrareze, M. Demoliner, P. R. de Almeida, J. S. Gularte, A. W. Hansen, M. N. Weber, J. D. Fleck, R. A. Zimerman, L. Kmetzsch, F. R. Spilki, C. E. Thompson, Genomic Epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. medRxiv 21249906 [Preprint] 26 January 2021. doi:10.1101/2021.01.21.21249906.

82. 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**, 1530–1534 (2020). doi:10.1093/molbev/msaa015 Medline

83. T. H. Jukes, C. R. Cantor, in Mammalian Protein Metabolism, H. N. Munro, Ed. (Academic Press, 1969), pp. 21–132.

84. R Core Team, *R*: A language and environment for statistical computing (R Core Team, Vienna, Austria, 2013).

85. M. Hasegawa, H. Kishino, T. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985). doi:10.1007/BF02101694 Medline

86. Z. Yang, Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994). doi:10.1007/BF00160154 Medline

87. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018). doi:10.1093/ve/vey016 Medline

88. D. L. Ayres, A. Darling, D. J. Zwickl, P. Beerli, M. T. Holder, P. O. Lewis, J. P. Huelsenbeck, F. Ronquist, D. L. Swofford, M. P. Cummings, A. Rambaut, M. A. Suchard, BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012). doi:10.1093/sysbio/syr100 Medline

89. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018). doi:10.1093/sysbio/syy032 Medline

90. M. Worobey, T. D. Watts, R. A. McKay, M. A. Suchard, T. Granade, D. E. Teuwen, B. A. Koblin, W. Heneine, P. Lemey, H. W. Jaffe, 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* **539**, 98–101 (2016). doi:10.1038/nature19827 Medline

91. M. Bletsa, M. A. Suchard, X. Ji, S. Gryseels, B. Vrancken, G. Baele, M. Worobey, P. Lemey, Divergence dating using mixed effects clock modelling: An application to HIV-1. *Virus Evol.* **5**, vez036 (2019). doi:10.1093/ve/vez036 Medline

92. B. Murrell, J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, S. L. Kosakovsky Pond, Detecting individual sites subject to episodic diversifying selection. *PLOS Genet.* **8**, e1002764 (2012). doi:10.1371/journal.pgen.1002764 Medline

93. O. A. MacLean, S. Lytras, S. Weaver, J. B. Singer, M. F. Boni, P. Lemey, S. L. Kosakovsky Pond, D. L. Robertson, Evidence of significant natural selection in the evolution of SARS-CoV-2 in bats, not humans. *bioRxiv* 2020.05.28.122366 doi:10.1101/2020.05.28.122366 (2020). Medline

94. A. G. Wrobel, D. J. Benton, P. Xu, C. Roustan, S. R. Martin, P. B. Rosenthal, J. J. Skehel, S. J. Gamblin, SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.* **27**, 763–767 (2020). doi:10.1038/s41594-020-0468-7 Medline

95. L. Schrödinger. *The PyMOL Molecular Graphics*, version 2.0 (2015).

96. P. S. Peixoto, D. Marcondes, C. Peixoto, S. M. Oliva, Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil. *PLOS ONE* **15**, e0235732 (2020). doi:10.1371/journal.pone.0235732 Medline

97. T. Mellan, H. Hoeltgebaum, S. Mishra, C. Whittaker, R. P. Schnekenberg, A. Gandy, H. J. H. Unwin, M. A. Vollmer, H. Coupland, I. Hawryluk, N. R. Faria, J. Vesga, H. Zhu, M. Hutchinson, O. Ratmann, M. Monod, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, N. Brazeau, G. Charles, L. V. Cooper, Z. Cucunuba, G. Cuomo-Dannenburg, A. Dighe, B. Djaafara, J. Eaton, S. L. van Elsland, R. FitzJohn, K. Fraser, K. Gaythorpe, W. Green, S. Hayes, N. Imai, B. Jeffrey, E. Knock, D. Laydon, J. Lees, T. Mangal, A. Mousa, G. Nedjati-Gilani, P. Nouvellet, D. Olivera, K. V. Parag, M. Pickles, H. A. Thompson, R. Verity, C. Walters, H. Wang, Y. Wang, O. J. Watson, L. Whittles, X. Xi, L. Okell, I. Dorigatti, P. Walker, A. Ghani, S. Riley, N. Ferguson, C. A. Donnelly, S. Flaxman, S. Bhatt, Report 21: Estimating COVID-19 cases and reproduction number in Brazil. (Imperial College London, 2020); doi:10.25561/78872.

98. R. Bellman, T. Harris, *On Age-Dependent Binary Branching Processes*. In: The Annals of Mathematics (1952).

99. W. Feller, On the Integral Equation of Renewal Theory. *Ann. Math. Stat.* **12**, 243–267 (1941). doi:10.1214/aoms/1177731708

100. I. Hawryluk, T. A. Mellan, H. Hoeltgebaum, S. Mishra, R. P. Schnekenberg, C. Whittaker, H. Zhu, A. Gandy, C. A. Donnelly, S. Flaxman, S. Bhatt, Inference of COVID-19 epidemiological distributions from Brazilian hospital data. *J. R. Soc. Interface* **17**, 20200596 (2020). doi:10.1098/rsif.2020.0596 Medline

101. J. Hellewell, T. W. Russell, R. Beale, G. Kelly, C. Houlihan, E. Nastouli, A. J. Kucharski, SAFER Investigators, Field Study Team, Crick COVID-19 Consortium, et al, Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. medRxiv 20229948 [Preprint] 24 December 2020. doi:10.1101/2020.11.24.20229948.

102. B. Borremans, A. Gamble, K. C. Prager, S. K. Helman, A. M. McClain, C. Cox, V. Savage, J. O. Lloyd-Smith, Quantifying antibody kinetics and RNA detection during early-phase SARS-CoV-2 infection by time since symptom onset. *eLife* **9**, e60122 (2020). doi:10.7554/eLife.60122 Medline

Supplementary Methods

Epidemiological Context

The HCFMUSP is composed of nine medical specialty institutes: Central Institute (ICHC), Heart Institute (InCor), Cancer Institute (ICESP), Children's Institute (ICr), Radiology Institute (InRad), Psychiatry Institute (IPq), Physical Medicine and Rehabilitation Institute (IMREA), Orthopaedics and Traumatology Institute (IOT), and Long-term auxiliary Hospital (HAS).

Institute B became an institute dedicated to the care of COVID patients. Patients began to be transferred to other institutes and structures were created to receive 300 ICU beds and 300 ward beds, with approximately 6000 HCW. A crisis committee was created and external assistance teams of doctors, nurses and physiotherapists came to help in patient care. All were trained in the correct use of PPE, which included private clothing, N95 masks, gloves and aprons, hats and face shields, and clothing and de-dressing techniques.

Personal protective equipment (PPE) was made available to all HCW. HCW who provided direct patient care to COVID-19 patients wore N95 respirators and scrubs during their entire shifts. When examining or touching patients they added disposable gloves and a gown. During aerosol-generating procedures, they used N95 respirators, a gown, gloves, and eye protection (face shield or goggles). HCW used the same N95 respirator between patients and these were reused by the same HCW for seven shifts or until damaged or soiled. The cleaning staff wore N95 respirators during their entire shift. HCW were trained to don and doff PPE in face-to-face sessions and with videos and posters. All symptomatic HCW were evaluated at a dedicated health service (located in a separate building) and, if indicated, oro-nasopharyngeal swabs were collected. If COVID-19 was confirmed, the HCW received paid leave for 14 days from the onset of symptoms. Clinical and epidemiological data collection (age, sex, home address, occupation, unit of work within the hospital, date of onset of symptoms, symptoms, need for

hospitalization, and clinical outcome) involved the collaboration of several teams that matched sample identification numbers to medical records from patients and health workers on each institute's electronic system.

Supplementary Panel 1 – COVID-19 response measures taken by each institute from the HCFMUSP according to epidemiological week.

Epil week	**Institute A**	**Institute B**	**Institute C**
Week 10 (07/03/2020 - 13/03/2020)	COVID and non-COVID areas; HCW were not allowed to move between areas; For aerosol forming procedures: N95 + glasses + face shield; Triage of patients, HCW and carers (temperature and symptoms)		
Week 11 (14/03/2020 - 20/03/2020)	Restriction of in person events and meetings; No visitors allowed in COVID areas and limited in all other areas;	Entrance areas, elevators, pantries, cafeterias were separated to ensure the safety of health professionals.	
Week 12 21/03/2020 - 27/03/2020)		Outpatient care was suspended	COVID and non-COVID areas
Week 13 (28/03/2020 - 06/04/2020)		Became COVID-only. Guideline use of PPE available for all HCW	
Week 15 (04/04/2020 - 10/04/2020)	Mandatory masking. Surgical masks for all HCW; administrative workers could wear cloth masks.	Focus on identifying symptomatic healthcare workers and immediate leave	HCW servicing patients - surgical masks in inpatient wards and N95 in ICUs;1 visitor per patient per day
Week 16 (11/04/2020 - 17/04/2020)		HCW not allowed to transit between institutes. Professionals were dedicated exclusively to Institute B	

Week 17 (18/04/2020 - 24/04/2020)		Training on the use of PPE for multidisciplinary teams such as nutritionists, hygiene staff, maintenance engineering, psychologists, social workers	Universal masking mandatory (surgical mask)
Week 18 (25/04/2020 - 01/05/2020)	N95 for all critical emergency areas, ICUs and surgical centre	Training on the use of PPE for administrative personnel due to an outbreak	5 Covid-19 transmission measures; Daily audits until 10/05
Week 19 (02/05/2020 - 08/05/2020)		Mandatory use of surgical masks for all professionals, including administrative personnel.	

SARS-CoV-2 genome amplification

For SARS-CoV-2 whole-genome sequencing, we used a tilling-amplicon multiplex PCR technique as previously described (3–5). First, the cDNA was synthesized from positive RNA samples using the ProtoScript II First-Strand cDNA synthesis kit (New England Biolabs, UK) and random primers or SuperScriptIV First-Strand Synthesis System (Thermo Fisher Scientific, USA). Subsequently, cDNA was subjected to amplification using the V2 ARTIC scheme (https://artic.network/ncov-2019) and Q5 High-Fidelity DNA polymerase (New England Biolabs, UK). After amplification, the AmpureXP purification beads (Beckman Coulter, United Kingdom) were used for product purification and the Qubit dsDNA High Sensitivity Assay on the Qubit 3.0 instrument (Life Technologies, USA) to quantify the amplicons.

Library preparation and whole-genome sequencing

Sequencing libraries were prepared using a total input of 100ng. The normalized amplicons were submitted to barcode ligation using the EXP-NBD 104 (1–12) and EXP NBD 114 (13–24) Native Barcoding Kits (Oxford Nanopore Technologies, UK). Sequencing libraries were generated using the SQK-LSK109 Kit (Oxford Nanopore Technologies, UK). Finally, 20ng of the library, containing 23

samples and one negative control, were loaded onto an R9.4.1 flow-cell on the MinION device and sequenced using MinKNOW 1.15.1 (Oxford Nanopore Technologies, UK).

Bioinformatic analysis

Guppy software v2.2.7 (Oxford Nanopore Technologies, UK) was used to basecall, demultiplex, and trim the FAST5 files. FASTQ files were mapped to the reference genome of SARS-CoV-2 isolate Wuhan-Hu 1 (GenBank Accession Number MN908947) using minimap2 v.2.28.0 to generate the consensus genomes and SAMtools to convert to a sorted BAM file (Li et al., 2009).

Length filtering and the quality test was performed for each barcode using artic guppyplex (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html). The genome statistics were obtained from SAMtools and Tablet viewer (Milne et al., 2010). To recover consensus sequences, called variants were detected with Nanopolish. Genome regions with a depth of <20-fold were not included in final consensus sequences, and these positions are represented with N characters. Runs with negative control presenting any contamination were discarded.

Collation of genomic datasets and sequence quality control

SARS-CoV-2 sequences from Brazil with collection date up to the 20th May 2020 (oldest collection date in our HCW dataset) (n=1860) were downloaded from GISAID (6–8) and appended to a previously described global dataset of 1,182 viral genomes (4). As previously described (4), we further filtered down our dataset by maintaining only sequences with at least 75% consensus sequence coverage. The resulting dataset was aligned to the reference NC_045512.2 using MAFFT v 7.450 (9) and manually edited using AliView. 3' and 5' untranslated regions of each sequence were discarded.

A Maximum likelihood tree was inferred using IQ-TREE v.2.0 (10) under the best substitution model as determined by ModelFinder (34) implemented in the IQ-TREE pipeline. TempEst v.1.5.3 (35) analyses and visual inspection of the alignment in Aliview were used to identify and remove sequences with unusual divergence for a given date of collection and/or long stretches of polymorphisms. Our final

dataset consisted of 2,550 sequences, including 340 sequences from HCFMUSP, 67 novel genomes and 273 previous GISAID submissions from our group (dataset 1). Pangolin version V3.1.11 (36) was used for lineage assignment. For accurate cluster identification, sequences with >90% coverage (n=2259) were maintained for subsequent phylogenetic analysis and initial cluster identification, including 234 sequences from this study. Finally, for Bayesian phylogenetic analysis, dataset B was subsampled to include 200 randomly selected sequences from other countries, all sequences from Sao Paulo state (n=407), and all sequences from this study with coverage >90% (n=234) (dataset 3, n=841 sequences). Only sequences from Sao Paulo state were maintained given that all sequences from this study clustered amongst them.

Phylogenetic and Phylogeographic analysis

Maximum likelihood trees for all datasets were inferred using IQ-TREE v.2.0, with the best substitution model determined by ModelFinder implemented in the IQ-TREE pipeline: GTR+F+R2 (Datasets A and B), GTR+F+R3 (dataset 3). Sequences from this study were scanned for recombination using all methods available on RDP4 (37) and no recombination signal was found. Bayesian time-rooted phylogenies were estimated using Beast v1.10.4 (38) and BEAGLE (39), under an HKY+Γ nucleotide substitution model, a strict molecular clock and an exponential growth coalescent tree prior (40); (27). For population size, a log-normal prior with mu=1 and sigma=0 was used and for growth rate, a la place prior with mean=0 and scale=100 was set. Analyses were run in triplicates for 200 million Markov Chain Monte Carlo (MCMC) steps. Run convergence was assessed using Tracer v.1.7.1 (41). Log and tree files were combined after removal of 10% burn-in from each run using LogCombiner v1.10.4 and summary Maximum Clade Credibility trees were generated from the combined tree files using TreeAnnotator v1.10.4. LogCombiner v1.10.4 was also used to generate a resampled distribution of 1,000 from the combined tree files, which was used for subsequent phylogeographic analysis.

To understand the dynamics of SARS-CoV-2 spread across the institutes we used 4 different discrete trait schemes. For the first trait, location (k=5), sequences were assigned one of five locations (k=5): (I)

Institute A, (II) Institute B, (III) Institute C, (IV) São Paulo (other sequences from São Paulo state) or (V) Other (international sequences). For the discretization scheme two (k=3), location2, all institute sequences were assigned one single location (I) HC; while other sequences were assigned as (II) São Paulo and (III) Other To incorporate the transmission dynamics between HCW and patients, discretization scheme 3 (k=4) assigned sequences to either (I) HCW, (II) patient, (III) São Paulo, and (IV) Other (k=4). Finally, discretization scheme 4 (k=8) accounted for both the institute and HCW/Patient trait information by assigning sequences to either (I) Institute A_HCW (II) Institute A_patient, (III) Institute B_HCW, (IV) Institute B_patient, (V) Institute C_HCW, (VI) Institute C_Patient, (VII) São Paulo and (VIII) Other. The number of migration events between each location considered, as well as the total number of imports and exports for each location, were estimated using a Markov Jumps count approach (42) implemented on Beast v1.10.4, under an asymmetric CTMC model for discrete trait reconstruction (43). Migration rates were inferred on a separate run under the same asymmetric model with a Bayesian Stochastic Search Variable Selection procedure (BSSVS) (44), which identifies and limits the number of rates to only one that can actually explain the diffusion process. The proportion (%) of imports for each institute was estimated by considering the maximum number of imports for each trait as equal to the total number of sequences for each trait included in the analysis. By considering imports as any transition from another trait, the remaining proportion, adding up to 100%, would be the proportion of transmission happening within a particular trait (non-imports), and thus representing transmission within each institute or HCW/Patient. To estimate the expected proportion of imports and transitions in a scenario in which clustering was not related to the traits, locations for institute sequences in schemes 1 and 4 were randomly reordered ten times and ten independent simulations were run for each scheme, under the same evolutionary models described above. The average distribution for imports for each trait was estimated from the results of the ten simulations.

Cluster analysis

Considering the relatively low evolutionary rate of SARS-CoV-2, which accumulates 2-3 mutations a month (27),(45) cluster analysis was performed on sequences with >90% coverage, in order to decrease the chances of incorrect assignment of sequences to transmission clusters. Initial identification of clusters was performed on dataset 2 under an ML phylogeny run on IQ-TREE v.2 (see methods). Clusters were confirmed on dataset 3 under both ML, Bayesian time-rooted trees and Bayesian trait-referenced time-rooted trees. Clusters were defined according to the content of HC sequences (sequences from this study) and according to the statistical support obtained from phylogenetic analysis. HC clusters were considered when >75% of the sequences were from HC, when they were supported by a minimum SH-like approximate likelihood ratio test (SH-aLRT) support of 75, minimum fast bootstrap support of 90, a minimum node posterior support of 0.9, and had at least one defining mutation separating clustered sequences from the nearest sequences in the phylogenetic tree. Statistical support thresholds were defined considering the support thresholds recommended by IQ-TREE and the agreement between the different support statistics across different phylogenetic methods.

The strength of the epidemiological link between clustered sequences was assessed using both hospital-associated metadata and the geocoded residential addresses for patients and HCW. An "epidemiological link" was asssigned when individuals worked/were hospitalized in the same ward/floor at the time of symptom onset or when it involved HCWs from the same specialty/division. A "possible epidemiological link" was defined as being hospitalized/working in the same institute at the time of symptom onset, but not necessarily in the same ward/floor or being from the same division/specialty. Finally, an "unclear epidemiological link" was defined as individuals who worked/were hospitalized at different institutes at the time of symptom onset and no clear epidemiological link could be established.

Compartmentalization analysis

To investigate the association of specific traits and genetic diversity of SARS-CoV-2 in our dataset, we used Simmond's Association Index implemented in the Hypothesis Testing Using Phylogenies (HYPHY) (46). AI calculates an association index for genetic diversity according to different

compartments (traits). It assesses population structure by weighting the contribution of each node while also running a bootstrap to provide support for the association index. In our analysis, we considered three separate sets of compartments: Institutes (Institute A, Institute B, Institute C), occupations (doctor, nurse, administration, other, and patient), and Patient/HCW. Doctors and nurses comprise all different levels of training of medical doctors and nursing professionals (including technicians). Analysis was performed in two datasets: (I) all HC sequences >90% coverage (n=234) and (I) clustered sequences only (n=73). Runs were performed under the following conditions: ten relabellings (default), 1000 bootstraps, and a significance threshold of <⅔ (default).

Statistical analysis

Descriptive analyses are shown as arithmetic means, median, and range. To assess traits that may affect the clustering of genetic sequences, we performed binomial logistic regression analyses, having clustered or non-clustered as the outcome variable, and analyses were controlled for sex and age. For model 1, traits location (Institute A, Institute B, and Institute C) and HCW/patient were used. Baseline variables were Institute B and patient. For model 2, one single trait incorporating both location and HCW/patient was used: HCW.Institute A, patient.Institute A, HCW.Institute B, patient.Institute B, HCW.Institute C, patient.Institute C. Baseline variable was patient.Institute B. For model 3, traits location (Institute A, Institute B, and Institute C) and occupation (administration, doctor, medical resident, nurse, nurse technician, others, and patient. Baseline variables were Institute B and patient. Finally, model 4 contained one single trait with all possible combinations of location and occupations used in model 3 (k=21). The baseline variable was patient. Institute B. Results were reported as the odds ratio (OR) over the baseline variables and p values <0.05 were considered statistically significant. For the household geographical distances analysis, a Mann-Whitney U test was performed in R Studio 1.2.1335.

Data sharing

Raw virus reads, and consensus sequences generated in this study can be found at https://github.com/CADDE-CENTRE/… XMLs GISAID IDs are available in Table S3.

Supplementary Figures

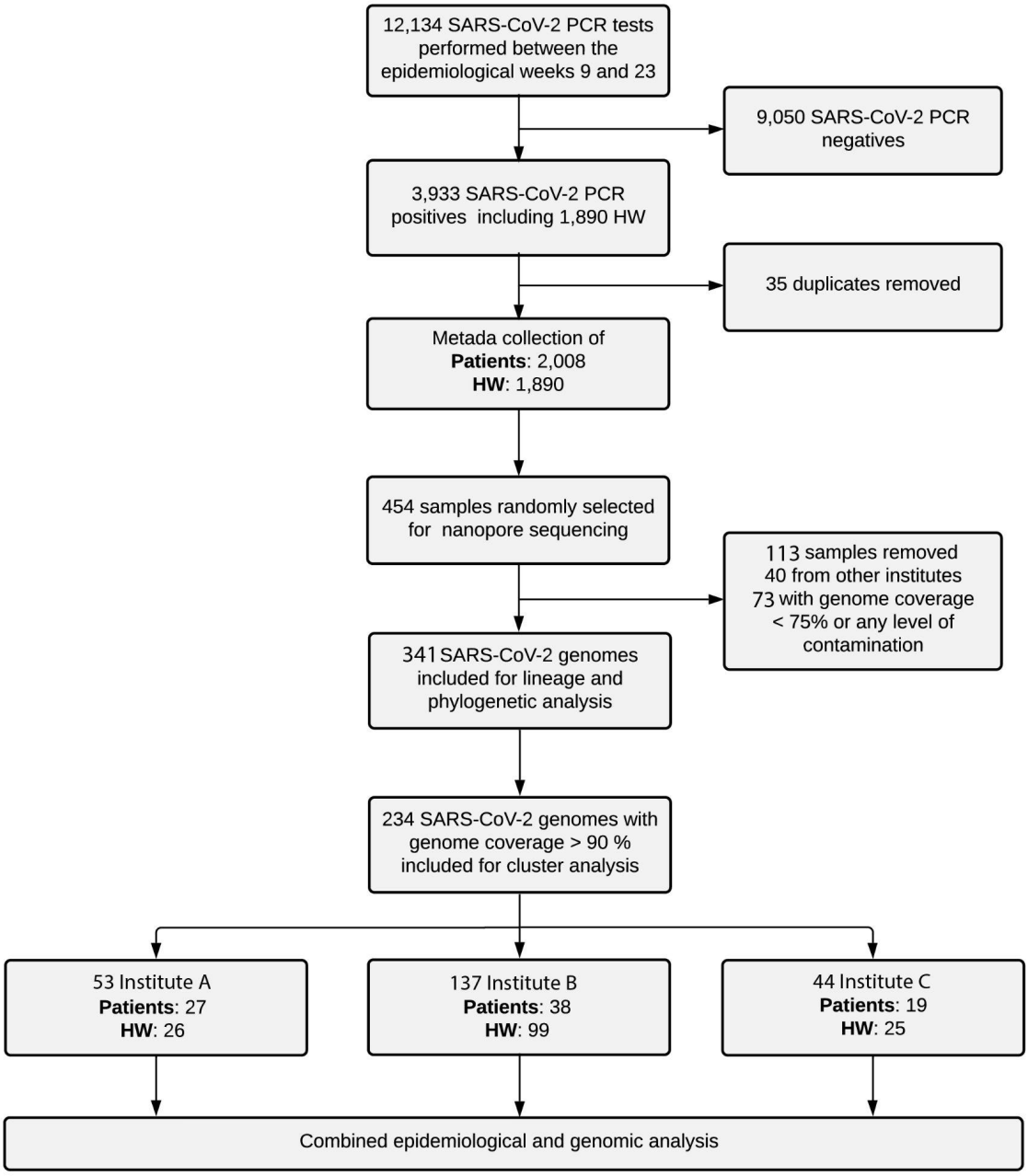


Figure S1. **Fluxogram contains information on the study design**. Out of 3,933 SARS-CoV-2 positive individuals from all HC institutes, minimum metadata was successfully collected for a total of 3,898. Of these, 454 samples from Institute B, Institute A, and Institute Cwere randomly selected for nanopore genome sequencing using the ARTIC protocol. 340 samples passed our quality control analysis and were submitted to lineage assignment using Pangolin COVID-19 Lineage assigner. 234 samples with coverage >90% were used for cluster analysis. For metadata collection and genome sequencing and quality control, see methods.

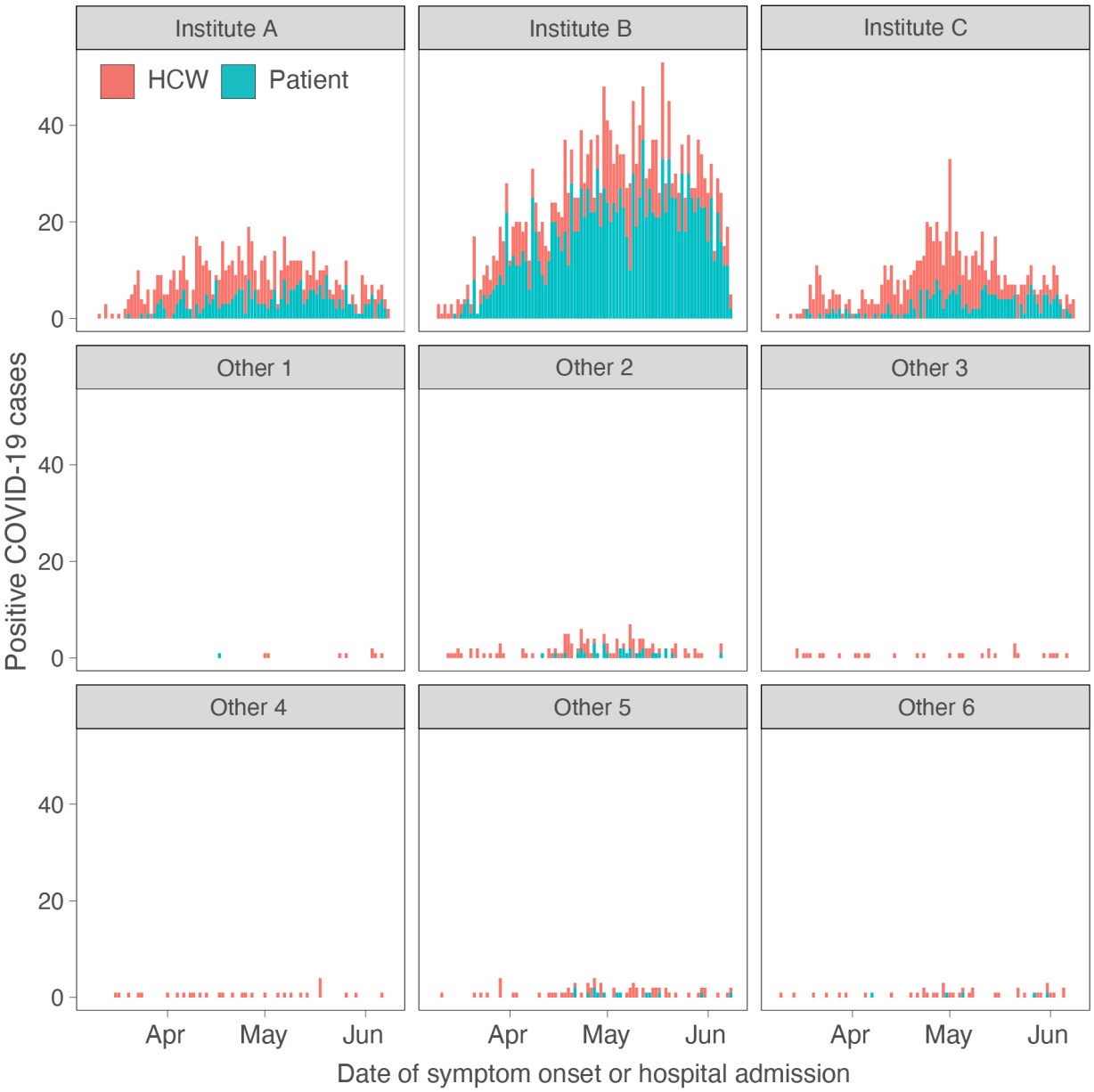


Figure S2. Time series of 3,898 HC COVID-19-positive cases by the institute of origin. Colors distinguish patients (blue) from HCW (red). The date of symptom onset was used for health workers and patients who were hospitalized before symptom onset. For community-acquired patients, the date of hospitalization was used (see methods). Institute A, Institute B, and Institute Cconcentrate 91.8% of all HC reported cases.

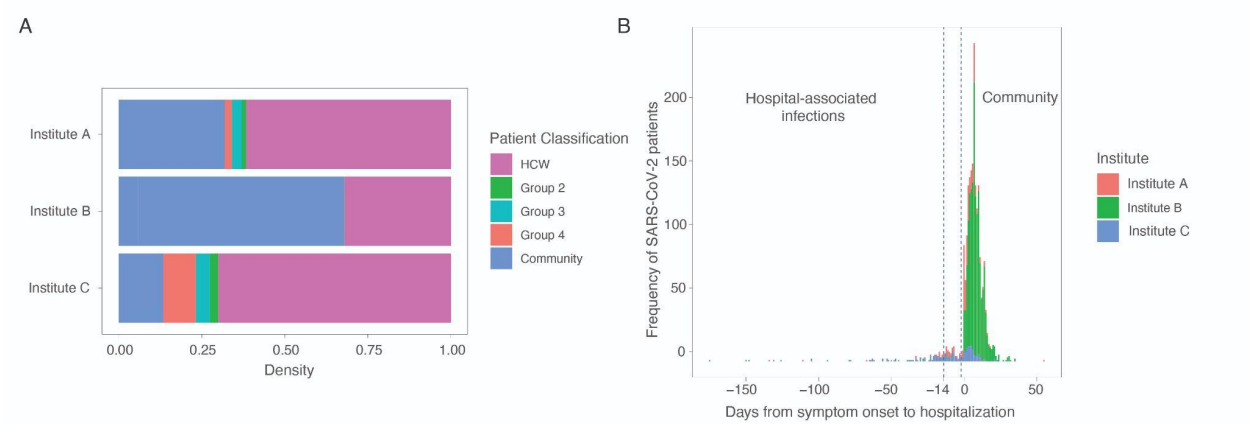


Fig S3. Classification of Covid-19 positive patients according to the gap between symptom onset and hospitalization. (A) Proportion of patients belonging to each category per institute. (B) Distribution of COVID-19 positive patients according to the time gap (in days) between hospitalization and symptom onset. Negative values mean that patients were hospitalized prior to symptom onset, while positive values mean that patients were hospitalized after symptom onset. Dotted lines mark patients from groups 3 (symptom onset >2 and <8 days after hospitalization) and 4 (symptom onset ≥8 and <15 days after hospitalization), to which hospital-associated infection is suspected but inconclusive.

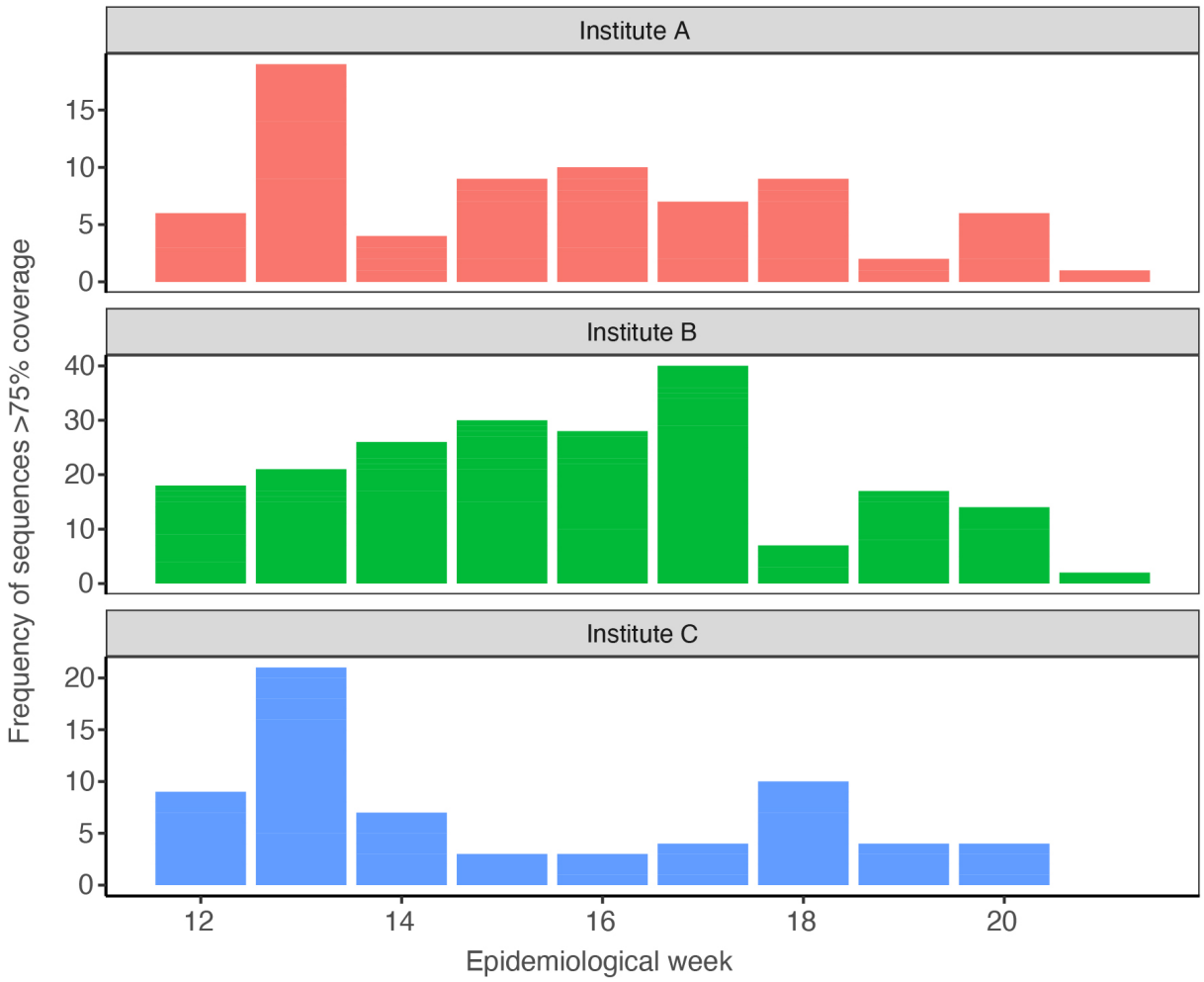


Fig S4. Time series of COVID-19 sequences with coverage >75% per week per institute of origin.

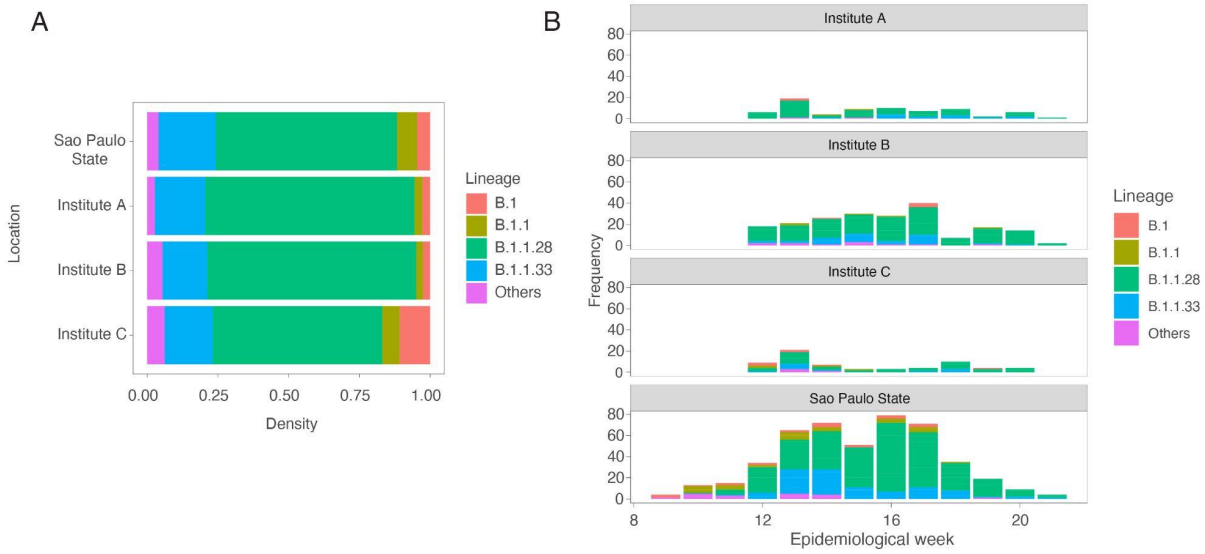


Fig. S5. Genetic diversity of HC SARS-CoV-2 samples. (A) Proportion of SARS-CoV-2 lineages per institute (n=340 sequences, >75% coverage) and São Paulo state (n=471 sequences). Lineage assignment was performed using Pangolin COVID-19 Lineage assigner. (B) Time series of lineages per institute/São Paulo per epidemiological week.

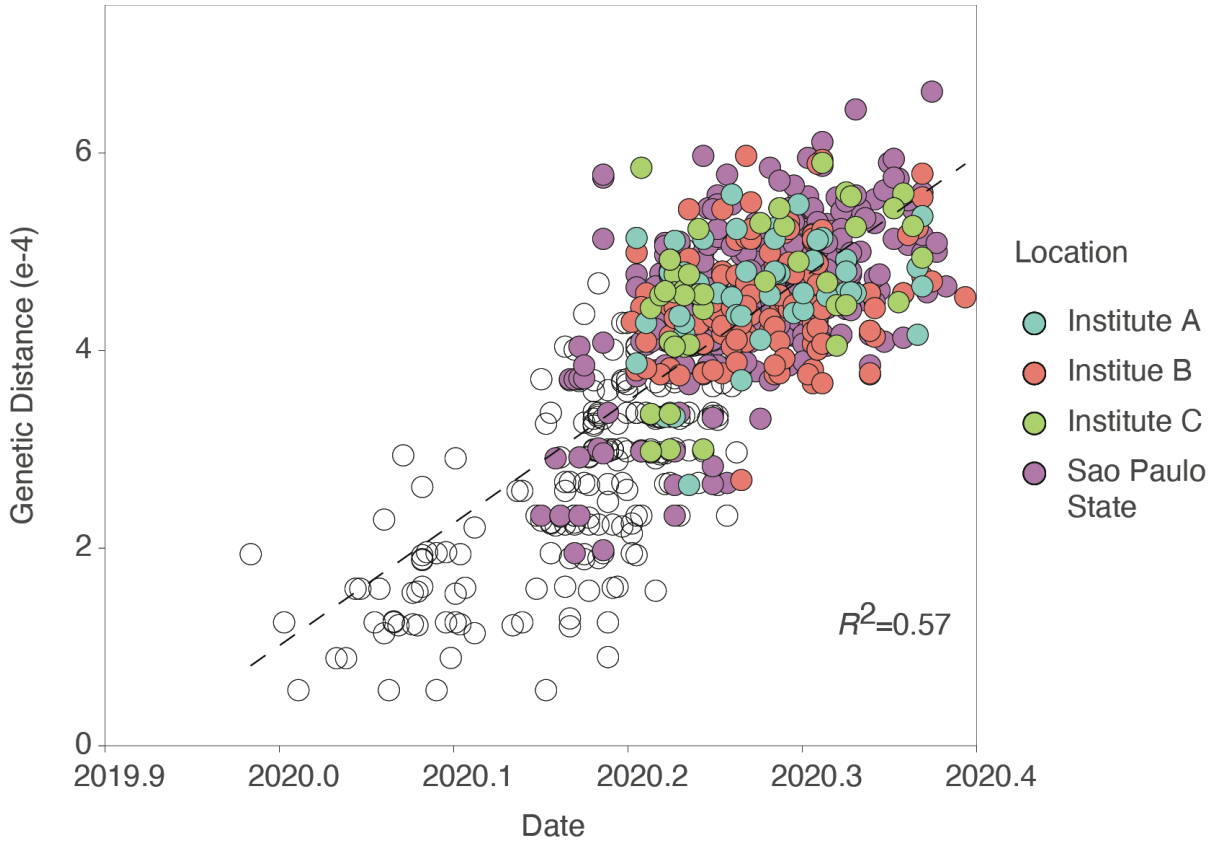


Fig S6. Regression of root-to-tip genetic diversity and sample collection dates for Dataset C. Dataset C is composed of 841 sequences, 234 sequences from HC, 407 sequences from the State of São Paulo in Brazil, and 200 international sequences (see methods). Circles are colored according to the location of collection: HC institutes (Institute A, Institute C, and Institute B), São Paulo State (SP, purple), international (white).

Fig. S7 (separate PDF file)
Detailed annotated maximum clade phylogenetic tree. Time-resolved maximum clade credibility phylogeny of 1,182 SARS-CoV-2 sequences, 490 from Brazil (red) and 692 from outside Brazil (blue). The largest Brazilian clusters are highlighted in grey (Clade 1, Clade 2 and Clade 3). States are defined by their 2-letter ISO 3166-1 codes. The MCC tree can be found at our Dryad repository (see Data Availability).

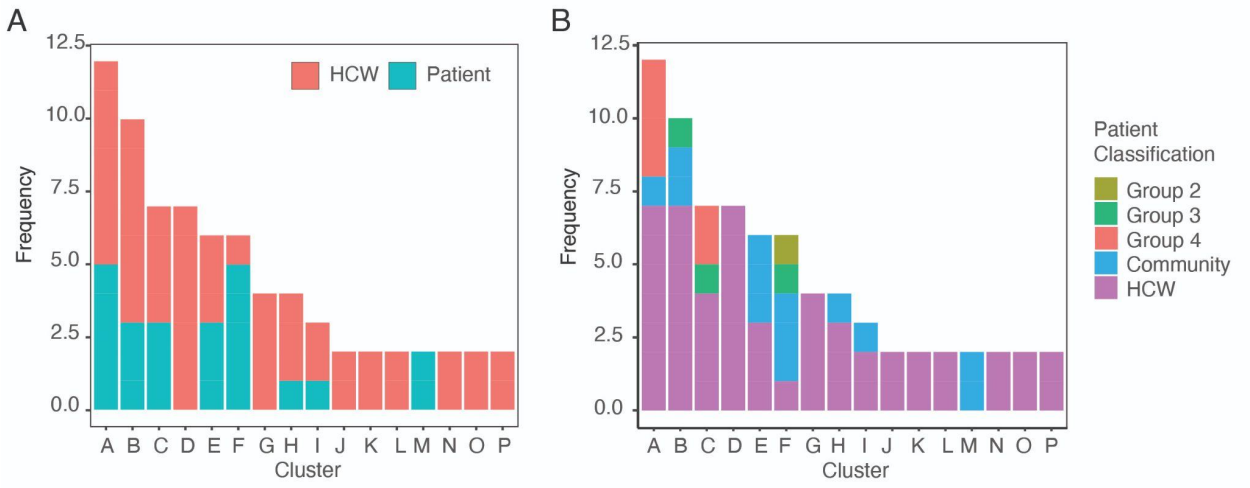


Fig S8. Characteristics of 16 potential hospital-associated HC transmission clusters. (A) Frequency of HCW and patients per cluster. (B) Patient classification according to hospitalization per cluster.

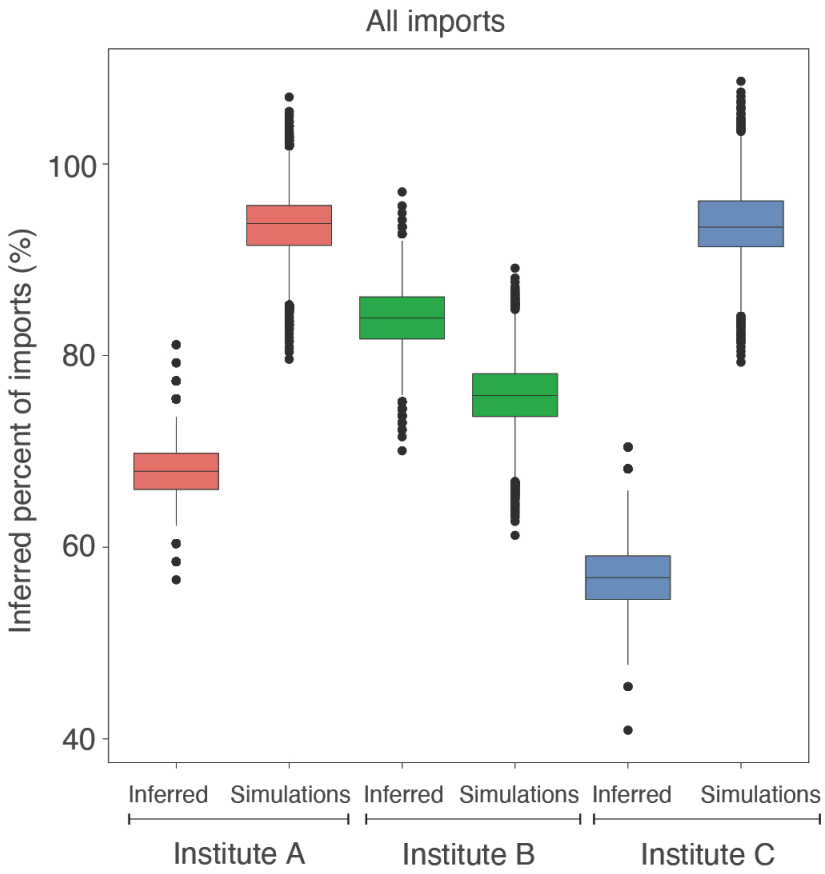


Fig S9. Inferred proportion of cases of patients and HCWs from HC institutes caused by imports from outside each institute. (A) Proportion (%) of inferred HCW imports to Institute A, Institute B, and Institute C. (B) Proportion (%) of inferred patient imports to Institute A, Institute B, and Institute C. Colours highlight boxplots from each institute Institute A (red), Institute B(green), Institute C(blue). Imports were inferred from a discrete trait analysis using Markov Jumps counts implemented on Beast v.1.10.4 (see methods for details).

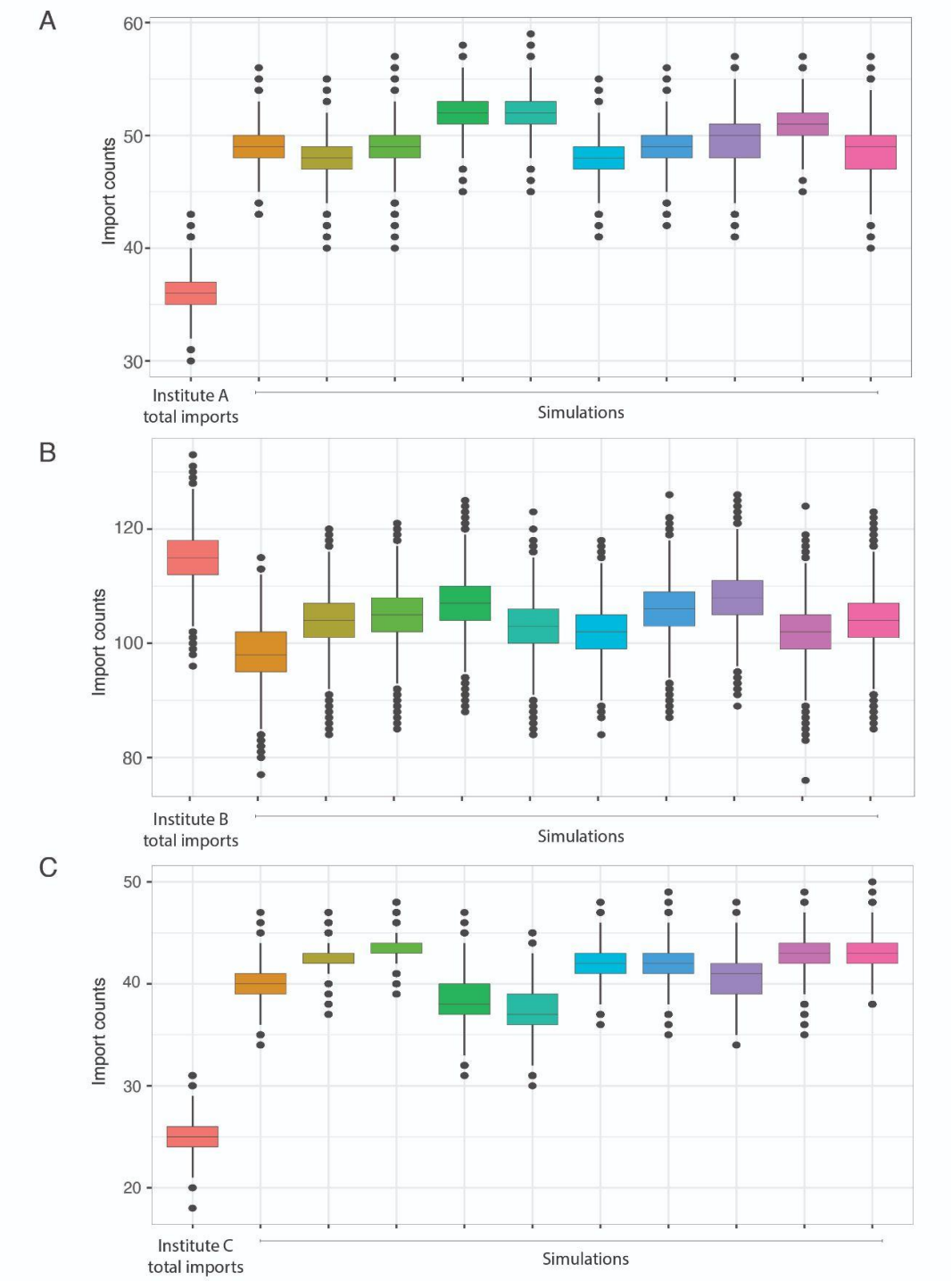


Fig S10. Distribution of total SARS-CoV-2 import cases per HC institute. Imports were inferred from dataset C (841 sequences) and using a Markov jumps approach implemented in Beast 1.10.4. An import is considered as any jumps coming from outside the institute. Simulations were performed by reshuffling the institute trait assigned to each HC sequence (see methods). (A) Total Institute A imports. (B) Total Institute Bimports. (C) Total Institute Cimports.

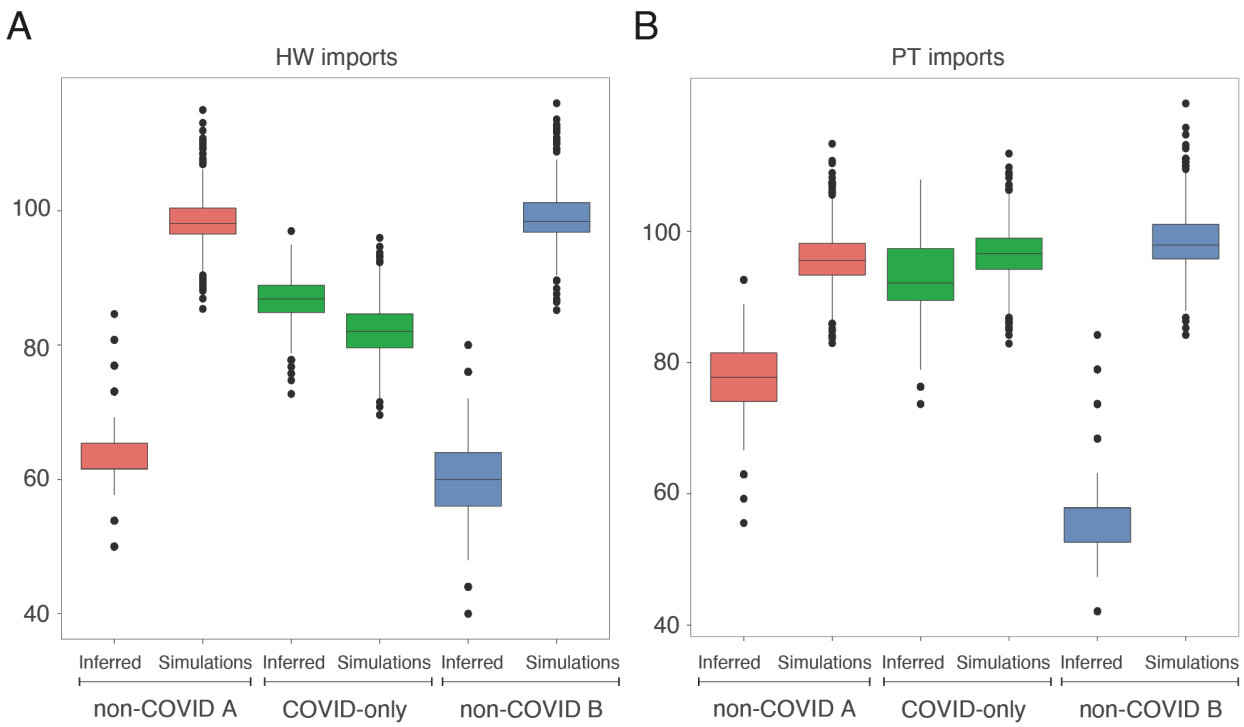


Fig S11. Inferred proportion of cases of patients and HCWs from HC institutes caused by imports from outside each institute. (A) Proportion (%) of inferred HCW imports to Institute A, Institute B, and Institute C. (B) Proportion (%) of inferred patient imports to Institute A, Institute B, and Institute C. Colours highlight boxplots from each institute Institute A (red), Institute B(green), Institute C(blue). Imports were inferred from a discrete trait analysis using Markov Jumps counts implemented on Beast v.1.10.4 (see methods for details).

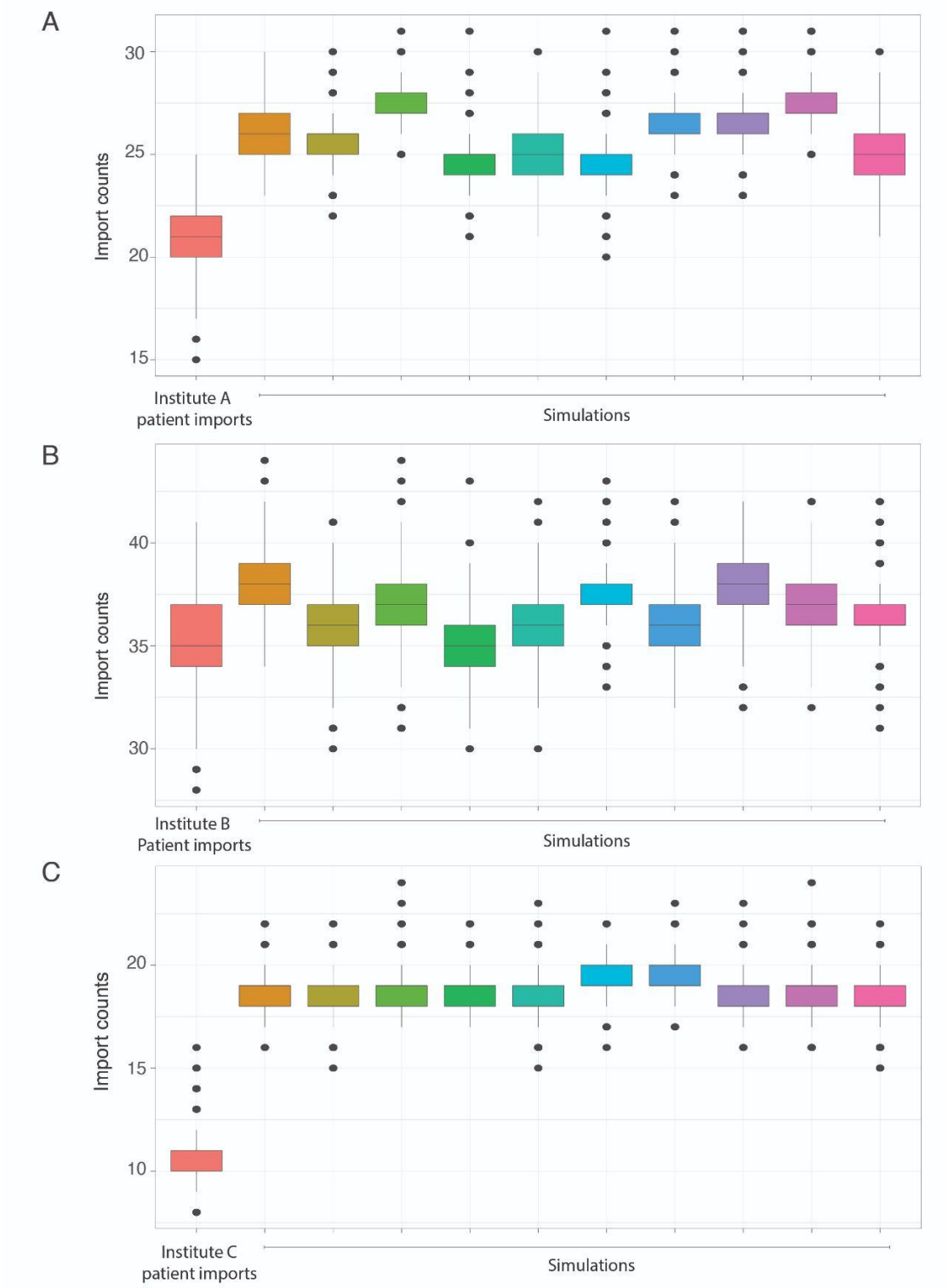


Fig S12. Distribution of patient SARS-CoV-2 import cases per HC institute. Imports were inferred from dataset C (841 sequences) and using a Markov jumps approach implemented in Beast 1.10.4. An import is considered as any jumps coming from outside the institute. Simulations were performed by reshuffling the institute trait assigned to each HC sequence (see methods). (A) Institute A patient imports. (B) Institute Bpatient imports. (C) Institute Cpatient imports.

Fig S13. Distribution of HCW SARS-CoV-2 import cases per HC institute. Imports were inferred from dataset C (841 sequences) and using a Markov jumps approach implemented in Beast 1.10.4. An import is considered as any jumps coming from outside the institute. Simulations were performed by reshuffling the institute trait assigned to each HC sequence (see methods). (A) Institute A HCW imports. (B) Institute BHCW imports. (C) Institute CHCW imports.

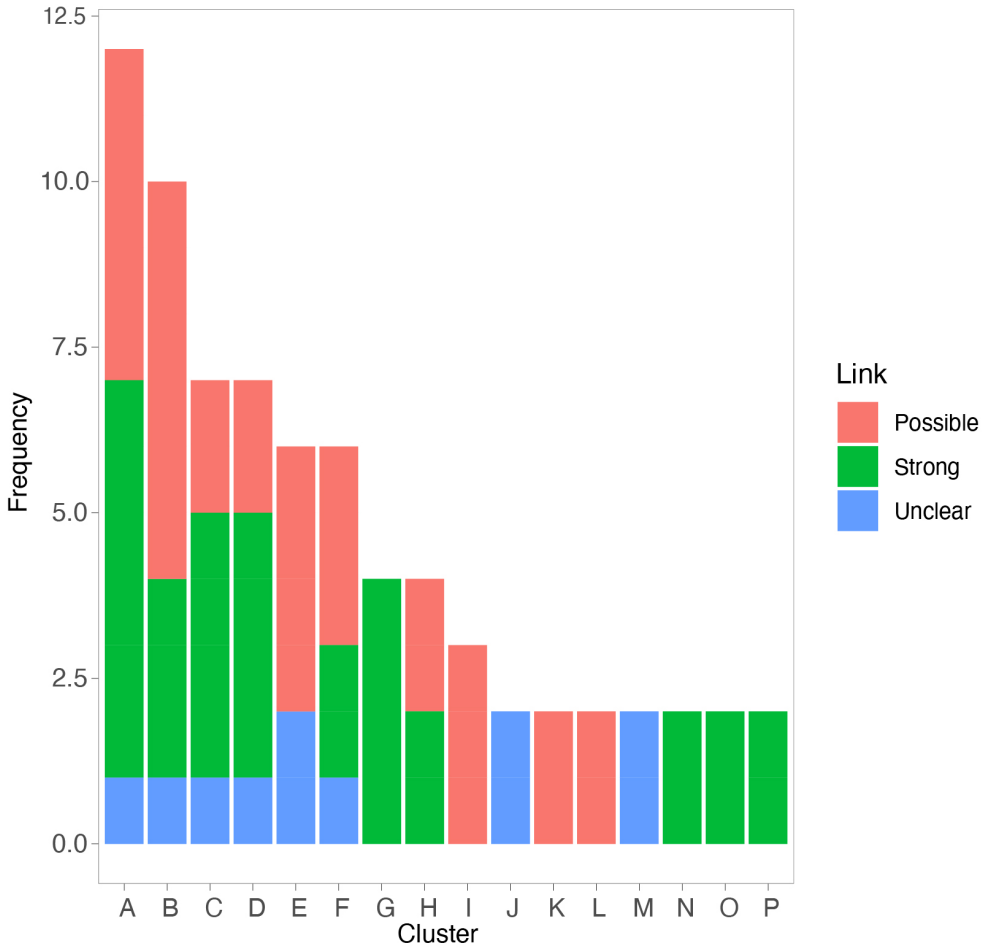
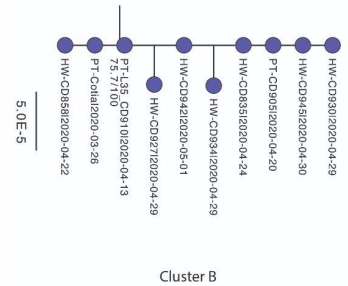


Fig S14. Frequency of epidemiological link strength by phylogenetic cluster.

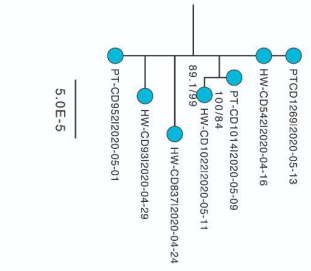
Cluster A

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD105	Inst C	Patient	Patient	Group5	18/03/2020	26/02/2020	Possible	Hospital acquired.
CD265	Inst C	Patient	Patient	Group5	28/03/2020	18/11/2019	Strong	Hospital acquired. CD98, 99, 88, 278, 267
CD278	Inst C	Patient	Patient	Group5	24/03/2020	20/02/2020	Strong	Hospital acquired. CD98, 99, 88, 267, 265
CD965	Inst C	Patient	Patient	Group5	30/04/2020	07/11/2019	Possible	Hospital acquired.
CD267	Inst C	Patient	Patient	Community	27/03/2020	30/03/2020	Possible	Onset before hospitalisation and hospitalisation before onset of other individuals from the same sector (CD98, 99, 88, 278, 265)
CD69	Inst C	HCW	Medical Resident	HCW	21/03/2020	NA	Possible	Same institute
CD98	Inst C	HCW	Nurse	HCW	21/03/2020	NA	Strong	Same unit as CD99, 88, 278, 267, 265
CD239	Inst C	HCW	Nurse Technician	HCW	22/03/2020	NA	Strong	Same unit as CD99, 88, 278, 267, 265
CD88	Inst C	HCW	Nursing Technician	HCW	22/03/2020	NA	Strong	Same unit as CD98, 99, 278, 267, 265
CD99	Inst C	HCW	Nursing Technician	HCW	21/03/2020	NA	Strong	Same unit as CD98, 88, 278, 267, 265
CD876	Inst B	HCW	Nursing Technician	HCW	18/04/2020	NA	Unclear	Different institute
CD891	Inst C	HCW	Kitchen Maid	HCW	17/04/2020	NA	Possible	Same institute



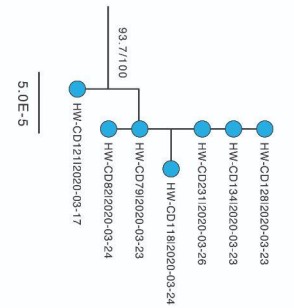
Cluster B

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD905	Inst A	Patient	Patient	Group4	20/04/2020	10/04/2020	Strong	Strong: 945 and 934; Group 4 patient
CD565	Inst C	Patient	Patient	Community	16/03/2020	28/03/2020	Unclear	Unclear: different institute; Transferred after symptom onset of HCW CD835 and CD858 from Inst B.
CD910	Inst A	Patient	Patient	Community	09/04/2020	12/04/2020	Possible	Possible link with Inst B HCW CD835 and CD858: Patient was seen at Inst B PSM on the 14/04, 4 and 5 days prior to their symptom onset.
CD858	Inst B	HCW	Medical Resident	HCW	18/04/2020	NA	Possible	Possible: same Institute as CD835, 1-day difference onset. Possible link with patient CD910 from Inst A.
CD835	Inst B	HCW	Nursing Technician	HCW	19/04/2020	NA	Possible	Possible: same institute as CD858, 1-day difference onset. Possible link with patient CD910 from Inst A.
CD927	Inst A	HCW	Nurse	HCW	27/04/2020	NA	Possible	Possible: same Institute
CD930	Inst A	HCW	Nursing Technician	HCW	25/04/2020	NA	Possible	Possible: same Institute
CD934	Inst A	HCW	Nursing Technician	HCW	26/04/2020	NA	Strong	Strong: same sector as CD945 and CD905.
CD942	Inst A	HCW	Nutrition Technician	HCW	25/04/2020	NA	Possible	Possible: same Institute
CD945	Inst A	HCW	Nursing Technician	HCW	26/04/2020	NA	Strong	Strong: 934 and 905.



Cluster C

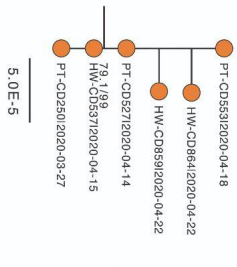
ID	Institute	HCW/patient	Occupation	Classificationn	Onset_date	Hospitalization date	Link	Epi link
CD1014	Inst C	Patient	Patient	Group5	09/05/2020	13/12/2019	Strong	Patient infected at the hospital. Same unit as CD952. Transferred to Inst B on 11/05/2020
CD1269	Inst C	Patient	Patient	Group5	13/05/2020	24/04/2020	Strong	Hospital acquired. Same unit as CD542. Transferred to Inst B 18/05/2020 U1CH
CD952	Inst C	Patient	Patient	Group4	26/04/2020	18/04/2020	Strong	Patient infected at the hospital. Same unit as CD1014 / Transferred to Inst B 04/05/2020 U1CH UTI - GASTRO 9° - 09AA05 / 11/05/2020 - E1CH - 9°CIRURGIA DO FIGADO - 9GN
CD542	Inst C	HCW	Nurse	HCW	13/04/2020	NA	Strong	Same unit as CD1269.
CD837	Inst B	HCW	Administration	HCW	19/04/2020	NA	Unclear	Different institute
CD932	Inst C	HCW	Medical Resident	HCW	26/04/2020	NA	Possible	Same institute
CD1022	Inst C	HCW	Administration	HCW	08/05/2020	NA	Possible	ICU doctor from different sector and floor



Cluster D

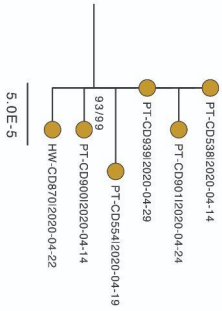
ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD121	Inst B	HCW	Medical Resident	HCW	12/03/2020	NA	Unclear	Different Institute
CD128	Inst A	HCW	Doctor	HCW	13/03/2020	NA	Strong	Same institute and occupation/ can be considered same sector
CD134	Inst A	HCW	Medical Resident	HCW	21/03/2020	NA	Strong	Same institute and medical specialty as CD118
CD79	Inst A	HCW	Doctor	HCW	22/03/2020	NA	Strong	Same institute and occupation/ can be considered same sector
CD82	Inst A	HCW	Doctor	HCW	22/03/2020	NA	Possible	Same institute and occupation
CD118	Inst A	HCW	Doctor	HCW	23/03/2020	NA	Strong	Same institute and medical specialty as CD134
CD231	Inst A	HCW	Doctor	HCW	23/03/2020	NA	Possible	Same institute and occupation

Fig S15. Maximum likelihood phylogenetic subtrees and epidemiological characteristics of samples in transmission clusters A, B, C, and D. Cluster subtrees were extracted from an ML tree inferred from Dataset C (see methods).



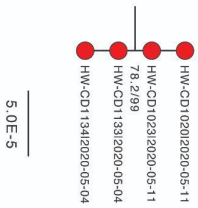
Cluster E

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD250	Inst C	Patient	Patient	Community	18/03/2020	27/03/2020	Possible	Symptom onset before hospitalisation. Community infection. Transferred to Inst B on the 26/03, PS and UMIN. Same institute as CD859 and CD864.
CD527	Inst A	Patient	Patient	Community	09/04/2020	13/04/2020	Unclear	Symptom onset before hospitalisation. Community infection.
CD553	Inst B	Patient	Patient	Community	05/04/2020	18/04/2020	Possible	Symptom onset before hospitalisation. Community infection. Same institute as CD859 and CD864.
CD537	Inst A	HCW	Nursing Technician	HCW	11/04/2020	NA	Unclear	Symptom onset before CD527 was admitted
CD859	Inst B	HCW	Pharmacy Technician	HCW	14/04/2020	NA	Possible	Same institute as CD864, CD553, CD250
CD864	Inst B	HCW	Medical Resident	HCW	18/04/2020	NA	Possible	Same institute as CD859, CD553, CD250



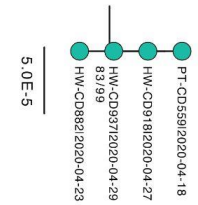
Cluster F

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD900	Inst A	Patient	Patient	Group3	21/04/2020	14/04/2020	Strong	Same unit to CD939. Onset after hospitalisation. Transferred to Inst B on the 26/04 EPS/4
CD939	Inst A	Patient	Patient	Group4	28/04/2020	18/04/2020	Strong	Same unit to CD900. Onset after hospitalisation.Transferred to Inst B on the 02/05 ECME/5
CD538	Inst A	Patient	Patient	Community	14/04/2020	14/04/2020	Possible	Community acquired. Transferred to Inst B on the 17/04 EPS/4, 4 days before symptom onset of CD870
CD554	Inst A	Patient	Patient	Community	03/04/2020	16/04/2020	Possible	Community acquired. Transferred to Inst B on the 17/04 EPS/4, 4 days before symptom onset of CD870
CD901	Inst A	Patient	Patient	Community	15/04/2020	22/04/2020	Unclear	Community acquired. Transferred to Inst B on the 25/04 EPS/4
CD870	Inst B	HCW	MedicalResident	HCW	18/04/2020	NA	Possible	Symptom onset 4 days after CD554 and CD538 were transferred to Inst B.



Cluster G

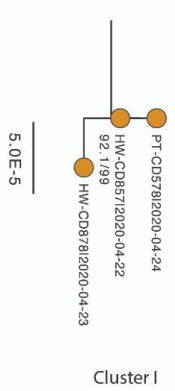
ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD1134	Inst B	HCW	Administration	HCW	30/04/2020	NA	Strong	Strong. Same sector as other individuals in the cluster
CD1133	Inst B	HCW	Administration	HCW	01/05/2020	NA	Strong	Strong. Same sector as other individuals in the cluster
CD1020	Inst B	HCW	Administration	HCW	04/05/2020	NA	Strong	Strong. Same sector as other individuals in the cluster
CD1023	Inst B	HCW	Administration	HCW	08/05/2020	NA	Strong	Strong. Same sector as other individuals in the cluster



Cluster H

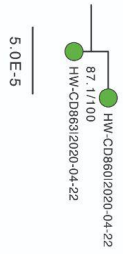
ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD559	Inst B	Patient	Patient	Community	12/04/2020	18/04/2020	Possible	Same institute as CD882
CD882	Inst B	HCW	Nursing Technician	HCW	15/04/2020	NA	Possible	Same institute as CD559
CD918	Inst C	HCW	Nurse	HCW	24/04/2020	NA	Strong	Same institute, similar occupation and same specialty to CD937/ same service
CD937	Inst C	HCW	Nursing Technician	HCW	23/04/2020	NA	Strong	Same institute, similar occupation and same specialty to CD918. Same occupation to CD882, possible link between institutes/ same service

Fig S16. Maximum likelihood phylogenetic subtrees and epidemiological characteristics of samples in transmission clusters E, F, G, and H. Cluster subtrees were extracted from an ML tree inferred from Dataset C (see methods).



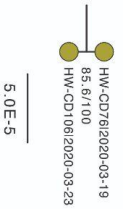
Cluster I

ID	Institute	HCW/patient	Occupation	Classification	Onset_date	Hospitalization _date	Link	Epi link
CD857	Inst B	HCW	Nursing Technician	HCW	15/04/2020	NA	Possible	Same Institute as CD878
CD878	Inst B	HCW	Nurse	HCW	19/04/2020	NA	Possible	Same Institute as CD857



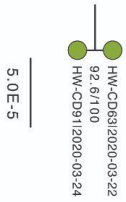
Cluster J

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD860	Inst B	HCW	Administration	HCW	16/04/2020	NA	Unclear	Same institute, different buildings
CD863	Inst B	HCW	Nursing Technician	HCW	16/04/2020	NA	Unclear	Same institute, different buildings



Cluster K

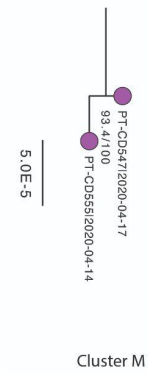
ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD76	Inst C	HCW	Doctor	HCW	19/03/2020	NA	Possible	Same institute and similar areas
CD106	Inst C	HCW	Doctor	HCW	21/03/2020	NA	Possible	Same institute and similar areas



Cluster L

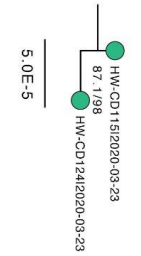
ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD63	Inst A	HCW	Doctor	HCW	22/03/2020	NA	Possible	Same institute and CD91 works on various floors
CD91	Inst A	HCW	Physiotherapist	HCW	21/03/2020	NA	Possible	Same institute and CD91 works on various floors

Fig S17. Maximum likelihood phylogenetic subtrees and epidemiological characteristics of samples in transmission clusters I, J, K, and L. Cluster subtrees were extracted from an ML tree inferred from Dataset C (see methods).



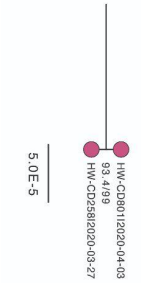
Cluster M

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD555	Inst A	Patient	Patient	Community	07/04/2020	14/04/2020	Unclear	No clear epi link: Transferred to Inst B on the 25/04/2020 8DN06; 29/04/2020 U4DN- UTI 4 DN 04DN11. Community acquired.
CD547	Inst B	Patient	Patient	Community	14/04/2020	17/04/2020	Unclear	No clear epi link; Patients with symptom onset prior to hospitalization. Hospitalization happens in different institutes. Community acquired.



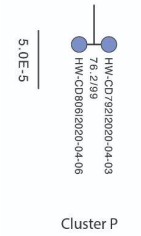
Cluster N

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD124	Inst B	HCW	Medical Resident	HCW	19/03/2020	NA	Strong	Same institute, same medical specialty and floor
CD115	Inst B	HCW	Medical Resident	HCW	21/03/2020	NA	Strong	Same institute, same medical specialty and floor



Cluster O

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD258	Inst B	HCW	Medical Resident	HCW	24/03/2020	NA	Strong	Same Institute and same medical specialty
CD801	Inst B	HCW	Medical Resident	HCW	31/03/2020	NA	Strong	Same Institute and same medical specialty



Cluster P

ID	Institute	HCW/patient	Occupation	Classification	Onset date	Hospitalization date	Link	Epi link
CD792	Inst B	HCW	Administration	HCW	30/03/2020	NA	Strong	Same institute and specialty/ part of the same service
CD806	Inst B	HCW	Nursing Technician	HCW	31/03/2020	NA	Strong	Same institute and specialty/ part of the same service

Fig S18. Maximum likelihood phylogenetic subtrees and epidemiological characteristics of samples in transmission clusters M, N, O, and P. Cluster subtrees were extracted from an ML tree inferred from Dataset C (see methods).

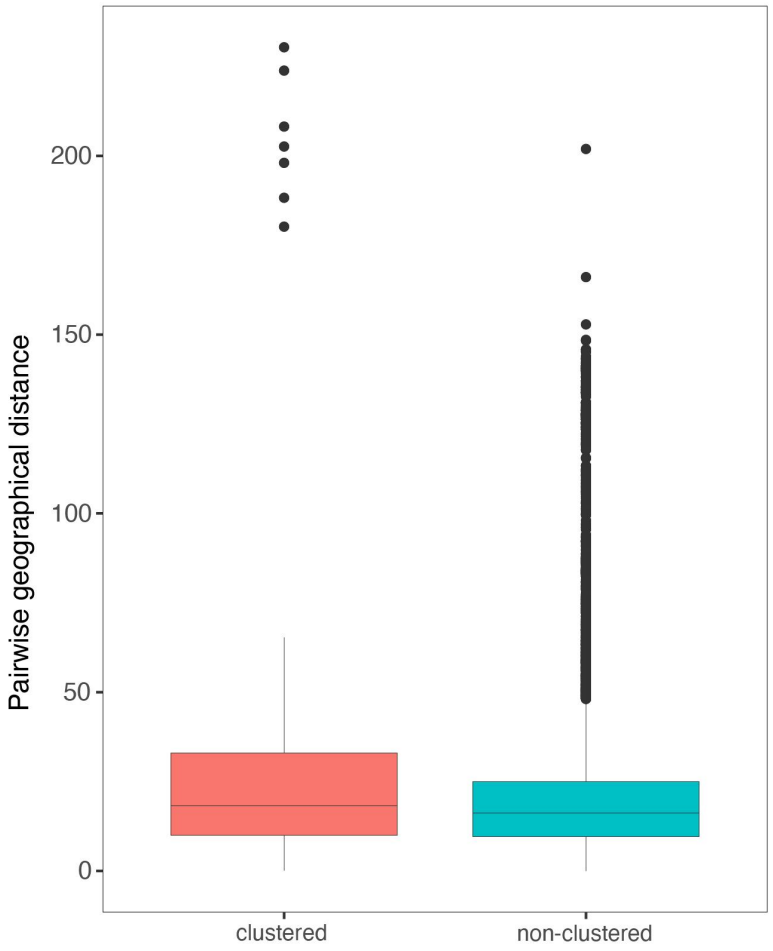


Fig S19. Comparison between pairwise geographical distances of households of clustered and non-clustered sequences. For clustered sequences, pairwise household geographical distances were estimated between sequences of the same cluster. For non-clustered sequences, pairwise household geographical distances were estimated between all sequences.

Table S1. Epidemiological and Demographic characteristics of all Hospital das Clínicas Complex (HC) and of Institute B, Institute C, Institute A, and other institutes.

	HC (n = 3898)				**Inst B (n = 2159, 55.4%)**				**Inst C (n = 716, 18.4%)**				**Inst A (n = 703, 18%)**			
	All	Patients	HW	p-value	All	Patients	HCW		All	Patients	HCW		All	Patients	HCW	
		(n = 2008)	(n = 1890)			(n = 1468)	(n = 691)			(n = 215)	(n = 501)			(n = 269)	(n = 434)	
Age	46 (0-101)	60 (0 - 101)	37 (17-84)	<0.0001	52 (0-97)	60 (0-101)	38 (17-71)	<0.0001	39 (0-93)	61 (0-93)	35 (20-67)	<0.0001	42 (15-92)	64 (15-92)	37 (19-66)	<0.0001
Sex																
Female	2251 (57.75%)	921 (45.9%)	1330 (70.4%)	<0.0001	1112 (51.5%)	665 (45.3%)	447 (64.7%)		456 (63.7%)	92 (42.7%)	364 (72.65%)	<0.0001	474 (67.4%)	140 (52.0%)	334 (76.9%)	<0.0001
Male	1647 (42.25%)	1087 (54.1%)	560 (29.6%)		1047 (48.5%)	803 (54.7%)	244 (35.3%)		258 (36.3%)	123 (57.2%)	137 (27.34%)		229 (32.6%)	129 (48.0%)	100 (23.1%)	
Occupation																
Nursing Technician	557 (14.3%)	-	557 (29.5%)	-	172 (8.0%)	-	172 (24.9%)		156 (21.8%)	-	156 (31.1%)	-	171 (24.3%)	-	171 (39.4%)	-
Doctor	421 (10.8%)	-	421 (22.3%)		188 (8.7%)	-	188 (27.2%)		96 (13.4%)	-	96 (19.2%)		65 (9.2%)	-	65 (15.0%)	
Administrative	295 (7.6%)	-	295 (15.6%)		130 (6.0%)	-	130 (18.8%)		64 (8.9%)	-	64 (12.8%)		56 (8.0%)	-	56 (8.0%)	
Nurse	282 (7.2%)	-	282 (14.9%)		70 (3.2%)	-	70 (10.1%)		95 (13.3%)	-	95 (19.0%)		81 (11.5%)	-	81 (18.7%)	
Physiotherapist	44 (1.3%)	-	44 (2.3%)		16 (0.7%)	-	16 (2.3%)		14 (1.9%)	-	14 (2.8%)		10 (1.4%)	-	10 (2.3%)	
KitchenMaid	33 (0.85%)	-	33 (1.7%)		13 (0.6%)	-	13 (1.8%)		16 (2.2%)	-	16 (2.8%)		4 (0.6%)	-	4 (0.9%)	

Cleaning	22 (0.6%)	-	22 (1.2%)		12 (0.55%)	-	12 (1.7%)		2 (0.3%)	-	2 (0.4%)		0 (0.0%)	-	0 (0.0%)	
RadiologyTechnician	22 (0.6%)	-	22 (1.2%)		5 (0.2%)	-	5 (0.7%)		5 (0.7%)	-	5 (1.0%)		4 (0.6%)	-	4 (0.9%)	
Others	219 (5.6%)	-	219 (10.4%)		76 (3.4%)	-	76 (10.6%)		51 (7.1%)	-	51 (10.2%)		42 (6.0%)	-	42 (9.7%)	
Unknown	15 (0.4%)	-	15 (0.9%)		12 (0.55%)	-	12 (1.7%)		2 (0.3%)	-	2 (0.4%)		1 (0.1%)	-	1 (0.2%)	
Sector																
Inpatient	1099 (28.2%)	933 (46.5%)	166 (8.7%)	<0.0001	815 (37.7%)	759 (51.7%)	56 (8.1%)		144 (20.1%)	90 (41.9%)	54 (10.8%)	<0.0001	103 (14.6%)	62 (23.0%)	41 (9.4%)	<0.0001
ICU	713 (18.3%)	422 (21.0%)	291 (15.4%)		406 (18.8%)	298 (20.3%)	165 (15.6%)		197 (27.5%)	87 (40.5%)	110 (21.9%)		90 (12.8%)	37 (13.7%)	53 (12.2%)	
Emergency	471 (12.1%)	357 (17.8%)	114 (6.0%)		230 (10.65%)	174 (11.8%)	56 (8.1%)		40 (5.6%)	14 (6.5%)	26 (5.2%)		185 (26.3%)	165 (61.3%)	20 (4.6%)	
Administration	86 (2.2%)	0 (0.0%)	86 (4.5%)		54 (2.5%)	0 (0.0%)	54 (7.8%)		11 (1.5%)	0 (0.0%)	11 (2.2%)		12 (1.7%)	0 (0.0%)	12 (2.8%)	
Outpatient	100 (2.6%)	10 (0.5%)	90 (4.8%)		60 (2.8%)	8 (0.5%)	39 (5.7%)		15 (2.1%)	2 (0.9%)	13 (2.6%)		5 (0.7%)	0 (0.0%)	11 (2.5%)	
Others	589 (15.1%)	59 (2.9%)	530 (28.0%)		180 (8.4%)	34 (2.6%)	165 (23.9%)		156 (21.8%)	18 (8.4%)	138 (27.5%)		153 (21.8%)	5 (1.8%)	148 (34.1%)	
Unknown	840 (21.5%)	227 (11.3%)	613 (32.4%)		414 (19.2%)	214 (14.6%)	200 (28.9%)		153 (21.4%)	4 (1.9%)	149 (29.7%)		149 (21.2%)	0 (0.0%)	149 (34.3%)	
Outcome																
Release	2946 (75.6%)	1061 (52.8%)	1885 (99.7%)	<0.0001	1504 (69.6%)	814 (55.4%)	690 (99.85%)		596 (83.2%)	98 (45.6%)	498 (99.4%)	<0.0001	543 (77.24%)	109 (40.5%)	434 (100.0%)	<0.0001
Death	577 (14.8%)	572 (28.5%)	5 (0.3%)		356 (16.5%)	355 (24.2%)	1 (0.15%)		78 (10.9%)	75 (34.9%)	3 (0.6%)		131 (18.6%)	131 (48.7%)	0 (0.0%)	

Transfer	36 (0.9%)	36 (2.0%)	0 (0.0%)		27 (1.2%)	27 (1.7%)	0 (0.0%)		5 (0.2%)	5 (2.2%)	0 (0.0%)		3 (0.4%)	3 (1.1%)	0 (0.0%)	
Unknown	339 (8.7%)	339 (16.9%)	0 (0.0%)		272 (12.6%)	272 (18.5%)	0 (0.0%)		37 (5.2%)	37 (17.2%)	0 (0.0%)		26 (3.7%)	26 (9.6%)	0 (0.0%)	
Municipality																
São Paulo City	2568 (65.9%)	1352 (67.3%)	1216 (64.3%)	<0.001	1534 (71.0%)	1017 (69.3%)	517 (74.8%)		431 (60.2%)	135 (62.8%)	296 (59.1%)	0.001	404 (57.5%)	157 (58.4%)	247 (56.9%)	0.0008
Others	1055 (27.1%)	656 (32.7%)	399 (21.1%)		605 (28.0%)	451 (30.7%)	154 (22.3%)		175 (24.4%)	80 (37.2%)	95 (19.0%)		211 (30.0%)	112 (41.6%)	99 (22.8%)	
Unknown	275 (7.0%)	0 (0.0%)	275 (14.6%)		20 (0.9%)	0 (0.0%)	20 (2.9%)		110 (15.4%)	0 (0.0%)	110 (21.9%)		88 (12.5%)	0 (0.0%)	88 (20.3%)	

Table S2. Epidemiological and demographic characteristics of groups 2, 3, and 4 patients from three HC complex institutes.

	Patients (n = 167)			p-value
	3 - 7 days (n=27)	8 - 14 days (n=54)	>14 days (n=86)	
	Group 2	Group 3	Group 4	
Age	64 (0 - 92)	63 (25 - 81)	60 (14 - 92)	0.16
Sex				
Female	10 (37.0%)	22 (40.7%)	42 (48.8%)	0.45
Male	17 (63%)	32 (59.3%)	44 (51.2%)	
Institute				
Inst B	0 (0.0%)	3 (5.5%)	1 (1.2%)	-
Inst C	18 (66.7%)	30 (55.5%)	70 (81.4%)	
Inst A	9 (33.3%)	21 (38.9%)	15 (17.4%)	
Sector				
Inpatient unit	13 (48.1%)	29 (53.7%)	41 (47.7%)	-
ICU	11 (40.7%)	18 (33.3%)	38 (44.2%)	
Emergency Room	1 (4.0%)	4 (7.4%)	5 (5.8%)	
Outpatients unit	0 (0.0%)	0 (0.0%)	2 (2.3%)	
Others	2 (7.4%)	1 (1.8%)	0 (0.0%)	
Unknown	0 (0.0%)	2 (3.7%)	0 (0.0%)	
Outcome				
Death	12 (44.5%)	25 (46.3%)	35 (40.7%)	0.93
Discharged	11 (40.7%)	19 (35.2%)	29 (33.7%)	
Transfer	0 (0.0%)	2 (3.7%)	4 (4.6%)	

Unknown	4 (14.8%)	8 (14.8%)	18 (20.9%)	

Table S3. Sequencing statistics for the Brazilian SARS-COV-2 genomes from this study (n=340).

Isolate	GISAID ID	New GISAID submission	Municipality	State	Collection Date	Mapped Reads	Average depth coverage	Bases covered >10x	Bases covered >25x	Reference covered (%)
CD100	EPI_ISL_476449	No	Sao Paulo	SP	2020-03-24	57472	717.771	27974	27153	89.9
CD1002	EPI_ISL_1172015	No	Sao Paulo	SP	2020-05-06	51838	645.976	26774	25293	81.8
CD1003	EPI_ISL_4463722	Yes	Sao Paulo	SP	2020-05-06	59284	746.913	26275	24424	79.6
CD1008	EPI_ISL_1172016	No	Sao Paulo	SP	2020-05-08	28954	361.532	26156	23357	79.2
CD101	EPI_ISL_476450	No	Sao Paulo	SP	2020-03-23	71837	888.494	29758	28960	95.2
CD1011	EPI_ISL_722006	No	Sao Paulo	SP	2020-05-09	63171	783.391	29222	29203	96.9
CD1014	EPI_ISL_721995	No	Sao Paulo	SP	2020-05-09	52807	657.624	29196	29016	96.2
CD1016	EPI_ISL_4463723	Yes	Sao Paulo	SP	2020-04-23	75669	941.732	29032	29019	96.1
CD1020	EPI_ISL_4463724	Yes	Sao Paulo	SP	2020-05-11	57547	719.699	28007	27286	89.6
CD1022	EPI_ISL_4463725	Yes	Cotia	SP	2020-05-11	131119	1622.29	29449	29209	97.0
CD1023	EPI_ISL_4463726	Yes	Sao Paulo	SP	2020-05-11	103425	1288.57	29268	28537	94.4
CD1025	EPI_ISL_4463727	Yes	Sao Paulo	SP	2020-04-27	86029	1068	26783	24805	85.4
CD105	EPI_ISL_476452	No	Sao Paulo	SP	2020-03-21	20454	252.938	29431	28406	95.9
CD106	EPI_ISL_476453	No	Sao Paulo	SP	2020-03-23	77578	959.98	29158	28215	93.5
CD107	EPI_ISL_476454	No	Sao Paulo	SP	2020-03-23	65714	830.558	26697	25044	83.4
CD109	EPI_ISL_476455	No	Sao Paulo	SP	2020-03-24	41571	523.947	26899	24987	81.8
CD110	EPI_ISL_476456	No	Sao Paulo	SP	2020-03-23	54967	679.223	29177	28384	94.3
CD111	EPI_ISL_476457	No	Sao Paulo	SP	2020-03-21	32555	406.751	27655	25681	87.3
CD1129	EPI_ISL_4463728	Yes	Poa	SP	2020-05-04	212992	2381.59	29232	29028	96.3
CD113	EPI_ISL_476459	No	Sao Paulo	SP	2020-03-20	43796	550.345	27191	25287	83.2
CD1130	EPI_ISL_4463729	Yes	Sao Paulo	SP	2020-04-27	222158	1752.33	23615	22723	75.0
CD1131	EPI_ISL_4297851	Yes	Sao Paulo	SP	2020-04-30	215848	1559.04	27664	25739	87.3
CD1132	EPI_ISL_1171871	No	Sao Paulo	SP	2020-05-04	273611	2804.66	29276	29072	96.3
CD1133	EPI_ISL_4297852	Yes	Sao Paulo	SP	2020-05-04	186352	2297.73	29286	29208	97.2
CD1134	EPI_ISL_4297853	Yes	Embu das Artes	SP	2020-05-04	202136	2529.59	29466	29278	97.2
CD1135	EPI_ISL_4297854	Yes	Sao Paulo	SP	2020-05-05	157917	1920.49	29238	29071	96.3
CD114	EPI_ISL_476460	No	Guarulhos	SP	2020-03-20	41489	521.837	26433	24987	80.7

CD115	EPI_ISL_476461	No	Sao Paulo	SP	2020-03-23	67205	830.874	29834	29184	96074
CD116	EPI_ISL_476462	No	Sao Paulo	SP	2020-03-23	65774	815.213	29615	28988	96.8
CD118	EPI_ISL_476469	No	Sao Paulo	SP	2020-03-24	128482	1626.99	29836	29836	98.6
CD121	EPI_ISL_476471	No	Jandira	SP	2020-03-17	104908	1326.95	29588	29384	96.9
CD122	EPI_ISL_476472	No	Sao Paulo	SP	2020-03-19	178589	2139.45	29147	28704	95.3
CD124	EPI_ISL_476473	No	Sao Paulo	SP	2020-03-23	181356	1898.82	27552	26950	89.1
CD126	EPI_ISL_476475	No	Sao Paulo	SP	2020-03-24	177196	2017.32	28628	28352	92.7
CD1261	EPI_ISL_4463730	Yes	Guarulhos	SP	2020-05-14	18289	226.612	29213	29206	96.9
CD1262	EPI_ISL_4297855	Yes	Sao Paulo	SP	2020-05-14	22872	284.946	26814	25000	83.2
CD1264	EPI_ISL_4297856	Yes	Sao Paulo	SP	2020-05-14	19422	241.971	25801	22681	75.7
CD1265	EPI_ISL_4297857	Yes	Sao Paulo	SP	2020-05-14	12256	152.797	27388	24988	86.2
CD1268	EPI_ISL_4297858	Yes	Sao Paulo	SP	2020-05-14	24900	311.296	27942	26239	87.0
CD1269	EPI_ISL_4297859	Yes	Santo Andre	SP	2020-05-13	19845	244.974	29167	29094	96.4
CD1276	EPI_ISL_4297860	Yes	Sao Paulo	SP	2020-05-12	9924	122.826	28702	27446	92.2
CD1278	EPI_ISL_4297861	Yes	Sao Paulo	SP	2020-05-14	24601	303.813	29458	29457	98.0
CD1279	EPI_ISL_4297862	Yes	Franco da Rocha	SP	2020-05-14	19482	246.765	25865	22679	77.4
CD128	EPI_ISL_476477	No	Sao Paulo	SP	2020-03-23	116345	1473.48	28847	28564	93.5
CD1283	EPI_ISL_4297863	Yes	Sao Paulo	SP	2020-05-15	34797	450.457	25431	22832	76.0
CD1284	EPI_ISL_4297864	Yes	Sao Paulo	SP	2020-05-15	85433	1077.47	28014	26851	88.1
CD1285	EPI_ISL_4297865	Yes	Embu das Artes	SP	2020-05-15	16228	202.795	29270	29003	97.1
CD1287	EPI_ISL_4297866	Yes	Sao Paulo	SP	2020-05-15	110708	1384.22	28880	27227	89.6
CD1288	EPI_ISL_4297867	Yes	Carapicuiba	SP	2020-05-15	70865	886.936	29039	29025	96.2
CD1289	EPI_ISL_4297868	Yes	Sao Paulo	SP	2020-05-15	69833	869.803	26346	23972	80.3
CD1292	EPI_ISL_4463731	Yes	Sao Paulo	SP	2020-05-14	32859	417.543	25827	23644	77.9
CD1293	EPI_ISL_4297869	Yes	Sao Paulo	SP	2020-05-15	104135	1298.02	29458	29269	97.1
CD1297	EPI_ISL_4297870	Yes	Sao Paulo	SP	2020-05-14	101437	1269.58	29274	28864	96.3
CD1298	EPI_ISL_4297871	Yes	Guarulhos	SP	2020-05-15	98654	1234.54	29273	29254	97.1
CD1300	EPI_ISL_4297872	Yes	Diadema	SP	2020-05-15	57045	705.346	29002	28100	94.0
CD1303	EPI_ISL_4297873	Yes	Sao Paulo	SP	2020-05-14	73375	921.437	28821	27582	92.4
CD1305	EPI_ISL_4297874	Yes	Osasco	SP	2020-05-17	76351	952.286	26871	25362	85.4
CD1306	EPI_ISL_4297875	Yes	Sao Paulo	SP	2020-05-20	81809	1004.05	27076	25471	85.4

CD131	EPI_ISL_476479	No	Sao Paulo	SP	2020-03-19	59589	788.547	26108	24493	80.6
CD1316	EPI_ISL_4297876	Yes	Sao Paulo	SP	2020-05-17	47548	595.197	27875	27049	89.0
CD133	EPI_ISL_476480	No	Taboao da Serra	SP	2020-03-23	112350	1434.81	28401	26871	87.1
CD134	EPI_ISL_476481	No	Sao Paulo	SP	2020-03-23	60637	766.611	28958	28353	92.7
CD135	EPI_ISL_476482	No	Sao Paulo	SP	2020-03-24	70126	895.417	28290	27135	90.7
CD139	EPI_ISL_476486	No	Sao Paulo	SP	2020-03-24	114744	1452.87	29408	28978	95.3
CD231	EPI_ISL_476237	No	Francisco Morato	SP	2020-03-26	101131	1272.4	29053	28318	93.4
CD232	EPI_ISL_476238	No	Sao Paulo	SP	2020-03-26	116552	1282.01	28367	26189	87.4
CD234	EPI_ISL_476239	No	Guarulhos	SP	2020-03-26	83874	1050.7	28362	26806	89.8
CD235	EPI_ISL_476240	No	Santo Andre	SP	2020-03-26	103474	1234.57	28585	26618	89.6
CD236	EPI_ISL_476241	No	Carapicuiba	SP	2020-03-24	63095	679.685	26516	23297	78.8
CD237	EPI_ISL_476242	No	Osasco	SP	2020-03-26	102358	1103.6	27206	25196	83.8
CD239	EPI_ISL_476243	No	Mogi das Cruzes	SP	2020-03-23	91282	1159.16	29300	28991	95.3
CD240	EPI_ISL_476244	No	Sao Paulo	SP	2020-03-26	66959	848.48	29675	29675	97.9
CD241	EPI_ISL_476245	No	Embu das Artes	SP	2020-03-26	82197	1044.52	29371	28662	95.7
CD242	EPI_ISL_476246	No	Sao Paulo	SP	2020-03-27	126660	1601.78	29675	29640	97.9
CD243	EPI_ISL_476247	No	Sao Paulo	SP	2020-03-27	95802	1201.52	29046	28048	92.6
CD246	EPI_ISL_476249	No	Sao Paulo	SP	2020-03-27	90130	1049.37	27902	26164	88.1
CD250	EPI_ISL_476250	No	Taboao da Serra	SP	2020-03-25	52724	667.049	29486	29038	96.3
CD252	EPI_ISL_476251	No	Ferraz de Vasconcelos	SP	2020-03-21	72855	784.277	25639	22858	75.8
CD255	EPI_ISL_476252	No	Sao Paulo	SP	2020-03-27	69676	868.835	27424	25565	86.5
CD257	EPI_ISL_476253	No	Sao Paulo	SP	2020-03-27	75996	976.113	29039	27664	93.6
CD258	EPI_ISL_476254	No	Sao Paulo	SP	2020-03-27	71401	904.888	29490	29478	97.1
CD260	EPI_ISL_476256	No	Osasco	SP	2020-03-27	58654	742.317	29674	29670	97.9
CD262	EPI_ISL_476257	No	Sao Paulo	SP	2020-03-24	64287	807.039	27683	25098	85.3
CD263	EPI_ISL_476258	No	Sao Paulo	SP	2020-03-27	54650	662.635	28553	27830	92.5
CD265	EPI_ISL_476259	No	Sao Paulo	SP	2020-03-29	14932	178.592	29667	29363	97.9
CD266	EPI_ISL_476260	No	Sao Paulo	SP	2020-03-27	41575	497.046	29367	28972	97.1
CD267	EPI_ISL_476261	No	Sao Paulo	SP	2020-03-29	36392	434.083	29399	28329	95.3
CD268	EPI_ISL_476262	No	Sao Paulo	SP	2020-03-29	38441	460.513	29668	29201	97.1

CD270	EPI_ISL_476263	No	Sao Paulo	SP	2020-03-30	41739	500.181	29379	28405	94.4
CD271	EPI_ISL_476264	No	Sao Paulo	SP	2020-03-30	32654	400.351	26509	24127	81.2
CD272	EPI_ISL_476265	No	Sao Paulo	SP	2020-03-30	38024	454.495	29673	29659	97.9
CD273	EPI_ISL_476266	No	Sao Paulo	SP	2020-03-30	46695	559.593	29164	28698	96.0
CD274	EPI_ISL_476267	No	Franco da Rocha	SP	2020-03-30	11820	141.811	29178	28727	96.9
CD277	EPI_ISL_476269	No	Sao Paulo	SP	2020-03-31	7798	941.438	28663	26476	88.8
CD278	EPI_ISL_476270	No	Franco da Rocha	SP	2020-03-30	33977	408.999	29118	28443	94.2
CD280	EPI_ISL_476271	No	Guaruja	SP	2020-03-30	7473	903.933	28103	25610	88.1
CD281	EPI_ISL_476272	No	Sao Paulo	SP	2020-03-30	38254	462.456	29001	28463	94.3
CD282	EPI_ISL_476273	No	Sao Paulo	SP	2020-03-30	26792	321.717	28721	27563	93.4
CD283	EPI_ISL_476274	No	Embu das Artes	SP	2020-03-30	48921	586.647	29434	28749	96.1
CD284	EPI_ISL_476275	No	Sao Paulo	SP	2020-03-30	20633	250.424	27038	24094	81.1
CD285	EPI_ISL_476276	No	Embu das Artes	SP	2020-03-30	19201	233.625	25884	22849	78.6
CD286	EPI_ISL_476277	No	Sao Paulo	SP	2020-03-30	19110	228.81	29346	27597	92.9
CD292	EPI_ISL_4297877	Yes	Sao Paulo	SP	2020-03-30	597600	3175.86	29446	29134	96.1
CD293	EPI_ISL_4297878	Yes	Sao Paulo	SP	2020-03-26	519575	4319.42	29838	29689	97.9
CD37	EPI_ISL_4297879	Yes	Sao Paulo	SP	2020-03-16	88136	1104.21	29605	29287	97.4
CD38	EPI_ISL_476488	No	Santana de Parnaiba	SP	2020-03-16	47079	579.429	29515	28991	96.9
CD39	EPI_ISL_4297880	Yes	Sao Paulo	SP	2020-03-16	98513	1230.56	29327	29078	97.4
CD41	EPI_ISL_1084733	No	Sao Paulo	SP	2020-03-16	48675	603.222	29287	29040	97.4
CD43	EPI_ISL_4297881	Yes	Sao Paulo	SP	2020-03-16	24035	303.04	29287	28888	96.5
CD44	EPI_ISL_4297882	Yes	Sao Paulo	SP	2020-03-17	45850	573.558	26872	25549	82.9
CD45	EPI_ISL_4297883	Yes	Sao Paulo	SP	2020-03-17	22577	279.997	29169	27716	94.6
CD46	EPI_ISL_4297884	Yes	Sao Paulo	SP	2020-03-17	42829	533.776	28149	27170	89.3
CD473	EPI_ISL_476372	No	Sao Paulo	SP	2020-04-01	76230	966.8	27332	25152	83.7
CD474	EPI_ISL_476373	No	Sao Paulo	SP	2020-04-01	229448	2825.63	29622	29483	98.7
CD475	EPI_ISL_476374	No	Sao Paulo	SP	2020-04-02	73561	932.902	29204	28712	96.1
CD476	EPI_ISL_476375	No	Osasco	SP	2020-03-20	77453	972.914	29463	29457	97.9
CD477	EPI_ISL_476376	No	Cubatao	SP	2020-04-01	100531	1265.78	29192	28564	95.2
CD479	EPI_ISL_476377	No	Sao Paulo	SP	2020-04-01	137720	1666.51	28386	27507	92.7
CD48	EPI_ISL_4348336	Yes	Sao Paulo	SP	2020-03-17	89974	1140.18	29044	29034	96.4

CD481	EPI_ISL_476378	No	Ribeirao Preto	SP	2020-04-01	73322	929.21	26710	24749	80.4
CD483	EPI_ISL_476379	No	Sao Paulo	SP	2020-04-05	135173	1701.37	29460	29385	97.6
CD484	EPI_ISL_476380	No	Itaquaquecetuba	SP	2020-04-05	126972	1604.03	29467	29457	98.0
CD485	EPI_ISL_476381	No	Diadema	SP	2020-04-05	123389	1536.73	27368	26279	86.3
CD487	EPI_ISL_476383	No	Sao Paulo	SP	2020-04-04	137022	1728.5	29467	29458	98.0
CD488	EPI_ISL_476384	No	Santo Andre	SP	2020-04-03	39061	487.111	29245	27570	95.6
CD49	EPI_ISL_4348337	Yes	Sao Paulo	SP	2020-03-18	50478	634.468	29489	29070	96.5
CD490	EPI_ISL_476386	No	Sao Paulo	SP	2020-04-06	140378	1755.94	29465	29038	97.9
CD491	EPI_ISL_4297885	Yes	Sao Paulo	SP	2020-04-05	28197	342.312	29560	28505	95.2
CD492	EPI_ISL_4297886	Yes	Osasco	SP	2020-04-05	16177	205.257	25995	23778	78.5
CD495	EPI_ISL_4297887	Yes	Sao Paulo	SP	2020-04-03	11925	147.621	27027	23858	81.7
CD496	EPI_ISL_4297888	Yes	Francisco Morato	SP	2020-04-06	24325	299.001	28565	27680	91.7
CD497	EPI_ISL_4297889	Yes	Sao Paulo	SP	2020-04-07	33136	407.832	29432	28084	95.3
CD498	EPI_ISL_4297890	Yes	Sao Paulo	SP	2020-04-07	16591	206.238	26124	23909	80.8
CD499	EPI_ISL_4297891	Yes	Sao Paulo	SP	2020-04-07	30383	373.876	29567	28835	95.4
CD501	EPI_ISL_4297892	Yes	Sao Paulo	SP	2020-04-07	8399	104.272	27820	24799	81.9
CD502	EPI_ISL_4297893	Yes	Sao Paulo	SP	2020-04-07	39566	484.976	28399	26983	91.2
CD503	EPI_ISL_4297894	Yes	Sorocaba	SP	2020-04-08	22631	277.592	29228	28746	95.2
CD504	EPI_ISL_4297895	Yes	Poa	SP	2020-04-07	39593	487.6	29674	29657	98.9
CD505	EPI_ISL_4297896	Yes	Sao Paulo	SP	2020-04-08	24050	297.846	27953	26564	88.2
CD51	EPI_ISL_4468752	Yes	Sao Paulo	SP	2020-03-18	48029	609.5	27241	25667	85.6
CD511	EPI_ISL_4297898	Yes	Sao Paulo	SP	2020-04-10	38607	474.391	28154	26303	88.2
CD515	EPI_ISL_4297899	Yes	Sao Paulo	SP	2020-04-10	10191	122.774	25351	23254	77.2
CD518	EPI_ISL_4297900	Yes	Sao Paulo	SP	2020-04-11	12660	151.008	26360	24008	79.0
CD52	EPI_ISL_4348338	Yes	Sao Paulo	SP	2020-03-15	44635	554.686	29097	28600	96.6
CD520	EPI_ISL_1084739	No	Sao Paulo	SP	2020-04-12	19540	240.413	27693	26564	87.6
CD525	EPI_ISL_672687	No	Sao Paulo	SP	2020-04-11	109980	1380.34	29465	29461	97.9
CD527	EPI_ISL_672688	No	Sao Paulo	SP	2020-04-14	87200	1101.11	29484	29229	97.9
CD528	EPI_ISL_672689	No	Sao Paulo	SP	2020-04-10	77425	896.929	25683	24112	79.5
CD529	EPI_ISL_672690	No	Sao Paulo	SP	2020-04-13	61332	769.405	29048	27807	93.4
CD53	EPI_ISL_4297901	Yes	Sao Paulo	SP	2020-03-18	32518	421.449	26742	25640	84.2

CD530	EPI_ISL_672691	No	Sao Paulo	SP	2020-04-13	94341	1173.91	28787	27663	92.7
CD532	EPI_ISL_672692	No	Sao Paulo	SP	2020-04-14	99021	1152.67	25130	22673	77.1
CD533	EPI_ISL_672693	No	Carapicuiba	SP	2020-04-11	83537	1043.22	29220	28867	96.0
CD534	EPI_ISL_672694	No	Sao Paulo	SP	2020-04-14	113237	1421.98	28798	27919	93.5
CD537	EPI_ISL_672695	No	Carapicuiba	SP	2020-04-15	135664	1703.09	29455	29019	97.0
CD538	EPI_ISL_672696	No	Sao Paulo	SP	2020-04-14	61630	777.989	29263	28951	96.1
CD539	EPI_ISL_672697	No	Sao Paulo	SP	2020-04-15	83953	1053.95	28626	28473	94.4
CD54	EPI_ISL_476428	No	Taboao da Serra	SP	2020-03-18	123850	1543.76	29245	28629	96.1
CD540	EPI_ISL_672698	No	Sao Paulo	SP	2020-04-15	63383	807.511	27724	27514	90.1
CD541	EPI_ISL_672699	No	Carapicuiba	SP	2020-04-15	88326	1117.95	29250	28782	96.2
CD542	EPI_ISL_672722	No	Sao Paulo	SP	2020-04-16	111549	1412.65	29275	28997	96.2
CD543	EPI_ISL_672723	No	Sao Paulo	SP	2020-04-16	96021	1222.33	27501	26754	88.3
CD545	EPI_ISL_672724	No	Sao Paulo	SP	2020-04-17	127010	1563.7	28765	27769	91.7
CD547	EPI_ISL_672725	No	Sao Paulo	SP	2020-04-17	113661	1421.56	29461	29456	97.8
CD548	EPI_ISL_672726	No	Sao Paulo	SP	2020-04-16	147928	1623.33	28358	26902	90.7
CD549	EPI_ISL_672727	No	Fortaleza	CE	2020-04-15	114951	1287.82	27184	24984	83.6
CD550	EPI_ISL_672728	No	Sao Paulo	SP	2020-04-18	134753	1682.33	29612	29257	97.1
CD551	EPI_ISL_672729	No	Santo Andre	SP	2020-04-18	44935	586.03	24853	23013	76.5
CD553	EPI_ISL_672730	No	Sao Paulo	SP	2020-04-18	92380	1175.43	28421	27528	92.1
CD554	EPI_ISL_672731	No	Jandira	SP	2020-04-16	111597	1394.08	29488	29460	97.9
CD555	EPI_ISL_672732	No	Guarulhos	SP	2020-04-14	140320	1727.59	28869	28425	94.4
CD556	EPI_ISL_672733	No	Sao Paulo	SP	2020-04-13	106419	1277.59	27768	25804	88.3
CD557	EPI_ISL_672734	No	Sao Paulo	SP	2020-04-14	172160	2019.94	29461	29027	97.1
CD558	EPI_ISL_672735	No	Ferraz de Vasconcelos	SP	2020-04-20	157736	1984.08	29461	29458	97.9
CD559	EPI_ISL_672736	No	Sao Paulo	SP	2020-04-18	116739	1476.24	29435	29185	96.9
CD56	EPI_ISL_1084735	No	Sao Paulo	SP	2020-03-18	23237	286.746	28809	26612	89.3
CD561	EPI_ISL_672737	No	Sao Paulo	SP	2020-04-20	135627	1441.47	27196	25652	86.7
CD564	EPI_ISL_672738	No	Sao Paulo	SP	2020-04-20	96139	1019.69	25985	24357	79.7
CD565	EPI_ISL_672739	No	Sao Paulo	SP	2020-03-26	66925	848.104	29209	29207	97.0
CD566	EPI_ISL_672740	No	Sao Paulo	SP	2020-04-21	131952	1664.68	29223	28368	93.5
CD567	EPI_ISL_672669	No	Sao Paulo	SP	2020-04-20	134644	1535.12	27745	26550	89.0

CD568	EPI_ISL_672741	No	Sao Paulo	SP	2020-04-19	234778	2795.74	29466	29457	97.9
CD57	EPI_ISL_476430	No	Sao Paulo	SP	2020-03-20	41754	535.563	25413	23064	76.8
CD570	EPI_ISL_672742	No	Sao Paulo	SP	2020-04-22	158065	1932.58	29067	27974	93.6
CD572	EPI_ISL_672743	No	Sao Paulo	SP	2020-04-23	96941	1221.04	27323	25921	86.5
CD578	EPI_ISL_672745	No	Guarulhos	SP	2020-04-20	27160	341.853	28602	27367	91.7
CD61	EPI_ISL_476431	No	Sao Paulo	SP	2020-03-21	73827	925.342	29049	28364	94.3
CD623	EPI_ISL_672700	No	Sao Paulo	SP	2020-04-12	93599	1178.02	29204	28526	95.2
CD63	EPI_ISL_476432	No	Sao Paulo	SP	2020-03-22	86292	1080.67	28981	28508	94.1
CD65	EPI_ISL_476433	No	Sao Paulo	SP	2020-03-18	66129	827.86	28648	27076	91.5
CD66	EPI_ISL_476434	No	Sao Paulo	SP	2020-03-25	108446	1370.37	28172	27686	90.0
CD67	EPI_ISL_476435	No	Sao Paulo	SP	2020-03-19	206616	2584.53	29836	29835	98.7
CD69	EPI_ISL_476436	No	Sao Paulo	SP	2020-03-23	77974	977.901	28569	27104	90.1
CD70	EPI_ISL_476437	No	Sao Paulo	SP	2020-03-23	126712	1589.14	29209	28996	96.0
CD71	EPI_ISL_476438	No	Sao Paulo	SP	2020-03-23	124416	1510.78	28105	27099	88.3
CD72	EPI_ISL_476439	No	Sao Paulo	SP	2020-03-24	186287	2333	29836	29835	98.6
CD73	EPI_ISL_476440	No	Sao Paulo	SP	2020-03-24	78479	1019.37	26043	23749	79.8
CD74	EPI_ISL_476441	No	Sao Paulo	SP	2020-03-25	133179	1665.57	29815	29361	96.9
CD75	EPI_ISL_476442	No	Sao Paulo	SP	2020-03-25	99343	1222.57	26806	24969	81.1
CD76	EPI_ISL_476443	No	Sao Paulo	SP	2020-03-19	116959	1463.61	29833	29385	97.9
CD788	EPI_ISL_722057	No	Niteroi	RJ	2020-04-02	58354	737.503	28131	27542	90.8
CD789	EPI_ISL_722080	No	Curitiba	PR	2020-04-02	79447	997.21	29053	28143	93.5
CD79	EPI_ISL_476445	No	Sao Paulo	SP	2020-03-23	93416	1181.27	28769	27609	90.0
CD790	EPI_ISL_722108	No	Sao Paulo	SP	2020-04-02	42148	529.6	26583	24991	83.3
CD791	EPI_ISL_722004	No	Sao Paulo	SP	2020-04-03	45349	569.229	29453	28808	96.8
CD792	EPI_ISL_722078	No	Taboao da Serra	SP	2020-04-03	40231	503.691	28594	28166	93.5
CD795	EPI_ISL_722079	No	Sao Paulo	SP	2020-04-03	59114	739.384	28831	27964	93.6
CD797	EPI_ISL_722065	No	Sao Paulo	SP	29/03/2020	46821	592.291	28623	27888	91.8
CD798	EPI_ISL_722023	No	Sao Paulo	SP	2020-04-03	53399	669.793	29458	28854	97.8
CD800	EPI_ISL_722119	No	Sao Paulo	SP	2020-04-03	35556	444.949	27822	26197	86.6
CD801	EPI_ISL_722012	No	Sao Paulo	SP	2020-04-03	44273	554.611	29452	28980	97.0
CD802	EPI_ISL_722082	No	Caierias	SP	2020-04-06	35736	448.464	29298	28291	94.0

CD803	EPI_ISL_721987	No	Sao Paulo	SP	2020-04-06	20822	260.668	29457	28282	95.1
CD806	EPI_ISL_722029	No	Cruzeiro	SP	2020-04-06	54086	672.156	29459	29436	97.9
CD807	EPI_ISL_722122	No	Sao Paulo	SP	2020-04-06	30347	378.962	27571	25775	87.7
CD81	EPI_ISL_476446	No	Sao Paulo	SP	2020-03-24	115174	1440.66	29836	29675	98.6
CD812	EPI_ISL_722084	No	Sao Paulo	SP	2020-04-06	30509	383.491	29028	28567	94.4
CD813	EPI_ISL_722063	No	Sao Paulo	SP	2020-04-06	42268	530.851	28099	26720	90.1
CD814	EPI_ISL_722099	No	Sao Paulo	SP	2020-03-16	39044	496.104	26753	24608	81.5
CD815	EPI_ISL_722067	No	Sao Paulo	SP	2020-04-08	34343	431.028	29163	28106	92.5
CD816	EPI_ISL_722068	No	Sao Paulo	SP	2020-04-08	31183	390.924	28599	27360	92.5
CD818	EPI_ISL_722030	No	Sao Paulo	SP	2020-04-08	93195	1166.59	29469	29457	98.0
CD819	EPI_ISL_722031	No	Sao Paulo	SP	2020-04-08	85293	1074.33	29461	29453	98.0
CD82	EPI_ISL_476447	No	Sao Paulo	SP	2020-03-24	118677	1483.26	29809	29182	96.0
CD820	EPI_ISL_722055	No	Sao Paulo	SP	2020-04-08	60449	761.781	27968	27477	90.3
CD822	EPI_ISL_722005	No	Sao Paulo	SP	2020-04-06	70936	893.995	29284	28978	96.9
CD823	EPI_ISL_722000	No	Guarulhos	SP	2020-04-08	114662	1446.86	29085	29059	96.3
CD825	EPI_ISL_722013	No	Mogi das Cruzes	SP	2020-04-09	65666	827.279	29457	29062	97.0
CD827	EPI_ISL_722059	No	Sao Paulo	SP	2020-04-09	93612	1179.65	28447	27272	90.0
CD828	EPI_ISL_722042	No	Sao Paulo	SP	2020-04-09	89017	1122.76	29463	29254	97.9
CD829	EPI_ISL_1172014	No	Sao Paulo	SP	2020-04-09	119250	1501.91	29461	29446	98.0
CD83	EPI_ISL_476448	No	Sao Paulo	SP	2020-03-24	42305	529.617	29455	28602	95.1
CD830	EPI_ISL_722043	No	Sao Paulo	SP	2020-04-06	57666	725.969	29460	29456	97.9
CD832	EPI_ISL_722018	No	Sao Paulo	SP	2020-04-08	84262	1064.81	29459	29260	97.1
CD833	EPI_ISL_722075	No	Mongagua	SP	2020-04-09	71419	907.93	28650	27139	92.7
CD834	EPI_ISL_722090	No	Sao Paulo	SP	2020-04-09	42274	550.586	25131	23123	74.7
CD835	EPI_ISL_722087	No	Sao Paulo	SP	2020-04-24	98892	1248.82	29066	28410	94.5
CD836	EPI_ISL_722032	No	Sao Paulo	SP	2020-04-13	109400	1379.24	29606	29459	97.9
CD837	EPI_ISL_722044	No	Sao Paulo	SP	2020-04-24	92592	1171.59	29461	29184	98.0
CD838	EPI_ISL_722054	No	Embu das Artes	SP	2020-04-14	113913	1422.86	28187	27051	90.1
CD839	EPI_ISL_722033	No	Carapicuiba	SP	2020-04-14	142469	1802.65	29464	29458	97.9
CD840	EPI_ISL_722045	No	Sao Paulo	SP	2020-04-15	104844	1318.98	29570	29461	98.0
CD841	EPI_ISL_721992	No	Sao Paulo	SP	2020-04-15	79095	981.914	29008	28618	95.4

CD842	EPI_ISL_722127	No	Sao Paulo	SP	2020-04-15	86274	1065.42	28390	27236	89.1
CD847	EPI_ISL_722083	No	Sao Paulo	SP	2020-04-16	49956	617.73	28793	28084	93.3
CD848	EPI_ISL_722085	No	Sao Paulo	SP	2020-04-16	45678	566.059	28817	28592	94.4
CD849	EPI_ISL_722034	No	Vargem Grande Paulista	SP	2020-04-16	70820	878.6	29615	29462	97.9
CD851	EPI_ISL_722101	No	Sao Paulo	SP	2020-04-17	42820	531.941	26149	24430	81.5
CD852	EPI_ISL_722014	No	Sao Paulo	SP	2020-04-17	46461	575.073	29445	29242	97.1
CD853	EPI_ISL_722035	No	Sao Paulo	SP	2020-04-17	94658	1172.55	29595	29458	97.9
CD854	EPI_ISL_722114	No	Sao Paulo	SP	2020-04-17	62519	785.366	27799	26314	85.5
CD855	EPI_ISL_722072	No	Itapevi	SP	2020-04-22	65683	814.038	28746	27487	91.8
CD856	EPI_ISL_722009	No	Sao Paulo	SP	2020-04-22	36582	447.696	29459	29215	97.0
CD857	EPI_ISL_722003	No	Sao Paulo	SP	2020-04-22	75045	926.763	29072	29058	96.3
CD858	EPI_ISL_722036	No	Sao Paulo	SP	2020-04-22	62722	771.889	29451	29209	97.0
CD859	EPI_ISL_722025	No	Sao Paulo	SP	2020-04-22	27069	334.514	29459	29457	98.0
CD86	EPI_ISL_476463	No	Sao Paulo	SP	2020-03-19	107501	1307.77	27651	25580	85.7
CD860	EPI_ISL_722015	No	Sao Paulo	SP	2020-04-22	54638	674.279	29021	28588	96.1
CD861	EPI_ISL_722053	No	Santos	SP	2020-04-22	49020	617.106	27943	26378	88.5
CD862	EPI_ISL_722056	No	Itapecerica da Serra	SP	2020-04-22	24229	297.033	27983	26747	89.4
CD863	EPI_ISL_722073	No	Betim	MG	2020-04-22	63248	787.19	28258	27941	91.8
CD864	EPI_ISL_722088	No	Taboao da Serra	SP	2020-04-22	72866	902.116	28811	28388	93.6
CD865	EPI_ISL_722052	No	Sao Paulo	SP	2020-04-22	19277	241.341	27845	26993	88.9
CD867	EPI_ISL_722024	No	Sao Paulo	SP	2020-04-22	56958	698.706	29346	29022	97.0
CD868	EPI_ISL_722112	No	Sao Paulo	SP	2020-04-22	33239	414.859	26864	25696	83.0
CD869	EPI_ISL_722026	No	Sao Paulo	SP	2020-04-22	58877	727.822	29464	29459	97.9
CD87	EPI_ISL_476464	No	Franco da Rocha	SP	2020-03-22	119209	1474.11	29660	29359	96.9
CD870	EPI_ISL_722001	No	Sao Paulo	SP	2020-04-22	171146	2083.47	29098	29074	96.2
CD872	EPI_ISL_4463732	Yes	Sao Paulo	SP	2020-04-22	167208	2078.27	29283	28732	95.3
CD873	EPI_ISL_722046	No	Carapicuiba	SP	2020-04-23	100380	1261.42	29584	29457	98.0
CD874	EPI_ISL_722021	No	Mogi das Cruzes	SP	2020-04-23	132896	1662.19	29461	29063	97.2
CD875	EPI_ISL_722037	No	Sao Paulo	SP	2020-04-23	167929	2082.99	29486	29464	98.0
CD876	EPI_ISL_722047	No	Sao Paulo	SP	2020-04-23	248441	3008.91	29472	29459	97.9

CD877	EPI_ISL_722048	No	Sao Paulo	SP	2020-04-23	108612	1366.6	29485	29471	97.9
CD878	EPI_ISL_721989	No	Sao Paulo	SP	2020-04-23	153805	1892.97	29035	28247	95.3
CD879	EPI_ISL_722027	No	Guarulhos	SP	2020-04-23	127196	1592.73	29458	29457	97.9
CD88	EPI_ISL_476465	No	Cotia	SP	2020-03-23	58372	723.188	28909	28088	93.5
CD880	EPI_ISL_722038	No	Guarulhos	SP	2020-04-23	160425	1979.84	29489	29475	98.0
CD881	EPI_ISL_722129	No	Ribeirao Pires	SP	2020-04-23	205101	2523.22	29625	29463	98.7
CD882	EPI_ISL_722028	No	Sao Paulo	SP	2020-04-23	58656	736.312	29443	29434	97.9
CD883	EPI_ISL_722049	No	Barueri	SP	2020-04-23	121355	1522.82	29462	29456	97.9
CD89	EPI_ISL_476466	No	Osasco	SP	2020-03-24	26016	324.992	28090	25963	88.2
CD891	EPI_ISL_721996	No	Sao Paulo	SP	2020-04-24	46890	585.906	29271	29018	96.2
CD892	EPI_ISL_721997	No	Sao Paulo	SP	2020-04-24	50420	631.056	29271	29018	96.2
CD894	EPI_ISL_722110	No	Sao Paulo	SP	2020-04-24	16008	201.505	27307	24934	83.7
CD895	EPI_ISL_722102	No	Sao Paulo	SP	2020-04-24	42099	530.277	27031	25288	82.0
CD896	EPI_ISL_722091	No	Carapicuiba	SP	2020-04-24	22799	294.833	25057	23352	75.2
CD898	EPI_ISL_722061	No	Osasco	SP	2020-04-24	60293	751.503	29013	26341	91.4
CD899	EPI_ISL_721999	No	Sao Paulo	SP	2020-04-23	26940	338.261	29272	29018	96.2
CD90	EPI_ISL_476489	No	Sao Paulo	SP	2020-03-24	81155	1026.92	29834	29673	97.9
CD900	EPI_ISL_722039	No	Sao Paulo	SP	2020-04-14	53573	669.285	29461	29457	97.9
CD901	EPI_ISL_722008	No	Carapicuiba	SP	2020-04-24	38875	485.618	29457	29204	97.0
CD902	EPI_ISL_722051	No	Sao Paulo	SP	2020-04-24	24677	309.225	29459	29207	97.9
CD905	EPI_ISL_722076	No	Itapecerica da Serra	SP	2020-04-20	38183	480.91	29051	28036	93.5
CD907	EPI_ISL_4297902	Yes	Franco da Rocha	SP	2020-04-24	22528	286.073	25343	23500	76.8
CD91	EPI_ISL_476490	No	Sao Paulo	SP	2020-03-24	75295	953.394	29836	29744	98.6
CD910	EPI_ISL_722019	No	Sao Paulo	SP	2020-04-13	59049	738.823	29278	28999	97.1
CD917	EPI_ISL_722124	No	Sao Paulo	SP	2020-04-27	31168	393.726	28312	26547	88.3
CD918	EPI_ISL_722081	No	Itapevi	SP	2020-04-27	50142	629.492	29183	28061	93.5
CD92	EPI_ISL_476203	No	Sao Paulo	SP	2020-03-24	98925	1246.21	27521	26976	88.3
CD920	EPI_ISL_722109	No	Sao Paulo	SP	2020-04-27	33615	426.912	27836	25533	83.7
CD921	EPI_ISL_722016	No	Carapicuiba	SP	2020-04-25	69007	854.987	29281	29264	97.1
CD923	EPI_ISL_722060	No	Sao Paulo	SP	2020-04-27	101458	1256.98	28156	26849	91.0

CD924	EPI_ISL_722125	No	Jandira	SP	2020-04-28	52450	660.218	27756	26494	89.0
CD926	EPI_ISL_722126	No	Sao Paulo	SP	2020-04-28	87786	1086.25	27998	26905	89.0
CD927	EPI_ISL_722020	No	Sao Paulo	SP	2020-04-28	65734	810.487	29284	29273	97.1
CD928	EPI_ISL_722071	No	Sao Paulo	SP	2020-04-27	82248	1026.99	28853	27425	92.6
CD930	EPI_ISL_722064	No	Sao Paulo	SP	2020-04-29	66769	827.223	28360	27643	91.8
CD931	EPI_ISL_722103	No	Francisco Morato	SP	2020-04-29	52672	662.221	26580	24364	82.0
CD932	EPI_ISL_721993	No	Sao Paulo	SP	2020-04-29	122577	1493.13	28898	28861	95.4
CD934	EPI_ISL_721994	No	Sao Paulo	SP	2020-04-29	98723	1214.15	29068	28857	95.4
CD935	EPI_ISL_722116	No	Sao Paulo	SP	2020-04-29	51982	655.416	27557	25811	86.4
CD937	EPI_ISL_722041	No	Taboao da Serra	SP	2020-04-29	59889	738.907	29451	29243	97.9
CD939	EPI_ISL_721998	No	Sao Paulo	SP	2020-04-29	46828	578.545	29266	28995	96.2
CD94	EPI_ISL_476205	No	Sao Paulo	SP	2020-03-19	107126	1355.36	28832	27759	94.3
CD940	EPI_ISL_722094	No	Sao Paulo	SP	2020-04-29	25961	326.325	25629	23820	76.7
CD941	EPI_ISL_722092	No	Sao Paulo	SP	2020-05-01	30014	383.3	25513	21950	75.3
CD942	EPI_ISL_722086	No	Sao Paulo	SP	2020-04-30	54293	680.262	28846	28414	94.4
CD943	EPI_ISL_722107	No	Sao Paulo	SP	2020-04-29	50676	635.477	26943	25040	83.6
CD945	EPI_ISL_722062	No	Sao Paulo	SP	2020-04-30	78566	974.789	28325	27208	91.1
CD952	EPI_ISL_722040	No	Sao Paulo	SP	2020-05-01	45297	561.732	29460	29457	97.9
CD954	EPI_ISL_722117	No	Sao Paulo	SP	2020-05-01	47146	585.119	28070	26173	86.6
CD957	EPI_ISL_722098	No	Sao Paulo	SP	2020-03-05	53716	691.788	26088	23803	79.9
CD96	EPI_ISL_476207	No	Sao Paulo	SP	2020-03-21	121343	1517.52	27674	27049	88.3
CD965	EPI_ISL_722022	No	Sao Paulo	SP	2020-04-30	70494	872.811	29429	29032	97.1
CD97	EPI_ISL_476208	No	Diadema	SP	2020-03-23	123574	1522.68	27458	26318	86.2
CD971	EPI_ISL_722120	No	Sao Paulo	SP	2020-05-05	94548	1186.08	27780	26518	87.2
CD972	EPI_ISL_722095	No	Aruja	SP	2020-05-05	51812	671.907	24563	23071	76.0
CD973	EPI_ISL_722097	No	Sao Paulo	SP	2020-05-05	50456	634.107	25404	23871	78.7
CD975	EPI_ISL_722111	No	Sao Paulo	SP	2020-05-05	70510	882.501	27093	25595	83.4
CD98	EPI_ISL_476467	No	Sao Paulo	SP	2020-03-23	92861	1154.77	29144	28115	95.2
CD981	EPI_ISL_722093	No	Sao Paulo	SP	2020-05-07	46698	606.117	24757	22924	75.8
CD986	EPI_ISL_722096	No	Sao Paulo	SP	2020-05-08	81399	950.075	25435	22492	76.1
CD99	EPI_ISL_476468	No	Sao Paulo	SP	2020-03-24	18420	227.765	28908	27906	92.5

CD990	EPI_ISL_722077	No	Sao Paulo	SP	2020-05-04	117215	1457.34	29027	28143	92.6
CD991	EPI_ISL_722105	No	Sao Paulo	SP	2020-05-04	75776	950.611	27264	25440	82.8
CD992	EPI_ISL_722121	No	Sao Paulo	SP	2020-05-04	70923	889.795	27524	26012	87.3
CD995	EPI_ISL_722113	No	Franco da Rocha	SP	2020-05-04	50108	630.562	27389	25525	83.5
CD996	EPI_ISL_4463733	Yes	Osasco	SP	2020-05-05	47399	598.43	25368	22970	75.1
CD997	EPI_ISL_1172017	No	Sao Paulo	SP	2020-05-05	76418	943.165	27850	26436	88.8

Table S4. Statistical support for 16 hospital-associated transmission clusters from HC complex.

	Dataset B				Dataset C				
Cluster	**alrt**	**FB**	**alrt_2**	**FB_2**	**alrt**	**FB**	**MCC**	**MCC DTA MJ**	**MCC DTA BSSVS**
A	78.4	100	84.8	100	86.5	98	0.9043	0.9015	0.9002
B	78.2	100	85.9	100	75.7	100	0.9899	0.9922	0.9921
C	90.2	99	88.2	100	89.1	99	1	1	1
D	76.4	100	85.5	100	93.7	100	1	1	1
E	75.9	100	78.6	100	79.1	99	0.9978	1	1
F	91.5	100	90.5	100	93	99	1	1	1
G	91.9	99	92.3	100	78.2	99	0.9906	0.9884	0.9892
H	93.1	100	92.4	100	83	99	0.9996	1	1
I	95	100	94.6	100	92.1	99	1	1	1
J	85.5	100	85.4	100	87.1	100	0.9996	1	1
K	92.4	100	92.2	100	85.6	100	0.9991	0.9988	0.9988
L	84.9	100	90.9	100	92.6	100	0.9994	0.9997	0.9992
M	92.2	100	90.4	99	93.4	100	1	1	1
N	85.7	99	88.1	99	87.1	98	0.9962	0.9969	0.9966
O	92	100	91.8	100	93.4	99	0.9999	1	1
P	77.2	100	76	100	76.2	99	1	1	1

Table S5. Summary of epidemiological and genetic characteristics of 16 hospital-associated transmission clusters from HC complex.

	Size	Institute	Pairwise divergence (SNP)				Collection date			Pairwise geo distance	Epidemiological link			Patient classification			
Cluster	n	Main	Mean	Median	Max	Min	Oldest	Youngest	Duration (days)	Median (m)	Strong	Possible	Unclear	Community	>14	HW	Others
A	12	Institute C	2.39	2	6	0	21/03/2020	30/04/2020	40	29331	6	5	1	1	4	7	0
B	10	Institute A	0.40	0	2	0	26/03/2020	01/05/2020	36	20702	3	6	1	2	0	7	1
C	7	Institute C	1.10	1	3	0	16/04/2020	13/05/2020	27	18574	4	2	1	0	2	4	1
D	7	Institute A	0.57	1	2	0	17/03/2020	26/03/2020	9	4574	4	2	1	0	0	7	0
E	5	Mixed	0.67	1	2	0	14/04/2020	22/04/2020	8	10330	0	2	3	2	0	3	0
F	6	Institute A	1.60	2	3	0	14/04/2020	29/04/2020	15	34590	2	3	1	3	0	1	2
G	4	Institute B	0.00	0	0	0	04/05/2020	11/05/2020	7	25264	4	0	0	0	0	4	0
H	4	Mixed	0.00	0	0	0	18/04/2020	29/04/2020	11	36230	2	2	0	1	0	3	0
I	3	Institute B	0.67	1	1	0	22/04/2020	24/04/2020	2	12541	0	3	0	1	0	2	0
J	2	Institute B	1.00	1	1	1	22/04/2020	22/04/2020	0	8410	0	0	2	0	0	2	0
K	2	Institute C	0.00	0	0	0	19/03/2020	23/03/2020	4	14290	0	2	0	0	0	2	0
L	2	Institute A	0.00	0	0	0	22/03/2020	24/03/2020	2	15633	0	2	0	0	0	2	0
M	2	Mixed	1.00	1	1	1	14/04/2020	17/04/2020	3	6899	0	0	2	2	0	0	0
N	2	Institute B	1.00	1	1	1	23/03/2020	23/03/2020	0	2750	2	0	0	0	0	2	0

O	2	Institute B	0.00	0	0	0	27/03/2020	03/04/2020	7	603	2	0	0	0	0	2	0
P	2	Institute B	0.00	0	0	0	03/04/2020	06/04/2020	3	12317	2	0	0	0	0	2	0

Table S6. Defining mutations of 16 hospital-associated transmission clusters from HC complex.

Cluster	Nucleotide	AA
A	C9733T	-
B	G28681T	N:E136D
C	C1912T	-
	C9479T	ORF1a:G3072C
	C14362T	-
D	G27240T	ORF:E13D
E	A1777G	-
F	G19086T	ORF1b:V1467I
	C24096T	S:A845V
G	T9093C	ORF1a:V2943A
H	C3293T	ORF1a:P1010S
I	G17866A	ORF1b:V1467I
	G29422T	-
J	C1884T	ORF1a:A540V
K	G18589T	ORF1b:V1708F
L	C26456T	E:P71L
M	C15738T	-
	C23481T	S:S640F
N	C3874T	-
O	C7869T	ORF1a:S2535L
P	C24023T	-

Table S7. Summary of epidemiological and genetic characteristics of 16 hospital-associated transmission clusters from HC complex per institute.

Institute	Clusters	Cluster size		Duration (days)		Pairwise geographical distance (m)	Epidemiological link (%)		
	n	Mean	Median	Mean	Median	median	Strong	Possible	Unclear
Institute B	6.00	2.50	2.00	2.83	1.50	10363.40	37.04	40.74	22.22
Institute A	4.00	6.25	6.50	15.50	12.00	18167.63	36.00	40.00	16.00
Institute C	3.00	7.00	7.00	23.67	27.00	18574.48	52.17	43.48	4.35

Table S8. Logistic Regression Models for prediction of outcomes clustered vs non-clustered sequences of 234 HC SARS-CoV-2 positive individuals.

Logistic Model Parameters	Level	aOdds Ratio	p-value
Model 1:			
Variables: Institute + HCW/patient			
+ Age + Sex			
Base level: Institute B; Patient	(Intercept)	0.17	**0.002**
	Institute A	3.48	**0.00074**
	Institute C	4.17	**0.0002**
	HCW	1.63	0.2111
Model 2:			
Variable: HW/PT per Institute + Age +			
Sex			
Base level: Patient.Institute B	(Intercept)	0.11	**0.0034**
	HCW.Institute A	7.95	**0.0033**
	HCW.Institute B	2.36	**0.01676**
	HCW.Institute C	7.49	**0.0043**
	Patient.Institute A	4.45	**0.0266**
	Patient.Institute C	7.43	**0.00498**
Model 3:			
Variables: Institute + Occupation			
+ Age + Sex			
Base level: Institute B; Patient	(Intercept)	0.0757	**0.0002**

	Institute A	4.4953	**0.0002**
	Institute C	5.1756	**0.0001**
		0.0000	0.9892
	Administration	2.8086	0.0936
	Doctor	1.2141	0.7186
	Medical Resident	6.7498	**0.0056**
	Nurse	3.7721	0.1157
	Nurse Technician	1.3628	0.5404
	Other	1.1490	0.8479
Model 4:			
Variable: Occupations per institute +			
Age + Sex			
Base level: Patient.Institute B	(Intercept)	0.0468	**0.0002**
	Administration.Institute A	0.0000	0.9931
	Administration.Institute B	5.7929	**0.0235**
	Administration.Institute C	6.5012	0.2213
	Doctor.Institute A	22.4661	**0.0012**
	Doctor.Institute B	0.7448	0.8067
	Doctor.Institute C	3.8426	0.1848
	Medical Resident.Institute A	317509584	0.9935
	Medical Resident.Institute B	10.1226	**0.0065**

	Medical Resident.Institute C	29.6604	**0.0185**
	Nurse.Institute A	10.4206	0.1323
	Nurse.Institute B	3.1991	0.3898
	Nurse.Institute C	107332973	0.9894
	Nurse Technician.Institute A	7.7753	**0.0325**
	Nurse Technician.Institute B	2.0146	0.3440
	Nurse Technician.Institute C	7.6647	**0.0318**
	Other.Institute A	6.4456	0.1067
	Other.Institute B	0.9923	0.9949
	Other.Institute C	10470286	0.9939
	Patient.Institute A	4.6535	**0.0241**
	Patient.Institute C	9.1490	**0.0025**

Table S9. Compartmentalization analysis results for 73 clustered sequences from HC complex according to different traits.

Dataset	n	Trait	AI*	Bootstraps
All clustered sequences	73	Institute	0.4517	1000
		HCW/patient	0.7699	993
		Occupation	0.7499	832
		Institute B vs Others	0.5583	998
		Institute A vs Others	0.3675	1000
		Institute C vs Others	0.439	999
Institute A clustered	23	HCW/patient	0.36435	999
		Occupation	0.4203	1000
Institute B clustered	27	HCW/patient	0.9275	433
		Occupation	0.8634	808
Institute C clustered	23	HCW/patient	0.862	540
		Occupation	0.92875	

*Simmond's Association Index

Table S10. Marjov Jumps counts, BSSVS rates and Bayes factors for all Location trait transitions.

Origin	Destination	Counts	Counts 95% BCI	Rates	Rates 95% BCI	Bayes Factor
Other	SP	30.465	27, 34	2.231	0.8533, 3.7905	**29353.794**
SP	Institute A	29.107	23, 34	1.333	0.4798, 2.3163	**29353.794**
SP	Institute B	104.918	94, 113	4.359	1.7017, 7.2442	**29353.794**
SP	Institute C	20.761	17, 24	1.003	0.3523, 1.7782	**29353.794**
Institute A	Institute B	4.405	2, 8	0.765	0.02241, 7.1799	**113.699**
Institute C	Institute B	4.472	1, 7	0.828	0.0824, 1.7819	**84.371**
Institute B	SP	9.763	1, 19	1.381	0.1162, 2.9797	**26.817**
Institute B	Institute A	5.410	0, 10	0.856	2.8247E-3, 2.1398	7.112
Other	Institute C	2.150	1, 4	0.618	2.0054E-4, 2.3907	4.249
Institute C	SP	2.163	0, 5	0.844	1.2088E-3, 2.6343	3.248
Institute C	Other	0.722	0, 2	0.637	8.6213E-4, 2.2403	2.946
Institute B	Institute C	2.200	0, 5	0.872	9.9104E-4, 2.8171	1.677
Institute C	Institute A	1.034	0, 7	0.984	3.5649E-5, 3.2025	0.858
Institute B	Other	0.167	0, 1	1.015	9.4685E-4, 3.2935	0.551
Institute A	SP	0.766	0, 3	1.079	5.6372E-4, 3.3461	0.481
SP	Other	0.703	0, 2	1.072	1.8106E-4, 3.3926	0.343
Institute A	Institute C	0.311	0, 1	1.156	1.0516E-4, 3.5092	0.280
Institute A	Other	0.058	0, 1	1.123	4.0284E-4, 3.4349	0.270
Other	Institute A	0.423	0, 1	1.169	6.4777E-6, 3.483	0.172
Other	Institute B	0.821	0, 2	1.159	5.5896E-4, 11.3425	0.156

Bibliography

1. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis. 2020 Nov;20(11):1263–72.

2. Hamilton WL, Tonkin-Hill G, Smith ER, Aggarwal D, Houldcroft CJ, Warne B, et al. Genomic epidemiology of COVID-19 in care homes in the east of England. eLife. 2021 Mar 2;10.

3. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017 Jun;12(6):1261–76.

4. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020 Sep 4;369(6508):1255–60.

5. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 2021 May 21;372(6544):815–21.

6. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017 Mar 30;22(13):30494.

7. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. 2017 Jan;1(1):33–46.

8. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. Gisaid's role in pandemic response. China CDC Wkly. 2021 Dec 3;3(49):1049–51.

9. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772–80.

10. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020 May 1;37(5):1530–4.

11. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun;14(6):587–9.

12. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016 Jan;2(1):vew007.

13. Rambaut A, Holmes EC, Hill V, OToole A, McCrone J, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. BioRxiv. 2020 Apr 19;

14. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 2015 May 26;1(1):vev003.

15. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018 Jan;4(1):vey016.

16. Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, et al. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol. 2012 Jan;61(1):170–3.

17. Hong SL, Lemey P, Suchard MA, Baele G. Bayesian phylogeographic analysis incorporating predictors and individual travel histories in BEAST. Curr Protoc. 2021 Apr;1(4):e98.

18. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in bayesian phylogenetics using tracer 1.7. Syst Biol. 2018 Sep 1;67(5):901–4.

19. O'Brien JD, Minin VN, Suchard MA. Learning to count: robust estimates for labeled distances between molecular sequences. Mol Biol Evol. 2009 Apr;26(4):801–14.

20. Edwards CJ, Suchard MA, Lemey P, Welch JJ, Barnes I, Fulton TL, et al. Ancient hybridization and an Irish origin for the modern polar bear matriline. Curr Biol. 2011 Aug 9;21(15):1251–8.

21. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006 May;4(5):e88.

22. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021 Jul;19(7):409–24.

23. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005 Mar 1;21(5):676–9.