medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Using multiple sampling strategies to estimate SARS-CoV-2 epidemiological parameters from genomic sequencing data

AUTHOR LIST AND AFFILIATIONS

Rhys P. D. Inward[1,6], Kris V. Parag[2,3,5,6], Nuno R. Faria[1,2,4,5,6]

1. Department of Zoology, University of Oxford, Oxford, UK
2. MRC Centre of Global Infectious Disease Analysis, Jameel Institute for Disease and Emergency Analytics, Imperial College London, London, UK
3. NIHR Health Protection Research Unit in Behavioural Science and Evaluation, University of Bristol, Bristol, UK
4. Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de Sao Paulo, Sao Paulo, Brazil
5. Jointly supervised this work
6. Corresponding author: E-mail: rhys.inward@zoo.ox.ac.uk, k.parag@imperial.ac.uk, n.faria@imperial.ac.uk

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

ABSTRACT

SARS-CoV-2 virus genomes are currently being sequenced at an unprecedented pace. The choice of viral sequences used in genetic and epidemiological analysis is important as it can induce biases that detract from the value of these rich datasets. This raises questions about how a set of sequences should be chosen for analysis, and which epidemiological parameters derived from genomic data are sensitive or robust to changes in sampling. We provide initial insights on these largely understudied problems using SARS-CoV-2 genomic sequences from Hong Kong and the Amazonas State, Brazil. We consider sampling schemes that select sequences uniformly, in proportion or reciprocally with case incidence and which simply use all available sequences (unsampled). We apply *Birth-Death Skyline* and *Skygrowth* methods to estimate the time-varying reproduction number (R_t) and growth rate (r_t) under these strategies as well as related R_0 and date of origin parameters. We compare these to estimates from case data derived from *EpiFilter*, which we use as a reference for assessing bias. We find that both R_t and r_t are sensitive to changes in sampling whilst R_0 and the date of origin are relatively robust. Moreover, we find that the unsampled datasets, which reflect an opportunistic sampling scheme, engender the most biased R_t and r_t estimates for both our Hong Kong and Amazonas case studies. We highlight that sampling strategy choices may be an influential yet neglected component of sequencing analysis pipelines. More targeted attempts at genomic surveillance and epidemic analyses, particularly in resource-poor settings with limited sequencing capabilities, are necessary to maximise the informativeness of virus genomic datasets.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped single-stranded zoonotic RNA virus belonging to the *Betacoronavirus* genus and *Coronaviridae* family[1]. It was first identified in late 2019 in a live food market in Wuhan City, Hubei Province, China[2]. Within a month, SARS-CoV-2 had disseminated globally through sustained human-to-human transmission. It was declared a public health emergency of international concern on the 30th of January 2020 by the World Health Organisation[3]. Those infected with SARS-CoV-2 have phenotypically diverse symptoms ranging from mild fever to multiple organ dysfunction syndromes and death[4].

Despite the implementation of non-pharmaceutical interventions (NPIs) by many countries to control their epidemics, to date over 418 million SARS-CoV-2 cases and 5.8 million deaths have been reported worldwide[5]. These NPIs can vary within and between countries and include restrictions on international and local travel, school closures, social distancing measures and the isolation of infected individuals and their contacts[6]. The key aim of NPIs is to reduce epidemic transmission, often measured by epidemiological parameters such as the time-varying effective reproduction number (R_t at time t) and growth rate (r_t), which both provide updating measures of the rate of spread of a pathogen (see Supplementary Table 1 for detailed definitions)[7,8].

However, there is currently great difficulty in estimating and comparing epidemiological parameters derived from case and death data globally due to disparities in molecular diagnostic surveillance and notification systems between countries. Further, even if data are directly comparable, the choice of epidemiological parameter can implicitly shape insights into how NPIs influence transmission potential[9,10]. As such, there is a need to supplement traditional estimates with information derived from alternative data sources, such as genomic data[11], to gain improved and more robust insights into viral transmission dynamics[12,13].

Phylodynamic analysis of virus genome sequences have increasingly been used for studying emerging infectious diseases, as seen during the current SARS-CoV-2 pandemic[14–17], recent Ebola virus epidemics in Western Africa[18] and the Zika epidemic in Brazil and the Americas[19,20]. Transmissibility parameters such as the basic reproduction number (R_0), R_t and r_t can be directly inferred from genomic sequencing data or from epidemiological data, while

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

other epidemiological parameters such as the date of origin of a given viral variant or lineage can only be estimated from genomic data. This is of particular importance for variants of concern (VOC), genetic variants with evidence of increased transmissibility, more severe disease, and/or immune evasion. VOC are typically detected through virus genome sequencing and only limited inferences can be made through epidemiological data alone[21].

Currently, SARS-CoV-2 virus genomes from COVID-19 cases are being sequenced at an unprecedented pace providing a wealth of virus genomic datasets[22]. There are currently over 8.4 million genomic sequences available on GISAID, an open-source repository for influenza and SARS-CoV-2 genomic sequences[23]. These rich datasets can be used to provide an independent perspective on pathogen dynamics and can help validate or challenge parameters derived from epidemiological data. Specifically, the genomic data can potentially overcome some of the limitations and biases that can result from using epidemiological data alone. For example, genomic data are less susceptible to changes at the government level such as alterations to the definition of a confirmed case and changes to notification systems[24,25]. Inferences from virus genomic data improve our understanding of underlying epidemic spread and can facilitate better-informed infection control decisions[26]. However, these advantages are not straightforward to realise. The added value of genomic data depends on two related variables: sampling strategy and computational complexity.

The most popular approaches used to investigate changes in virus population dynamics include the Bayesian Skyline Plot[27] and Skygrid[28] models and the Birth-Death Skyline (BDSKY)[29]. These integrate Markov Chain Monte Carlo (MCMC) procedures and often converge slowly on large datasets[30]. As such, currently available SARS-CoV-2 datasets containing thousands of sequences become computationally impractical to analyse and sub-sampling is necessary. Although previous studies have examined how sampling choices might influence phylodynamic inferences[30–34], this remains a neglected area of study[35], particularly concerning SARS-CoV-2 for which sequencing efforts have been unprecedented [36]. To our knowledge, there are no published studies concerning SARS-CoV-2 which explore the effect that sampling strategies have on the phylodynamic reconstruction of key transmission parameters. Incorrectly implementing a sampling scheme or ignoring its importance can mislead inferences and introduce biases[30,37].

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

This raises the important question that motivates our analysis: how should sequences be selected for phylodynamic analysis and which parameters are sensitive or robust to changes in different sampling schemes. Here we explore how diverse sampling strategies in genomic sequencing may affect the estimation of key epidemiological parameters. We estimate R_0, R_t, and r_t from genomic sequencing data under different sampling strategies from a location with higher genomic coverage represented by Hong Kong, and a location with lower genomic coverage represented by the Amazonas region, Brazil. We then compare our estimates against those derived from reference case data. By benchmarking genomic inferences against those from case data we can better understand the impact that sampling strategies may have on phylodynamic inference, bolster confidence in estimates of genomic-specific parameters such as the origin time and improve the interpretation of estimates from areas with heterogeneous genomic coverage.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

METHODS

Empirical Estimation of the Reproduction Number, Time-varying Effective Reproduction Number, and Growth Rate

Epidemiological Datasets

Two sources of data from the Amazonas region, Brazil and one source of data from Hong Kong were used to calculate empirical epidemiological parameters. For the Amazonas region, case data from the SIVEP-Gripe (Sistema de Informação de Vigilância Epidemiológica da Gripe) SARI (severe acute respiratory infections) database from the 30th of November 2020 up to 7th of February 2021 were used. Here we were interested in cases caused by the P.1/Gamma VOC first detected in Manaus, the number of P.1 cases was calculated by using the proportion of P.1/Gamma viral sequences uploaded to GISAID within each week (Supplementary Figure 1). For Hong Kong, all case data were extracted from the Centre of Health Protection, Department of Health, the Government of the Hong Kong Special Administrative region up to the 7th of May 2020. Due to lags in the development of detectable viral loads, symptom onset and subsequent testing[38]; the date on which each case was recorded was left shifted by 5 days within our models[39] to account for these delays in both datasets.

Basic Reproduction Number

The R_0 parameter was estimated using a time series of confirmed SARS-CoV-2 cases from both Hong Kong and the Amazonas region. To avoid the impact of NPIs, only data up to the banning of mass gathering in Hong Kong (27th March 2020) and until the imposition of strict restrictions in the Amazonas region (12th January 2021) were used. Weekly counts of confirmed cases were modelled using maximum likelihood methods. The weekly case counts were assumed to be Poisson distributed and were fitted to a closed Susceptible-Exposed-Infectious-Recovered (SEIR) model (Equation 1) by maximising the likelihood of observing the data given the model parameters (Table 1). Subsequently, the log-likelihood was used to calculate the R_0 by fitting β, the effective contact rate.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Equation 1:

$$\lambda = \frac{\beta I}{N}$$
$$\frac{dS}{dt} = -\lambda S$$
$$\frac{dE}{dt} = \lambda S - \gamma E$$
$$\frac{dI}{dt} = \gamma E - \sigma I$$
$$\frac{dR}{dt} = \sigma I$$

To generate approximate 95% confidence intervals (CIs) for R_0, non-parametric bootstrapping was used with 1000 iterations.

Table 1: This shows the parameter estimates used within the deterministic SEIR model.

Parameter	**Description**	**Value (source)**
$R_0 = \beta\alpha$	Basic Reproduction Number	Estimated
N	Population of Hong Kong	7,481,800 persons[40]
	Population of Amazonas Region	4,207,714 persons[41]
β	Effective Contact Rate	Estimated
α	Infectious Period	0.07 day^{-1} [42]
λ	Force of Infection	Estimated
γ	Progression from E to I	5.26 day^{-1} [43]
δ	Progression from I to R	14.3 day^{-1} [42]
S	Estimated number of Susceptibles	Estimated
E	Estimated number of Exposed	Estimated
I	Number of Infected	Weekly case counts
R	Estimated number of Recovered	Estimated

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Time-varying Effective Reproduction Number

To estimate R_t from case line list data the *EpiFilter* method[44] was used. *EpiFilter* describes transmission using a renewal model; a general and popular framework that can be applied to infer the dynamics of numerous infectious diseases from case incidence[45]. This model describes how the number of new cases (incidence) at time t depends on R_t at that specified time point and the past incidence, which is summarised by the cumulative number of cases up to each time point weighted by the generation time distribution, which we assume to be known. *Epifilter* integrates both Bayesian forward and backward recursive smoothing. This improves R_t estimates by leveraging the benefits of two of the most popular R_t estimation approaches: *EpiEstim* [46] and the Wallinga-Teunis method[47]. *EpiFilter* minimises the mean squared error in estimation and reduces dependence on prior assumptions, making it a suitable candidate for deriving reference estimates. We use these to benchmark estimates independently obtained from genomic data. We assume the generation time distribution is well approximated by the serial interval (SI) distribution[46] , which describes the times between symptom onsets between an infector–infectee pair.

Growth Rate

After R_t has been inferred, the Wallinga-Lipsitch equation for a gamma distributed generation time distribution (Equation 2) was used to estimate the exponential epidemic r_t[48]. The SI for Hong Kong was derived from a systematic review and meta-analysis[49] and a study exploring SI in Brazil was used for the Amazonas datasets[50]. The SI was assumed to be gamma distributed. The gamma distribution is represented by gamma (ε, γ) with ε and γ being the shape and scale parameters respectively.

Equation 2:

$$r_t = \varepsilon(R_t^{\left(\frac{1}{\gamma}\right)} - 1)$$

SARS-CoV-2 Brazilian Gamma VOC and Hong Kong datasets

All high-quality, complete SARS-CoV-2 genomes were downloaded from GISAID[23] for Hong Kong (up to 7th May 2020) and the Amazonas state, Brazil (from 30th November 2020 up to 7th February 2021). Using the Accession ID of each sequence, all sequences were screened and only sequences previously analysed and published in PubMed, MedRxiv,

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

BioRxiv, virological or Preprint repositories were selected for subsequent analysis. For both datasets, sequence alignment was conducted using MAFFTV.7[51]. The first 130 base pairs (bp) and last 50 bps of the aligned sequences were trimmed to remove potential sequencing artefacts in line with the Nextstrain protocol[52]. Both datasets were then processed using the Nextclade pipeline for quality control (https://clades.nextstrain.org/). Briefly, the Nextclade pipeline examines the completeness, divergence, and ambiguity of bases in each genetic sequence. Only sequences deemed 'good' by the Nextclade pipeline were selected for. Subsequently, all sequences were screened for identity and in the case of identical sequences, for those with the same location, collection date, only one such isolate was used. Moreover, PANGO lineage classification was conducted using the Pangolin[22] software tool (http://pangolin.cog-uk.io) on sequences from the Amazonas region and only those with the designated P.1/Gamma lineage were selected for (Supplementary Figure 1).

Maximum Likelihood tree reconstruction

Maximum likelihood phylogenetic trees were reconstructed using IQTREE2[53] for both datasets. A TIM2 model of nucleotide substitution with empirical base frequencies and a proportion of invariant sites was used as selected for by the ModelFinder application[54] for the Hong Kong dataset. For the Brazilian dataset, a TN model of nucleotide substitution[55] with empirical base frequencies was selected for. To assess branch support, the approximate likelihood-ratio test based on the Shimodaira–Hasegawa-like procedure with 1,000 replicates[56], was used.

Root-to-tip regression

To explore the temporal structure of both the Brazilian and Hong Kong dataset, TempEst v.1.5.3[57] was used to regress the root-to-tip genetic distances against sampling dates (yyyy-mm-dd). The 'best-fitting' root for the phylogeny was found by maximising the R^2 value of the root-to-tip regression (Supplementary Figure 2). Several sequences showed incongruent genetic diversity and were discarded from subsequent analyses. This resulted in a final dataset of N = 117 Hong Kong sequences and N = 196 Brazilian sequences. The gradient of the slopes (clock rates) provided by TempEst were used to inform the clock prior in the phylodynamic analysis.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Subsampling for analysis

Four retrospective sampling schemes were used to select a subsample of Amazonas and Hong Kong sequences. Each sampling period was broken up into weeks with each week being used as an interval according to a temporal sampling scheme (without replacement). This temporal sampling scheme was based on the number of reported cases of SARS-CoV-2.

The temporal sampling schemes that we explored were:

- **Uniform sampling:** All weeks have equal probability.
- **Proportional sampling**: Weeks are chosen with a probability proportional to the value of the number of cases in each epi-week.
- **Reciprocal-proportional sampling:** Weeks are chosen with a probability proportional to the reciprocal of the number of cases in each epi-week.
- **No sampling strategy applied:** All sequences were included without a sampling strategy applied (equivalent to opportunistic sampling).

These sampling schemes were inspired by those recommended by the WHO for practical use in different settings and scenarios[58]. Proportional sampling is equivalent to representative sampling, uniform sampling is equivalent to fixed sampling whilst the unsampled data includes all sampling strategies. Reciprocal-proportional sampling is not commonly used in practice as was used as a control within this study.

Bayesian Evolutionary Analysis

Date molecular clock phylogenies were inferred for all sampling strategies applied to the Amazonas and Hong Kong dataset using BEAST v1.10.4[59] with BEAGLE library v3.1.0[60] for accelerated likelihood evaluation. For both the Amazonas and Hong Kong datasets, a HKY substitution model with gamma-distributed rate variation among sites and four rate categories was used to account for among-site rate variation[61]. A strict clock molecular clock model was chosen. Both the Amazonas and Hong Kong dataset were analysed under a flexible non-parametric skygrid tree prior[62]. Four independent MCMC chains were run for both datasets. For the Amazonas dataset, each MCMC chain consisted of 250,000,000 steps with sampling every 50,000 steps. Meanwhile, for the Hong Kong dataset, each MCMC chain consisted of 200,000,000 steps with sampling every 40,000 steps. For both datasets, the four independent MCMC runs were combined using LogCombiner v1.10.4[59]. Subsequently, 10% of all trees were discarded as burn in, and the effective sample size of parameter estimates were

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

evaluated using TRACER v1.7.2[63]. An effective sample size of over 200 was obtained for all parameters. Maximum clade credibility (MCC) trees were summarised using Tree Annotator[59].

Phylodynamic Reconstruction

Estimation of the Basic and Time-varying Effective Reproduction Numbers

The Bayesian birth-death skyline (BDSKY) model[29] implemented within BEAST 2 v2.6.5[64] was applied to estimate the time-varying transmissibility parameter R_t (Table 2). A HKY substitution model with a gamma-distributed rate variation among sites and four rate categories[61] was used alongside a strict molecular clock model. The selected number of intervals for both datasets was 5, representing R_t changing every 2.5 weeks for the Hong Kong datasets and every 2 weeks for the Brazilian datasets, with equidistant intervals per step. An exponential distribution was used with a mean of $36.5y^{-1}$ for the rate of becoming infectious, assuming a mean duration of infection of 10 days[15]. A uniform distribution prior was used for the sampling proportion, which models changes in case ascertainment. Four independent MCMC chains were run for 50 million MCMC steps with sampling every 5000 steps for each dataset. These MCMC runs were combined using LogCombiner v2.6.5.[64] and the effective sample size of parameter estimates evaluated using TRACER v1.7.2[63]. We obtained an effective sample size above 200 for all parameters (indicating convergence) and plotted all results using the bdskytools R package (https://github.com/laduplessis/bdskytools).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Table 2: Values and priors for the parameters of the BDSKY model

Parameter	Dataset	Value or prior	Rationale/Assumption
Clock rate	Brazil	4.0×10^{-4} (subs/site/year)	Informed by root-to-tip regression
	Hong Kong	1.0×10^{-4} (subs/site/year)	
Death rate	Brazil and Hong Kong	36.5 y^{-1}	The period between infection and becoming uninfectious assumed an exponential distribution with a mean of 10 days[15]
Reproduction number	Brazil and Hong Kong	Lognormal (0.8, 0.5)	Median 2.2, 95% IQR 0.8 to 5.9
Time of origin	Brazil	Lognormal (-1.50, 0.4) y before present	Median 4th December 2020, 95% IQR 25th September 2020 to 12th January, 2021
	Hong Kong	Lognormal (-1.75, 0.4) y before present	Median 18th January 2020, 95% IQR 17th November 2019 to 15th February 2020
Sampling proportion	Brazil	Uniform (0, 0.024)	196 sequences from 8246 suspected P.1 cases as of 7th February, 2021
	Hong Kong	Uniform (0, 0.116)	117 sequences from 1012 confirmed cases as of 7th May, 2020

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Estimation of Growth Rates

For each dataset, a scaled proxy for r_t was obtained from the *skygrowth* method[65] within R. *Skygrowth* uses a non-parametric Bayesian approach to apply a first-order autoregressive stochastic process on the growth rate of the effective population size. The MCMC chains were run for one million iterations for each dataset on their MCC tree with an Exponential (10^{-5}) prior on the smoothing parameter. The *skygrowth* model was parameterised assuming that the effective population size of SARS-COV-2 could change every two weeks. To facilitate a comparison of the scaled proxy for r_t estimated by *skygrowth* and exponential r_t estimated by *EpiFilter*, the r_t estimated by the *skygrowth* method was rescaled to the exponential growth rate. This was achieved by adding a gamma rate variable to the scaled proxy for r_t, which assumed a mean duration of infection of 10 days[15], to calculate R_t. Subsequently, the Wallinga-Lipsitch equation (Equation 2) was used to convert R_t into the exponential growth rate[48].

Comparing Parameter Estimates from Genetic and Epidemiological Data

To compare estimates derived from epidemiological and genetic data the Jensen-Shannon divergence (D_{JS})[66], which measures the similarity between two probability mass functions (PMFs), was applied. The D_{JS} offers a formal information theoretic evaluation of distributions and is more robust than comparing Bayesian credible intervals (BCIs) since it considers both the shape and spread of a given distribution. The D_{JS} is essentially a symmetric and smoothed version of the Kullback-Leibler divergence (D_{KL}) and is commonly used in the fields of machine learning and bioinformatics. The D_{KL} between two PMFs, P and Q, is defined in Equation 3 below[67].

Equation 3:

$$D_{KL}\,(P||M) = \sum_x P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

To calculate the PMF for each epidemiological parameter, the cumulative probability density function (PDF) was extracted for each model, converted to a probability density function (PDF), and a discretisation procedure then applied τ represents the PDF and is discretized via Equation 4, where s = 0.05, 0.01….and $\tau(\upsilon)$ is the cumulative probability density of τ and i is the incidence.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Equation 4:

$$\tau_{Rt,rt,R0} = \int_{s-0.025}^{s+0.025} \tau(\upsilon)\tau i$$

The Jensen-Shannon distance (JSD) metric quantifies the D_{JS} by taking the square-root of the total D_{JS} and is the metric that we used to compare parameter estimations from differing sampling strategies. The JSD can be calculated using Equation 5 with P and Q representing the two probability distributions and D_{KL} as the KL divergence. A smaller JSD metric indicates that two probability distributions (P and Q) are more similar with a Jensen-Shannon distance of 0 indicating equivalence of the two distributions. The mean JSD was taken over all intervals for the BDSKY and *Skygrowth* models to obtain an overall measure of the level of estimated similarity.

Equation 5:

$$\text{JSD}\ (P||Q\ ) = \sqrt{\frac{1}{2}D_{KL}(P||M\ ) + \frac{1}{2}D_{KL}\ (Q||M\ )} \text{ where } M\ = \frac{1}{2}\ (P+Q)$$

Data availability

Please see https://github.com/rhysinward/Phylodyanmic-Subsampling for code and data used within this study.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

RESULTS

Sampling Schemes

Hong Kong

Hong Kong reacted rapidly upon learning of the emergence of SARS-CoV-2 in Wuhan, Hubei province, China, by declaring a state of emergency on the 25th of January 2020 and by mobilising intensive surveillance schemes in response to initial cases[68]. This appeared to be successful in controlling the first wave of cases. However, due to imported cases from Europe and North America, a second wave of SARS-CoV-2 infections emerged prompting stricter NPIs such as the closure of borders and restrictions on gatherings [68]. Following these measures, the incidence of SARS-CoV-2 rapidly decreased (Figure 1). Hong Kong has a high sampling intensity with 11.6% of confirmed cases sequenced during our study period. Further, Hong Kong has high quality case data with a high testing rate through effective tracing of close contacts, testing of all asymptomatic arriving travellers and all patients with pneumonia[69].

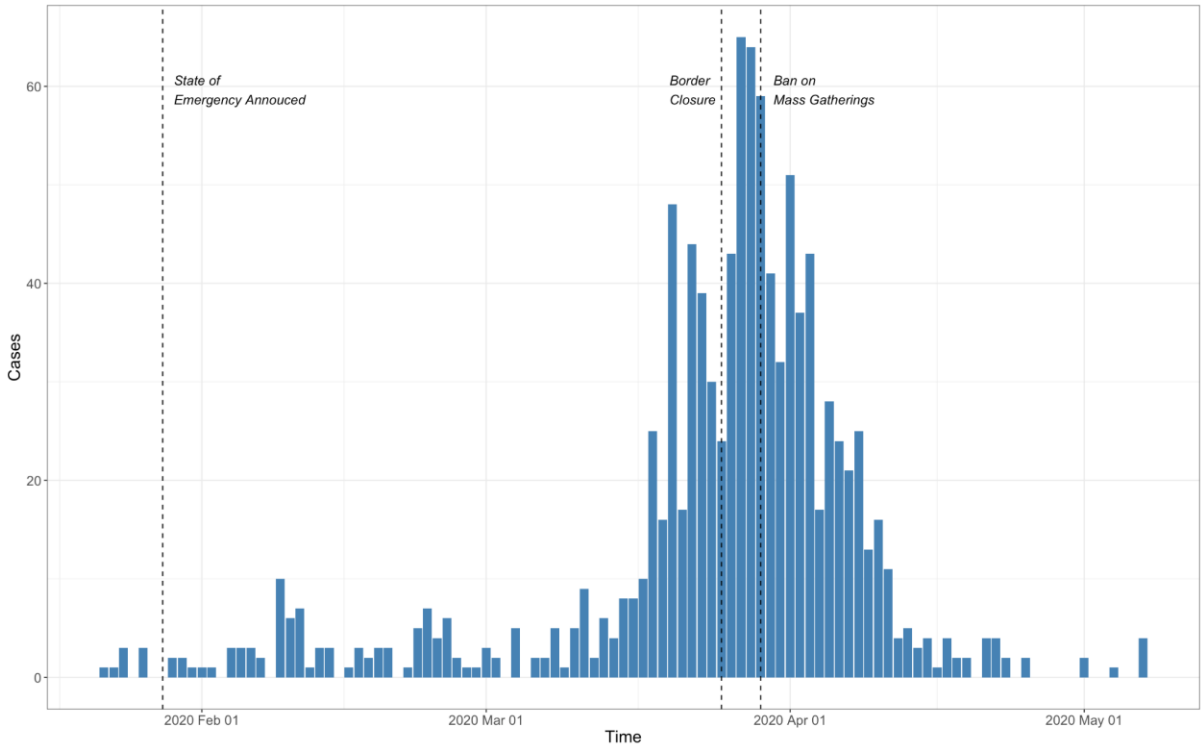


Figure 1. Confirmed SARS-CoV-2 cases from Hong Kong until 7th of May 2020. The dashed lines represent policy change-times[68].

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

The number of cases within Hong Kong for each week was used to inform the sampling schemes used within this study. This resulted in the unsampled scheme having N = 117 sequences, the proportional sampling scheme having N = 54 sequences, the uniform sampling scheme having N = 79 and the reciprocal-proportional sampling scheme having N = 84 sequences (Supplementary Figure 3).

Amazonas

The Amazonas state of Brazil had its first laboratory confirmed case of SARS-CoV-2 in March 2020 in a traveller returning from Europe[70]. After a first large wave of SARS-CoV-2 infections within the state that peaked in early May 2020 (Figure 2), the epidemic waned, cases dropped, remaining stable until mid-December 2020. The number of cases then started growing exponentially, ushering in a second epidemic wave. This second wave peaked in January 2021 (Figure 2) and coincided with the emergence of a new SARS-CoV-2 VOC, designated P.1/Gamma[14].

To combat this second wave, the Government of the Amazonas state suspended all non-essential commercial activities on the 23rd of December 2020 (http://www.pge.am.gov.br/legislacao-covid-19/). However, in response to protests, these restrictions were reversed, and cases continued to climb. On the 12th of January, when local transmission of P.1/Gamma was confirmed in Manaus, capital of Amazonas state[71], NPIs were re-introduced (http://www.pge.am.gov.br/legislacao-covid-19/) which seemed to be successful in reducing the case incidence in the state. However, cases remained comparatively high (Figure 4). Amazonas has a low sampling intensity with 2.4% of suspected P.1/gamma cases sequenced during our study period.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

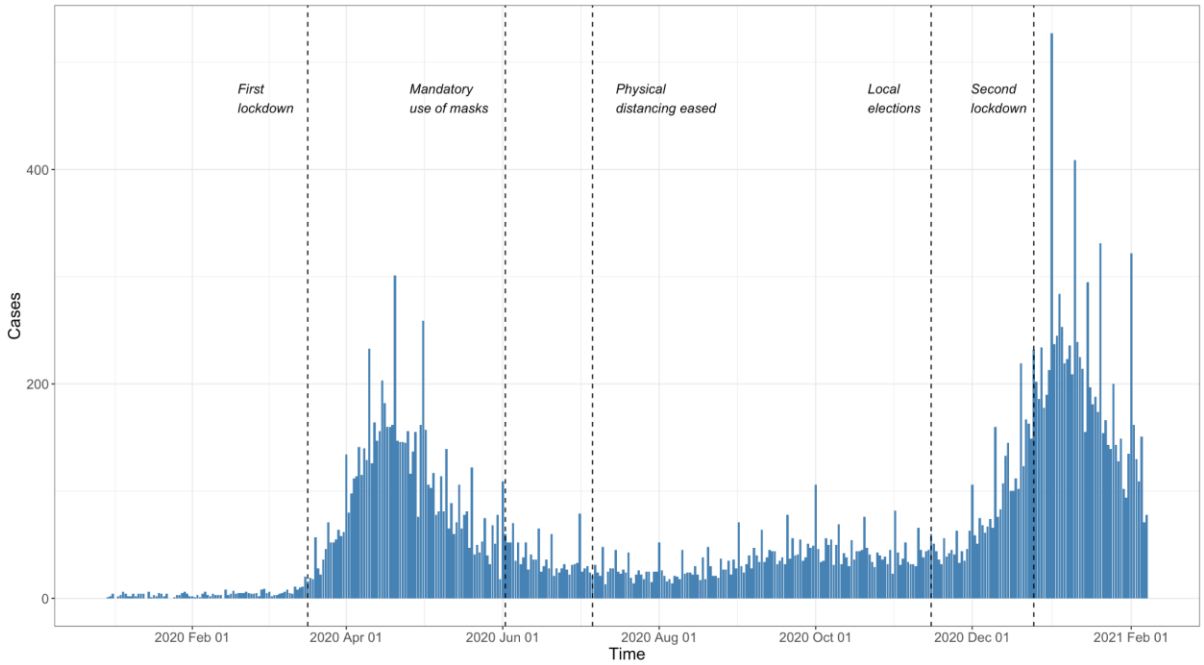


Figure 2. Confirmed SARS-CoV-2 cases from Amazonas state, north Brazil until 7th of February 2021. The dashed lines represent policy change-times [72].

The number of cases within the Amazonas region informed the sampling schemes used within this study. This resulted in the unsampled scheme having N = 196 sequences, the proportional sampling scheme having N = 168 sequences, the uniform sampling scheme having N = 150 and the reciprocal-proportional sampling scheme having N = 67 sequences (Supplementary Figure 4).

Root-to-tip Regression

The correlation (R^2) between genetic divergence and sampling dates for the Hong Kong datasets ranged between 0.36 and 0.52 and between 0.13 and 0.20 for the Amazonas datasets (Supplementary Figure 2). This implies that the Hong Kong datasets have a stronger temporal signal. This is likely due to the Hong Kong datasets have a wider sampling interval (106 days) compared to the Amazonas datasets (69 days). A wider sampling interval can lead to a stronger temporal signal[73]. The gradient (rate) of the regression ranged from 1.16×10^{-3} to 2.09×10^{-3} s/s/y for the Hong Kong datasets and 4.41×10^{-4} to 5.30×10^{-4} s/s/y for the Amazonas datasets.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Estimation of Evolutionary Parameters

The mean substitution rate (measured in units of number of substitutions per site per year, s/s/y) and the time to most common recent ancestor (TMRCA) was estimated in BEAST, for both datasets, and the estimation from all sampling schemes was compared.

Hong Kong

For Hong Kong, the mean substitution rate per site per year ranged from 9.16×10^{-4} to 2.09×10^{-3} with sampling schemes all having overlapped Bayesian credible intervals (BCIs) (Supplementary table 2; Supplementary Figure 5A). This indicates that the sampling scheme did not have a significant impact on the estimation of the clock rate. Moreover, the clock rate is comparable to estimations from the root-to-tip regression and to early estimations of the mean substitution rate per site per year of SARS-CoV-2 (Duchene et al., 2020).

Molecular clock dating of the Hong Kong dataset indicates that the estimated time of the most common recent ancestor was around December 2020 (Figure 3B; Supplementary Table 2). This is a few weeks before the first confirmed case which was reported on the 18th of January 2021. Once again, all sampling strategies have overlapped BCIs and with the range in means differing by around three weeks, a relatively short time scale, suggesting that the sampling scheme does not significantly impact the estimation of the TMRCA.

Brazil

For the P.1 lineage in the Amazonas region, the mean substitution rate ranged from 4.00×10^{-4} to 5.56×10^{-4} s/s/y with all sampling schemes having overlapped BCIs (Figure 3D, Supplementary Table 2; Supplementary Figure 5B). This indicates that sampling strategy does not impact the estimation of the clock rate, supporting findings from the Hong Kong dataset. This also supports estimations from the root-to-tip analysis (Supplementary Figure 2).

Molecular clock dating estimated a TMRCA mean around late October to early November (Figure 3D; Supplementary Table 2). This is around five weeks before the date of the first P.1 case identified in Manaus used in our study. All sampling schemes have overlapping BCI consistent with the conclusion from the Hong Kong data that TMRCA is relatively robust to sampling.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Estimation of Basic Reproduction Number

We found that Hong Kong had a significantly lower R_0 of 2.17 (95% credible interval (CI) = 1.43 - 2.83) when compared to Amazonas which had a R_0 of 3.67 (95% CI = 2.83 – 4.48). All sampling schemes for both datasets were characterised by similar R_0 values (Figure 3) indicating that the estimation of R_0 is robust to changes in sampling scheme.

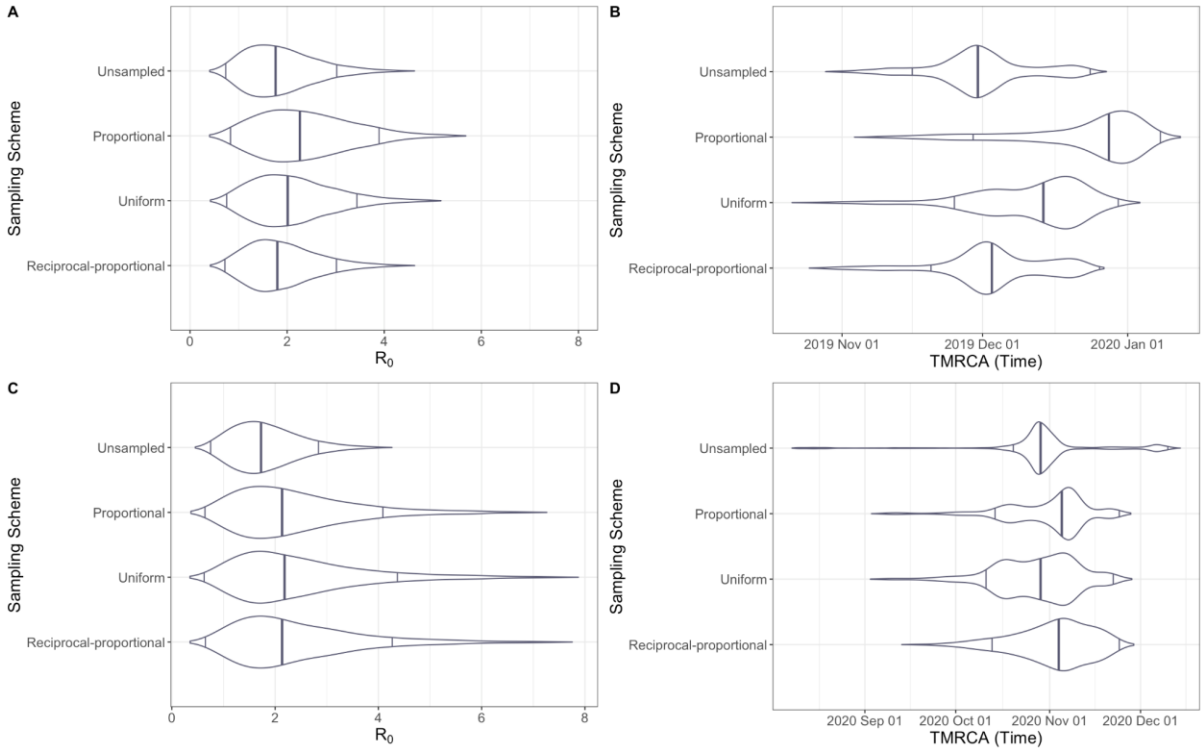


Figure 3. R_0 estimated from BDSKY and TMRCA for Hong Kong and Brazil. Figure 1A and B represent Hong Kong and Figure 1C and D represent the Amazonas.

Time-varying Reproduction number and Growth rate

We estimate R_t and r_t for local SARS-CoV-2 epidemics in Hong Kong and Amazonas, Brazil. Our main results showing these two parameters and JSD metrics are shown in figures 4-8.

Hong Kong

We applied the BDSKY model to estimate the R_t for each dataset subsampled according to the different sampling strategies (Figure 4). We compared these against the R_t from case data, derived from *EpiFilter*. Based on the proportional sampling scheme, which had the lowest

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

JSD (Figure 4E), we initially infer a super-critical R_t value, with a mean around 2, that appears to fall swiftly in response to the state of emergency and the rapid implementation of NPIs. A steady transmission rate subsequently persisted throughout the following weeks around the critical threshold ($R_t = 1$). This period is followed by a sharp increase in R_t, peaking at a mean value of 2.6. This is likely due to imported cases from North America and Europe[68]. This led to a ban on international travel resulting in a sharp decline in R_t (Figure 2). However, this decline lasted around a week with the mean R_t briefly increasing until more stringent NPIs such as the banning of major gatherings were implemented. Following this, the R_t continued its sharp decline falling below the critical threshold, with transmission becoming sub-critical (Figure 4).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

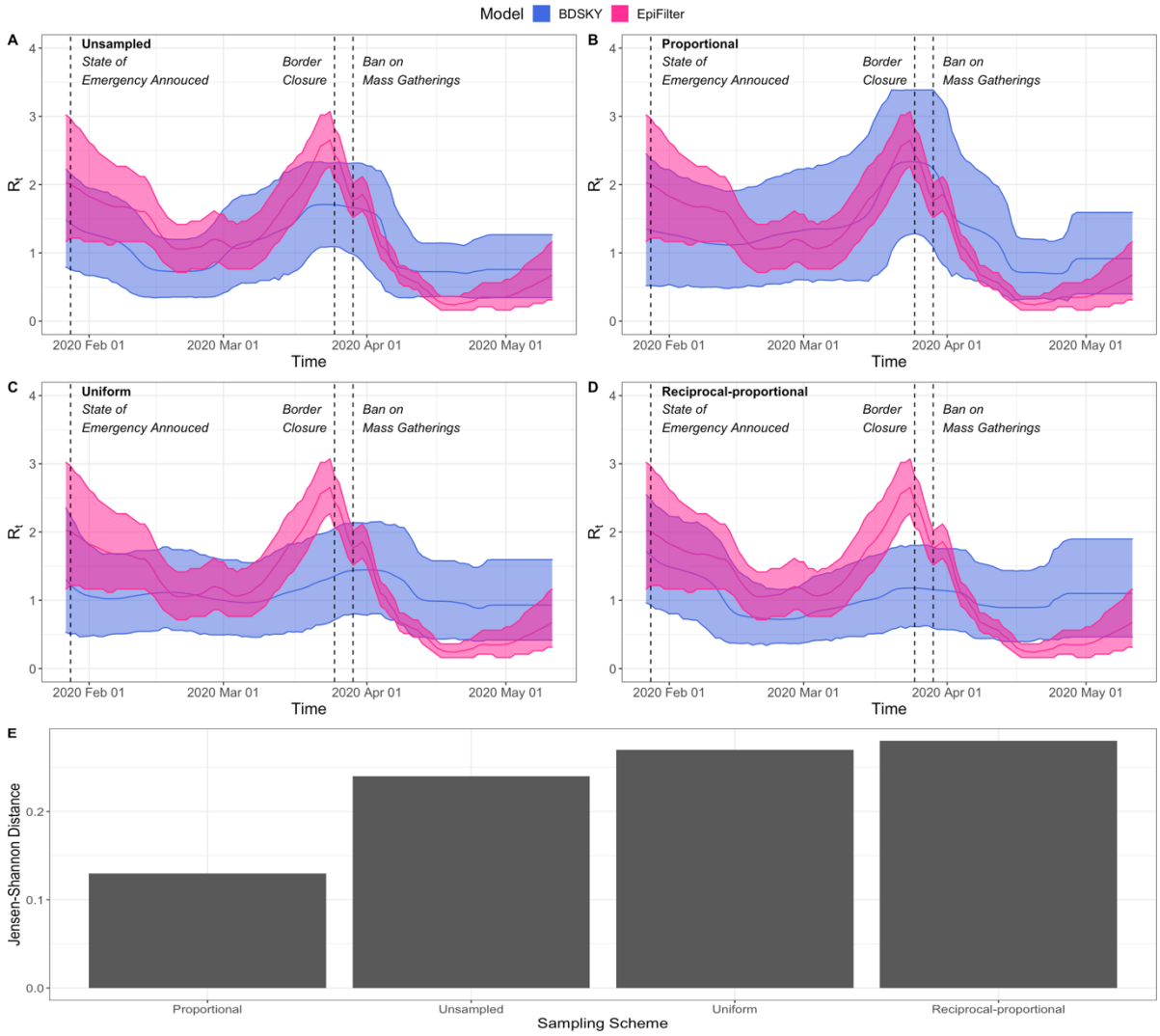


Figure 4: R_t estimated from both the BDSKY and *EpiFilter* methods for Hong Kong. The bold writing represents the sampling scheme used in panels A-D. The light-shaded area represents the 95% Highest Posterior Density Interval with the darker-shaded area presenting where the BDSKY and *EpiFilter* models overlap. The solid line represents the mean R_t estimate with *EpiFilter* in red and BDSKY in blue. The dashed lines represent policy change-times. The Jensen Shannon Distance is ordered from best to worse in panel E.

These results were mirrored in the estimation of $r_{t.}$ (Figure 5) for which the uniform and proportional sampling schemes showed the least divergence (Figure 5E). There was an initial decline in the r_t, which steadied at a value of ~ 0, indicating that epidemic stabilisation had occurred. This stable period is followed by an increase in r_t peaking at around a 0.050 d^{-1} (Figure 5). In response to NPIs, the r_t starts to decrease, falling below 0, indicating a receding epidemic. The rate of this decline peaks at around -0.075 d^{-1} (Figure 5).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

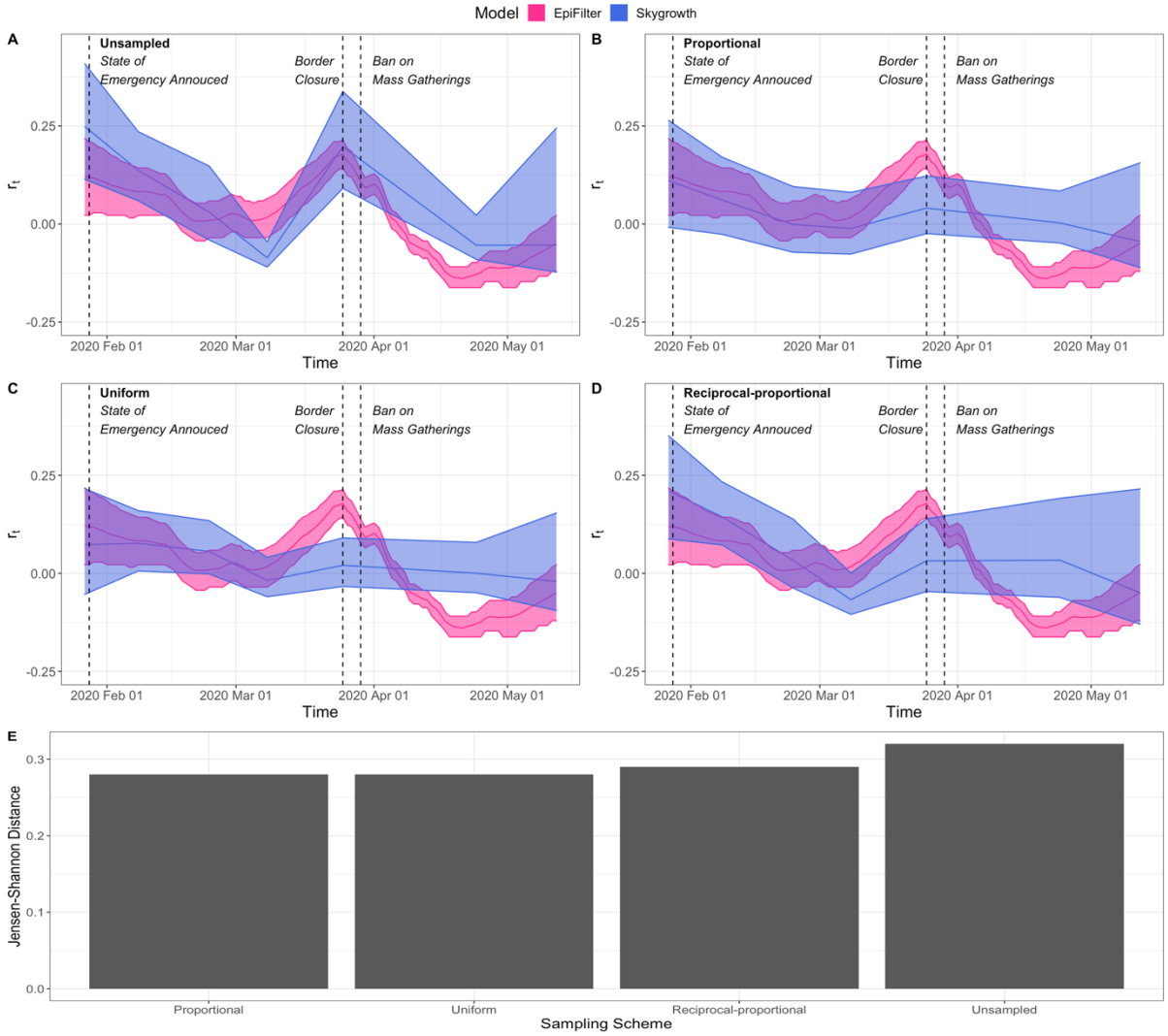


Figure 5: r_t estimated from both the *Skygrowth* and *EpiFilter* methods for Hong Kong. The bold writing represents the sampling scheme used in panels A-D. The light-shaded area represents the 95% Highest Posterior Density Interval with the darker-shaded area presenting where the BDSKY and *Skygrowth* models overlap. The solid line represents the mean r_t estimate with *Skygrowth* in red and BDSKY in blue. The dashed lines represent policy change-times. The JSD metric is ordered from best to worse in panel E.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Brazil

Based on the uniform sampling scheme, which had the lowest JSD (Figure 6E), we initially infer super-critical transmission ($R_t > 1$) with a mean value of 3 (Figure 6). From this point, the R_t declines, although it remains above the critical threshold ($R_t = 1$) for much of the study period. Sub-critical transmission ($R_t < 1$) was only reached after the re-imposition of NPIs. This implies that initial restrictions, such as the suspension of commercial activities, were likely insufficient for suppressing spread. Only after more stringent restrictions were imposed did R_t become sub-critical. However, there is no evidence of a sharp decrease in R_t once restrictions were re-imposed, which may suggest limited effectiveness.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

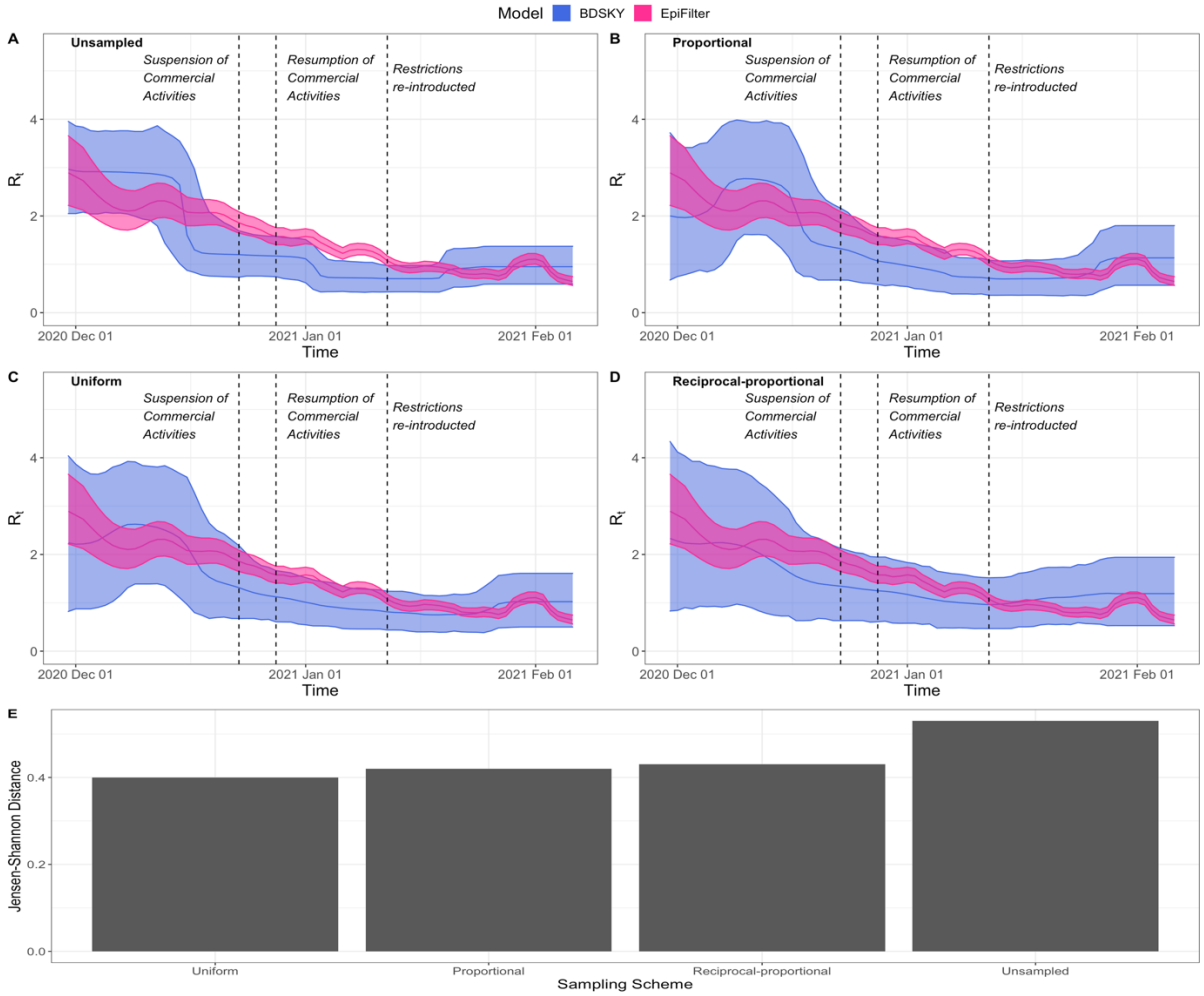


Figure 6: R_t estimated from both the BDSKY and *EpiFilter* methods forAmazonas, Brazil. The bold writing represents the sampling scheme used in panels A-D. The light-shaded area represents the 95% Highest Posterior Density Interval with the darker-shaded area presenting where the BDSKY and *EpiFilter* models overlap. The solid line represents the mean R_t estimate with *EpiFilter* in red and BDSKY in blue. The dashed lines represent policy change-times. The Jensen Shannon Distance is ordered from best to worse in panel E.

Based on the uniform sampling scheme, which had the lowest JSD (Figure 7E) we infer a steady decline in r_t which matches the pattern seen with the R_t value (Figure 7). The initial r_t implied a 0.250 d^{-1}. Subsequently, the r_t falls over the study period. r_t falls below 0 after the re-imposition of NPIs declining at -0.030 d^{-1} by the end of the study period. There is no evidence of any noticeable declines in r_t when interventions were introduced indicating that they might not have significantly impacted the growth rate of P.1/gamma.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

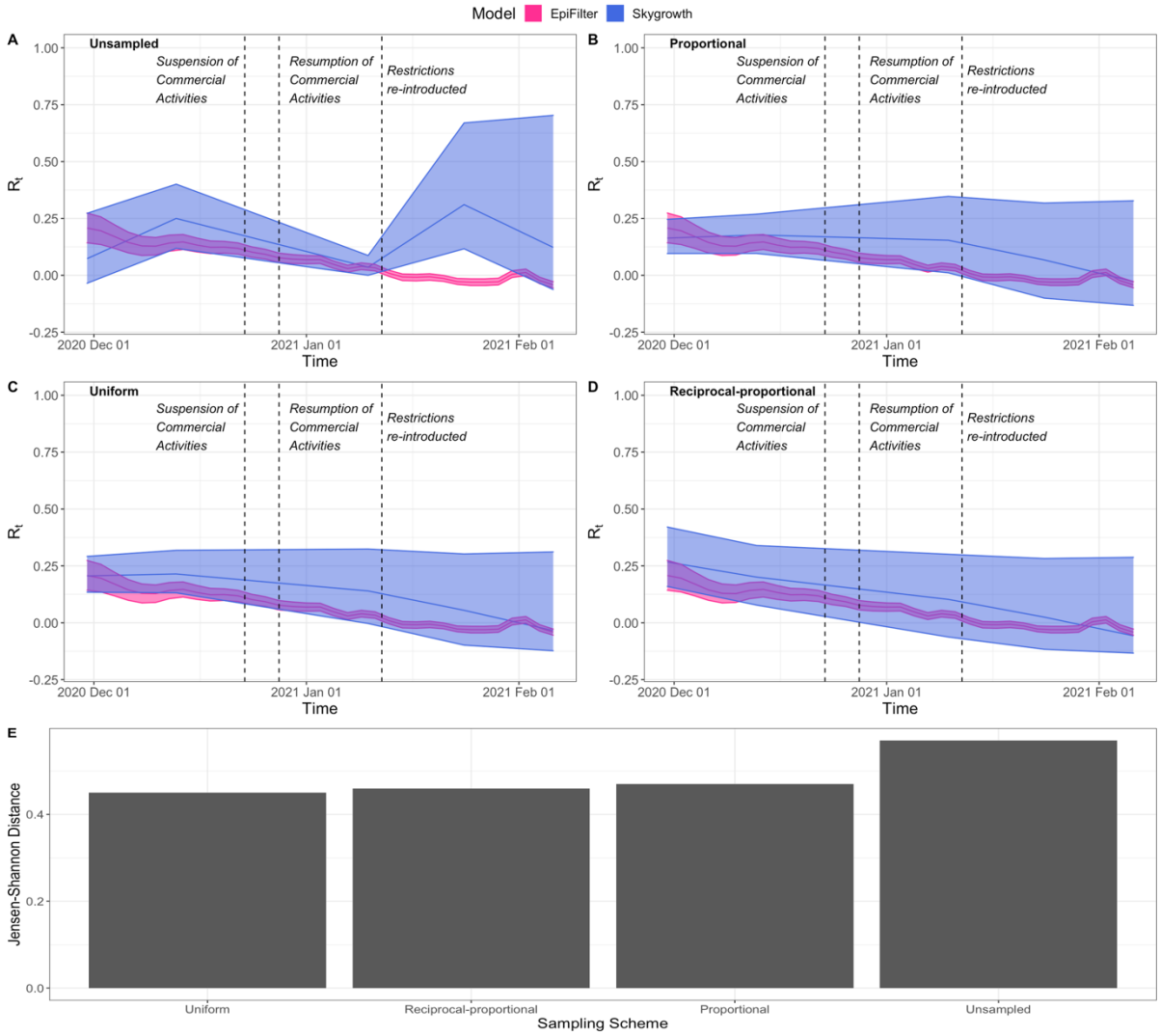


Figure 7: r_t estimated from both the *Skygrowth* and *EpiFilter* methods for Amazonas, Brazil. The bold writing represents the sampling scheme used in panels A-D. The light-shaded area represents the 95% Highest Posterior Density Interval with the darker-shaded area presenting where the BDSKY and *Skygrowth* models overlap. The solid line represents the mean r_t estimate with *Skygrowth* in red and BDSKY in blue. The dashed lines represent policy change-times. The JSD is ordered from best to worse in panel E.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Discussion

In this study, phylodynamic methods have been applied to available SARS-CoV-2 sequences from Hong Kong and the Amazonas region of Brazil to infer their key epidemiological parameters and to compare the impact that various sampling strategies have on the phylodynamic reconstruction of these parameters.

We estimated the basic reproductive number of SARS-CoV-2 in Hong Kong to be 2.17 (95% CI = 1.43-2.83). This supports previous estimates of the initial R_0 in Hong Kong[68,74] which estimates R_0 to be 2.23 (95% CI = 1.47-3.42). For the Amazonas region in Brazil, we estimated the R_0 to be 3.67 (95% CI = 2.83 – 4.48). Whilst the population of Amazonas State may not be fully susceptible to P.1/gamma[14], this should not affect the comparison among sampling schemes. We found that R_0 is robust to changes in sampling schemes (Figure 3A and C).

For the Hong Kong dataset, the proportional sampling scheme was superior to all other sampling schemes in estimating R_t. It successfully predicted the initial super-critical R_t, its decline in response to rapid NPIs, and subsequent increase and decline during the second wave of infections (Figure 4B). This was in comparison to the reciprocal-proportional scheme, which provided the worst (largest) JSD (Figure 4D) and an R_t estimate that was largely insensitive to NPIs. The proportional sampling scheme, alongside the uniform sampling scheme, best estimated r_t (Figure 5B and C). In contrast, for the Amazonas dataset, the uniform sampling scheme best estimated the R_t and r_t (Figure 6C and Figure 7C). It captured both its initial super-critical R_t and high r_t alongside their subsequent decline. Our R_t estimates are consistent with previous estimates of P.1 in Amazonas state[14]. This contrasted with the unsampled data in which the r_t increased at the end of the period (Figure 7A). This highlights that unlike R_0, both R_t and r_t are sensitive to changes in sampling and that even related epidemiological parameters like R_t and r_t may require different sampling strategies to optimise inferences.

Molecular clock dating of the Hong Kong and Amazonas dataset has revealed that the date of origin is relatively robust to changes in sampling schemes. For Hong Kong, SARS-CoV-2 likely emerged in mid-December 2019 around 5 weeks before the first reported case on the 22nd of January 2020[68]. The Amazonas dataset revealed that the date of the common ancestor of the P.1 lineage emerged around late October 2020 to early November, around 5 weeks before the first reported case on the 6th of December[14], with all BCI's overlapping for each

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

sampling strategy. Like the molecular clock dating, we found that the molecular clock rate was robust to changes in sampling strategies in both datasets with all sampling strategies having overlapped BCI's (Supplementary Table 2 and Supplementary Figure 5). For the Hong Kong dataset, its clock rate is comparable to early estimations of the mean substitution rate per site per year of SARS-CoV-2[13]. However, the clock rate estimated for the Brazilian dataset is lower than initial 8.00×10^{-4} s/s/y which is used in investigating SARS-CoV-2[75] and that has been used in previous analyses of P.1[76]. This initial estimation of evolutionary rate was estimated from genomic data taken over a short time span at the beginning of the pandemic introducing a time dependency bias. By using a more appropriate clock rate it can improve tree height and rooting resulting in more robust parameter estimations[77].

Treating sampling times as uninformative has been shown to be inferior to including them as dependent on effective population size and other parameters by several previous studies[30,31,34,78]. Whilst these studies did not consider the estimation of epidemiological parameters, they highlight the potential of systematic biases being introduced into the phylodynamic reconstruction by not using a sampling scheme or by assuming an incorrect model for how sampling schemes introduce information. This was supported by our results as phylodynamic inferences with no sampling strategy applied had the poorest performance for both Hong Kong and the Amazonas region. This implies that sampling has a significant impact on phylodynamic reconstruction, and that exploration of sampling strategies is needed to obtain the most robust parameter estimates.

While our results provide a rigorous underpinning and insight into the dynamics of SARS-CoV-2 and the impact of sampling strategies in the Amazonas region and Hong Kong, there are limitations. The *Skygrowth* and BDSKY models do not explicitly consider imports into their respective regions. This is particularly relevant for Hong Kong as most initial sequences from the region were sequenced from importation events[79] which can introduce error into parameter estimation. However, as the epidemic expanded, more infections were attributable to autochthonous transmission[79], and the risk of error introduced by importation events decreased. Moreover, while sampling strategies can account for temporal variations in genomic sampling fractions there is currently no way to account for non-random sampling approaches in either the BDSKY or *Skygrowth* models[80]. It is unclear how network-based sampling may affect parameter estimates obtained through these models[81] presenting a key

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

challenge in molecular and genetic epidemiology. Spatial heterogeneities were also not explored within this work. This represents the next key step in understanding the impact of sampling as spatial sampling schemes would allow the reconstruction of the dispersal dynamics and estimation of epidemic overdispersion (*k*), a key epidemiological parameter.

Finally, we compared our phylodynamic estimates against epidemiological inferences derived from case data from Hong Kong and Amazonas state, two settings with very different diagnostic capacity. While Hong Kong has high quality case data with a high testing rate through[69], there is a large underreporting of SARS-CoV-2 cases in the Amazonas state[72,82] . Future epidemiological modelling work is needed to compare parameter estimates obtained from case data, death data and excess death data across different settings.

This work has highlighted the impact and importance that applying temporal sampling strategies can have on phylodynamic reconstruction. Whilst more genomic datasets from a variety of countries and regions with different sampling intensities and proportions are needed to create a more generalisable sampling framework and to dissect any potential cofounders, it has been shown that genomic datasets with no sampling strategy applied can introduce significant uncertainty and biases in the estimation of epidemiological parameters. This finding identifies the need for more targeted attempts at performing genomic surveillance and epidemic analyses particularly in resource-poor settings which have a limited genomic capability.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Role of the Funding Sources: N.R.F. acknowledges support from Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z), Bill and Melinda Gates Foundation (INV-034540) and Medical Research Council-Sao Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0) (https://caddecentre.org). K.V.P. acknowledges support from grant reference MR/R015600/1, jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) and from the NIHR Health Protection Research Unit in Behavioural Science and Evaluation at University of Bristol.

CRediT authorship contribution statement: R.P.D.I, K.V.P and N.R.F conceived and designed the study, R.P.D.I wrote and performed the analyses. R.P.D.I wrote the manuscript which was edited and supervised by K.V.P and N.R.F. All authors have contributed to and approved the manuscript for submission.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

References

1. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* **5**, 536–544 (2020).
2. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine* **382**, 727–733 (2020).
3. World Health Organisation. *Public Health Emergency of International Concern (PHEIC)*. (2020).
4. Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet. Infectious diseases* **20**, 669–677 (2020).
5. World Health Organisation. Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update. *https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports* (2022).
6. European Centre for Disease Prevention and Control. *Guidelines for the implementation of non-pharmaceutical interventions against COVID-19 Key messages General considerations on NPI to control COVID-19*. (2020).
7. Anderson, Vegari, C., Baggaley, R., Hollingsworth, T. D. D. & Maddren, R. The Royal Society SET-C Reports. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation [report unpublished]. *The Royal Society* 1–86 (2020).
8. UK Health Security Agency. The R value and growth rate. *https://www.gov.uk/guidance/the-r-value-and-growth-rate* (2022).
9. Parag, K. v, Thompson, R. N. & Donnelly, C. A. Are epidemic growth rates more informative than reproduction numbers? *medRxiv* 2021.04.15.21255565 (2021) doi:10.1101/2021.04.15.21255565.
10. Dushoff, J. & Park, S. W. Speed and strength of an epidemic intervention. *Proceedings of the Royal Society B: Biological Sciences* **288**, 20201556 (2021).
11. World Health Organisation. *Genomic sequencing of SARS-CoV-2 A guide to implementation for maximum impact on public health*. (2021).
12. Jombart, T. *et al.* Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLOS Computational Biology* **10**, e1003457- (2014).
13. Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution* **6**, (2020).
14. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815 LP – 821 (2021).
15. Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S. & Stadler, T. The origin and early spread of SARS-CoV-2 in Europe. *Proceedings of the National Academy of Sciences* **118**, e2012008118 (2021).
16. Romano, C. M. & Melo, F. L. Genomic surveillance of SARS-CoV-2: A race against time. *The Lancet Regional Health - Americas* **0**, 100029 (2021).
17. Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75.e11 (2021).
18. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* (2017) doi:10.1038/nature22040.
19. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406–410 (2017).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

20. Grubaugh, N. D. *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* **546**, 401–405 (2017).
21. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **19**, 409–424 (2021).
22. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology* **5**, 1403–1407 (2020).
23. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **22**, 30494 (2017).
24. Tsang, T. K. *et al.* Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. *The Lancet. Public health* **5**, e289–e296 (2020).
25. de Souza, W. M. *et al.* Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. *Nature Human Behaviour* **4**, 856–865 (2020).
26. Dolan, P. T., Whitfield, Z. J. & Andino, R. Mapping the Evolutionary Potential of RNA Viruses. *Cell Host and Microbe* **23**, 435–446 (2018).
27. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution* **22**, 1185–1192 (2005).
28. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution* **30**, 713–724 (2013).
29. Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences* **110**, 228 LP – 233 (2013).
30. Hall, M. D., Woolhouse, M. E. J. & Rambaut, A. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: A simulation study. *Virus evolution* **2**, vew003–vew003 (2016).
31. Parag, K. v, du Plessis, L. & Pybus, O. G. Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences. *Molecular Biology and Evolution* **37**, 2414–2429 (2020).
32. Stack, J. C., Welch, J. D., Ferrari, M. J., Shapiro, B. U. & Grenfell, B. T. Protocols for sampling viral sequences to study epidemic dynamics. *Journal of the Royal Society, Interface* **7**, 1119–1127 (2010).
33. de Silva, E., Ferguson, N. M. & Fraser, C. Inferring pandemic growth rates from sequence data. *Journal of The Royal Society Interface* **9**, 1797–1808 (2012).
34. Karcher, M. D., Palacios, J. A., Bedford, T., Suchard, M. A. & Minin, V. N. Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference. *PLoS computational biology* **12**, e1004789–e1004789 (2016).
35. Frost, S. D. W. *et al.* Eight challenges in phylodynamic inference. *Epidemics* **10**, 88–92 (2015).
36. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
37. Hidano, A. & Gates, M. C. Assessing biases in phylodynamic inferences in the presence of super-spreaders. *Veterinary Research* **50**, 74 (2019).
38. Gostic, K. M. *et al.* Practical considerations for measuring the effective reproductive number, Rt. *PLOS Computational Biology* **16**, e1008409 (2020).
39. Pullano, G. *et al.* Underdetection of cases of COVID-19 in France threatens epidemic control. *Nature* **590**, 134–139 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

40. The World Bank. Population, total - Hong Kong SAR, China. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=HK (2021).
41. IBGE. Population Projections. https://www.ibge.gov.br/en/statistics/social/population.html (2020).
42. Byrne, A. W. *et al.* Inferred duration of infectious period of SARS-CoV-2: Rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open* **10**, 1–16 (2020).
43. McAloon, C. *et al.* Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. *BMJ Open* **10**, e039652 (2020).
44. Parag, K. v. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. *PLOS Computational Biology* **17**, e1009347 (2021).
45. Fraser, C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. *PLoS ONE* **2**, e758 (2007).
46. Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology* **178**, 1505–1512 (2013).
47. Wallinga, J. & Teunis, P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology* **160**, 509–516 (2004).
48. Wallinga & Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences* **274**, 599–604 (2007).
49. Rai, B., Shukla, A. & Dwivedi, L. K. Estimates of serial interval for COVID-19: A systematic review and meta-analysis. *Clinical epidemiology and global health* **9**, 157–161 (2021).
50. Prete, C. A. *et al.* Serial interval distribution of SARS-CoV-2 infection in Brazil. *Journal of travel medicine* **28**, 1–3 (2021).
51. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3066 (2002).
52. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
53. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).
54. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587–589 (2017).
55. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**, 512–526 (1993).
56. Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic biology* **60**, 685–699 (2011).
57. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution* **2**, vew007–vew007 (2016).
58. World Health Organisation. *Guidance for surveillance of SARS-CoV-2 variants Interim guidance*. (2021).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

59. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus evolution* **4**, vey016–vey016 (2018).
60. Ayres, D. L. *et al.* BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Systematic Biology* **68**, 1052–1061 (2019).
61. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160–174 (1985).
62. Hill, V. & Baele, G. Bayesian Estimation of Past Population Dynamics in BEAST 1.10 Using the Skygrid Coalescent Model. *Molecular Biology and Evolution* **36**, 2620–2628 (2019).
63. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic biology* **67**, 901–904 (2018).
64. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* **15**, e1006650 (2019).
65. Volz, E. M. & Didelot, X. Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. *Systematic Biology* **67**, 719–728 (2018).
66. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **37**, 145–151 (1991).
67. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86 (1951).
68. Cowling, B. J. *et al.* Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. *The Lancet Public Health* **5**, e279–e288 (2020).
69. Wu, P. *et al.* Suppressing COVID-19 Transmission in Hong Kong: An Observational Study of the First Four Months. *SSRN* (2020) doi:10.21203/rs.3.rs-34047/v1.
70. Nascimento, V. A. do *et al.* Genomic and phylogenetic characterisation of an imported case of SARS-CoV-2 in Amazonas State, Brazil. *Memórias do Instituto Oswaldo Cruz* **115**, (2020).
71. Faria, N. R. *et al.* Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological* https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586 (2021).
72. Sabino, E. C. *et al.* Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *Lancet (London, England)* **397**, 452–455 (2021).
73. Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G. Measurably evolving populations. *Trends in Ecology & Evolution* **18**, 481–488 (2003).
74. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* **92**, 214–217 (2020).
75. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine* (2020) doi:10.1038/s41591-020-0820-9.
76. Naveca, F. G. *et al.* COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. *Nature Medicine* **27**, 1230–1238 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

77. Boskova, V., Stadler, T. & Magnus, C. The influence of phylodynamic model specifications on parameter estimates of the Zika virus epidemic. *Virus Evolution* **4**, (2018).
78. Liu, Q. *et al.* Population Genetics of SARS-CoV-2: Disentangling Effects of Sampling Bias and Infection Clusters. *Genomics, Proteomics & Bioinformatics* **18**, 640–647 (2020).
79. Adam, D. C. *et al.* Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Medicine* **26**, 1714–1719 (2020).
80. Vasylyeva, T. I. *et al.* Phylodynamics helps to evaluate the impact of an HIV prevention intervention. *Viruses* **12**, 1–15 (2020).
81. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS computational biology* **9**, e1002947–e1002947 (2013).
82. Buss, L. *et al.* Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. *Science* **371**, 288–292 (2021).

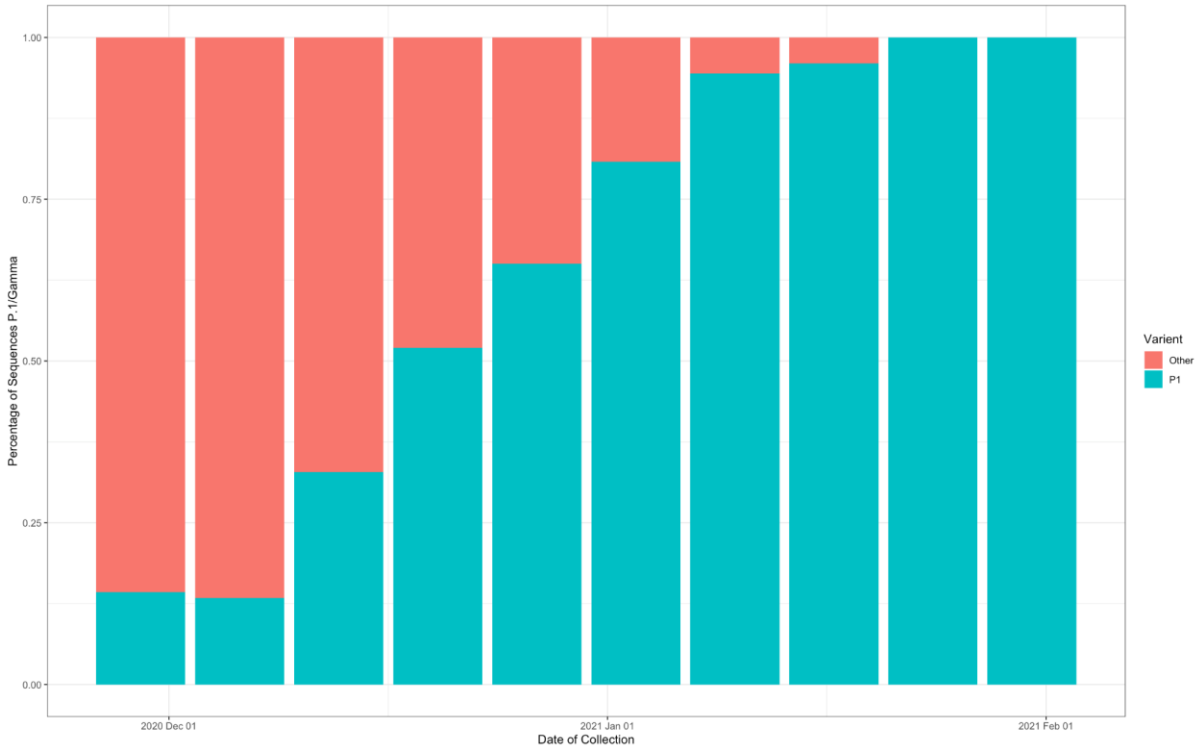
medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Supplementary Figures and Tables

Supplementary Table 1: Key parameters and definitions for SARS-CoV-2

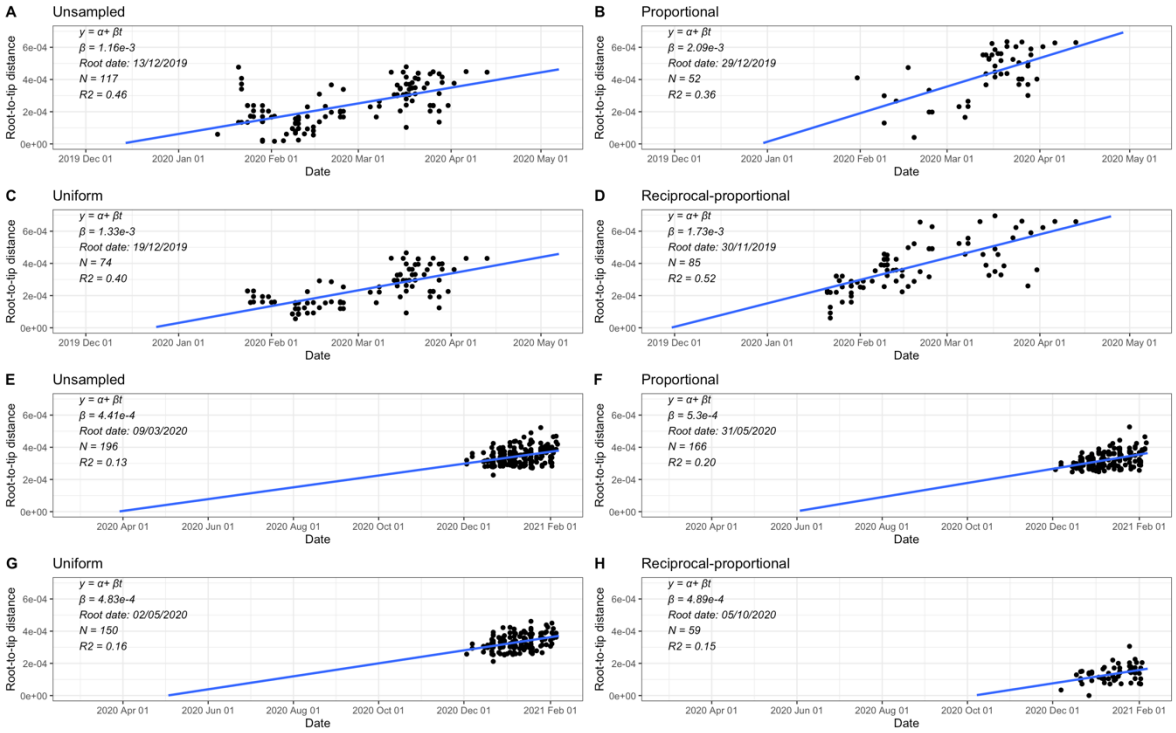
Parameter	Definition
Basic reproduction number (R_0)	Average number of individuals infected by a single infected person in a fully susceptible population
Time-varying or effective reproduction number (R_t)	Average number of secondary infections generated per effective primary case at a certain time point and in the presence of susceptible depletion or interventions
Growth rate (r_t)	Rate of change of the logarithm of the number of new cases per unit of time
Incubation period	Time between infection and symptom onset
Infectious period	Period in which an infectious host can transmit infectious agents to a susceptible individual
Generation interval	Time between infection events in an infector–infectee pair
Date of origin	Date in which viral variant is thought to have emerged
Serial Interval	Time between symptom onsets in an infector–infectee pair

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



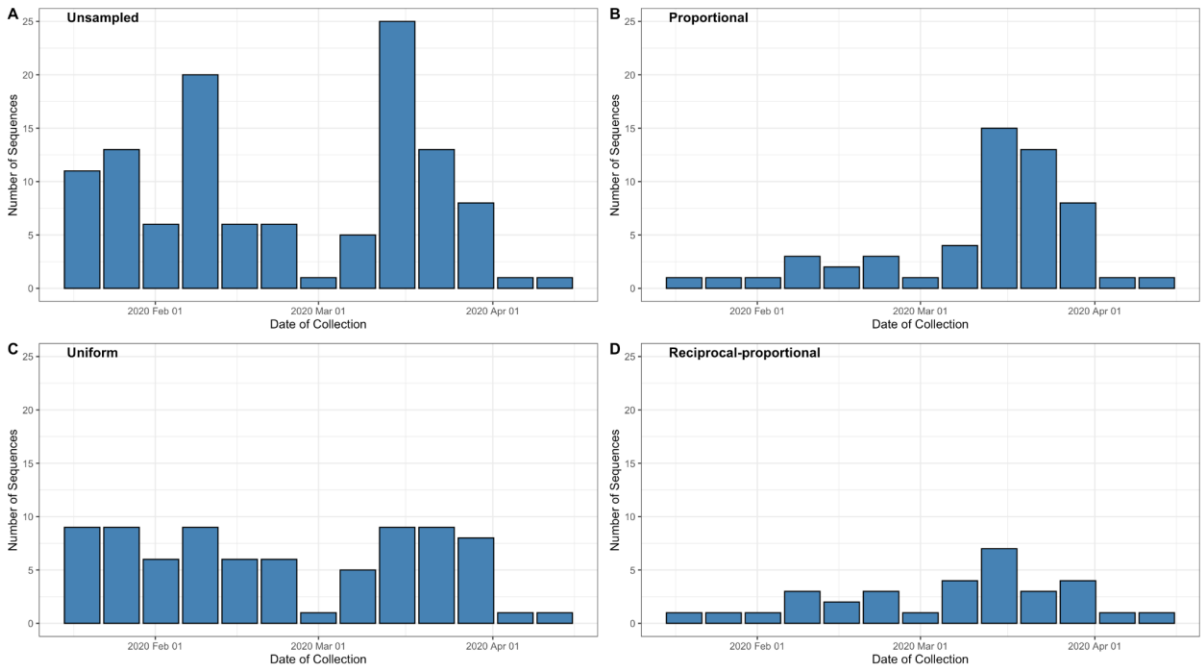
Supplementary Figure 1: The proportion of P.1 sequences compared to non-P.1 sequences found on GISaid (Shu and McCauley, 2017).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

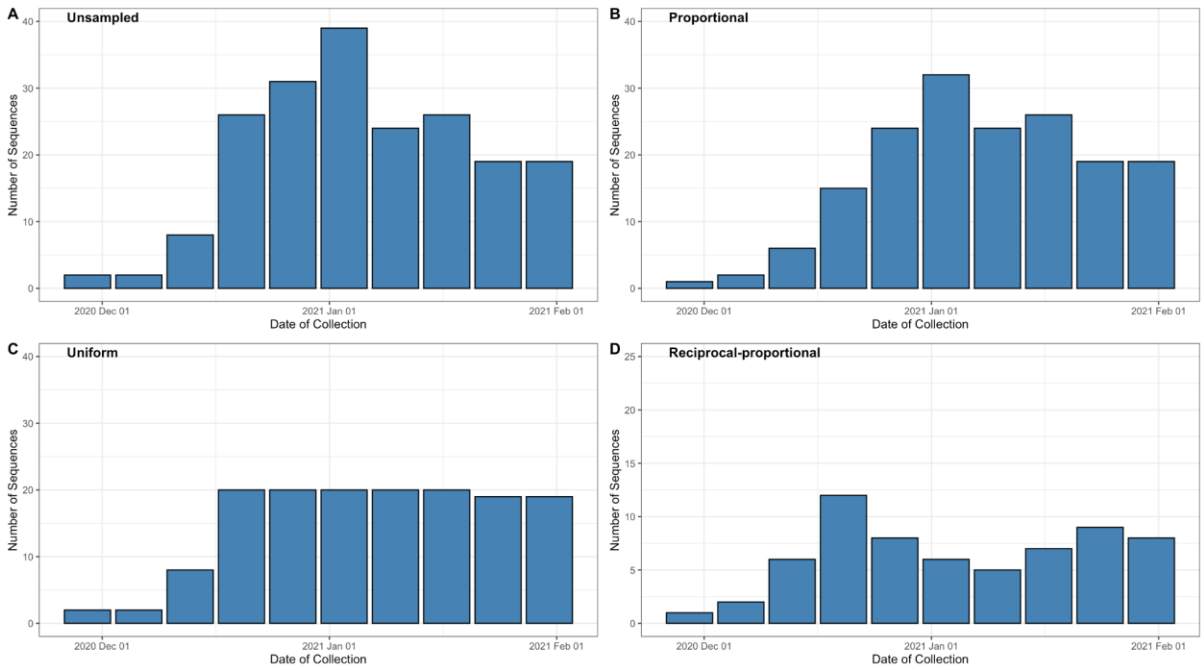


Supplementary Figure 2: Root-to-tip genetic distances to sample collection dates for the SARS-CoV-2 genome datasets used in this study: A-D represents Hong Kong and E-H represent Amazonas State. Plots are based on the maximum likelihood trees rooted by maximising R^2. The linear regression trend lines are shown to data points, corresponding to the genome sequences (represented with black dots).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



Supplementary Figure 3: Number of sequences for each week and sampling scheme for Hong Kong dataset.



Supplementary Figure 4: Number of sequences for each week and sampling scheme for Amazonas dataset.

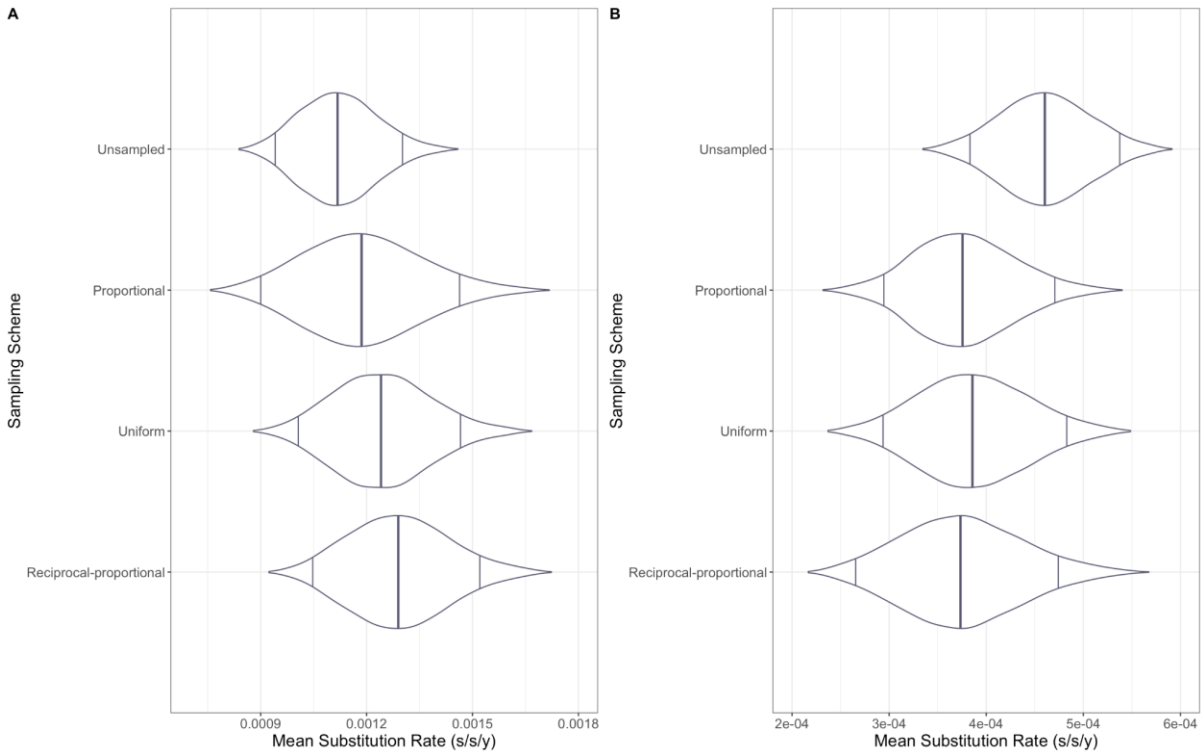
medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Supplementary Table 2: TMRCA and mean substitution rate both with 95% BCI for each sampling strategy for Hong Kong and Amazonas datasets alongside the Jensen-Shannon distance. Full posterior distribution of the TMRCA and substitution rates obtained under the different sampling strategies can be found in Figure 3B and D and Supplementary Figure 5.

Sampling Strategy	**Dataset**	**TMRCA (95% BCI)**	**Mean Substitution Rate (95% BCI, subs/site/year, s/s/y)**
Unsampled	Hong Kong	2nd December 2019 (10th November 2019 – 24th December 2019)	1.12×10^{-3} (9.16×10^{-4} – 1.35×10^{-3})
	Brazil	30th October 2020 (8th October 2020 – 13th December 2020)	4.58×10^{-4} (3.69×10^{-4} – 5.56×10^{-4})
Proportional	Hong Kong	24th December 2019 (21st November 2019 – 11th January 2020)	1.39×10^{-3} (9.28×10^{-4} – 2.48×10^{-3})
	Brazil	30th October 2020 (25th August 2020 – 29th November 2020)	4.60×10^{-4} (3.70×10^{-4} – 5.56×10^{-4})
Uniform	Hong Kong	13th December 2019 (18th November 2019 – 4th January 2020)	1.64×10^{-3} (1.22×10^{-3} – 2.09×10^{-3})

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

	Brazil	27th October 2020 (5th October 2020 – 25th November 2020)	4.60×10^{-4} (3.70×10^{-4} – 5.56×10^{-4})
Reciprocal-proportional	Hong Kong	6th December 2019 (10th November 2019 – 28th December 2019)	1.30×10^{-3} (1.03×10^{-3} – 1.59×10^{-3})
	Brazil	30th October 2020 (27th September 2020 – 25th November 2020)	4.00×10^{-4} (2.56×10^{-4} – 5.55×10^{-4})



Supplementary Figure 5: Mean substitution rate (s/s/y) for Hong Kong and Brazil. Figure 1A represents Hong Kong with Figure 1B representing the Amazonas.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Supplementary Table 3: Accession ID of each Hong Kong sequence for each sampling strategy used within this study

Unsampled	Proportional	Uniform	Reciprocal-proportional
EPI_ISL_ 412028	EPI_ISL_ 414517	EPI_ISL_ 412029	EPI_ISL_ 412028
EPI_ISL_ 412029	EPI_ISL_ 414519	EPI_ISL_ 414517	EPI_ISL_ 412029
EPI_ISL_ 412030	EPI_ISL_ 414527	EPI_ISL_ 414519	EPI_ISL_ 412030
EPI_ISL_ 414517	EPI_ISL_ 418815	EPI_ISL_ 414527	EPI_ISL_ 414517
EPI_ISL_ 414519	EPI_ISL_ 419224	EPI_ISL_ 414569	EPI_ISL_ 414519
EPI_ISL_ 414527	EPI_ISL_ 419229	EPI_ISL_ 414571	EPI_ISL_ 414527
EPI_ISL_ 414528	EPI_ISL_ 419232	EPI_ISL_ 416314	EPI_ISL_ 414528
EPI_ISL_ 414569	EPI_ISL_ 450404	EPI_ISL_ 417064	EPI_ISL_ 414569
EPI_ISL_ 414571	EPI_ISL_ 450405	EPI_ISL_ 417443	EPI_ISL_ 414571
EPI_ISL_ 416314	EPI_ISL_ 450410	EPI_ISL_ 419214	EPI_ISL_ 416314
EPI_ISL_ 417064	EPI_ISL_ 476801	EPI_ISL_ 419215	EPI_ISL_ 417064
EPI_ISL_ 417176	EPI_ISL_ 476802	EPI_ISL_ 419217	EPI_ISL_ 417176
EPI_ISL_ 417178	EPI_ISL_ 476803	EPI_ISL_ 419224	EPI_ISL_ 417178
EPI_ISL_ 417181	EPI_ISL_ 497769	EPI_ISL_ 419225	EPI_ISL_ 417181
EPI_ISL_ 417185	EPI_ISL_ 497773	EPI_ISL_ 419227	EPI_ISL_ 417185
EPI_ISL_ 417187	EPI_ISL_ 497775	EPI_ISL_ 419228	EPI_ISL_ 417187
EPI_ISL_ 417188	EPI_ISL_ 497784	EPI_ISL_ 419229	EPI_ISL_ 417188
EPI_ISL_ 417193	EPI_ISL_ 497786	EPI_ISL_ 419231	EPI_ISL_ 417193
EPI_ISL_ 417197	EPI_ISL_ 497791	EPI_ISL_ 419232	EPI_ISL_ 417197
EPI_ISL_ 417443	EPI_ISL_ 497796	EPI_ISL_ 419245	EPI_ISL_ 417443
EPI_ISL_ 418815	EPI_ISL_ 497799	EPI_ISL_ 419247	EPI_ISL_ 418815
EPI_ISL_ 419214	EPI_ISL_ 497806	EPI_ISL_ 419250	EPI_ISL_ 419214
EPI_ISL_ 419215	EPI_ISL_ 497808	EPI_ISL_ 419252	EPI_ISL_ 419215
EPI_ISL_ 419216	EPI_ISL_ 497810	EPI_ISL_ 434564	EPI_ISL_ 419216
EPI_ISL_ 419217	EPI_ISL_ 497811	EPI_ISL_ 434565	EPI_ISL_ 419217

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 419219	EPI_ISL_ 497818	EPI_ISL_ 434567	EPI_ISL_ 419219
EPI_ISL_ 419221	EPI_ISL_ 497819	EPI_ISL_ 434568	EPI_ISL_ 419221
EPI_ISL_ 419222	EPI_ISL_ 497821	EPI_ISL_ 434569	EPI_ISL_ 419222
EPI_ISL_ 419224	EPI_ISL_ 497823	EPI_ISL_ 434570	EPI_ISL_ 419224
EPI_ISL_ 419225	EPI_ISL_ 497824	EPI_ISL_ 434571	EPI_ISL_ 419225
EPI_ISL_ 419226	EPI_ISL_ 497840	EPI_ISL_ 450405	EPI_ISL_ 419226
EPI_ISL_ 419227	EPI_ISL_ 497845	EPI_ISL_ 450408	EPI_ISL_ 419227
EPI_ISL_ 419228	EPI_ISL_ 497846	EPI_ISL_ 450409	EPI_ISL_ 419228
EPI_ISL_ 419229	EPI_ISL_ 497847	EPI_ISL_ 450410	EPI_ISL_ 419229
EPI_ISL_ 419231	EPI_ISL_ 497850	EPI_ISL_ 450411	EPI_ISL_ 419231
EPI_ISL_ 419232	EPI_ISL_ 497856	EPI_ISL_ 476801	EPI_ISL_ 419232
EPI_ISL_ 419245	EPI_ISL_ 497865	EPI_ISL_ 476802	EPI_ISL_ 419245
EPI_ISL_ 419247	EPI_ISL_ 497870	EPI_ISL_ 476804	EPI_ISL_ 419247
EPI_ISL_ 419250	EPI_ISL_ 516798	EPI_ISL_ 497769	EPI_ISL_ 419250
EPI_ISL_ 419252	EPI_ISL_ 539820	EPI_ISL_ 497771	EPI_ISL_ 419252
EPI_ISL_ 434560	EPI_ISL_ 539850	EPI_ISL_ 497783	EPI_ISL_ 434563
EPI_ISL_ 434563	EPI_ISL_ 539851	EPI_ISL_ 497784	EPI_ISL_ 434564
EPI_ISL_ 434564	EPI_ISL_ 610167	EPI_ISL_ 497791	EPI_ISL_ 434565
EPI_ISL_ 434565	EPI_ISL_ 610168	EPI_ISL_ 497806	EPI_ISL_ 434566
EPI_ISL_ 434566	EPI_ISL_ 610169	EPI_ISL_ 497810	EPI_ISL_ 434567
EPI_ISL_ 434567	EPI_ISL_ 610170	EPI_ISL_ 497811	EPI_ISL_ 434568
EPI_ISL_ 434568	EPI_ISL_ 610171	EPI_ISL_ 497813	EPI_ISL_ 434569
EPI_ISL_ 434569	EPI_ISL_ 610172	EPI_ISL_ 497818	EPI_ISL_ 434570
EPI_ISL_ 434570	EPI_ISL_ 610173	EPI_ISL_ 497821	EPI_ISL_ 434571
EPI_ISL_ 434571	EPI_ISL_ 610174	EPI_ISL_ 497823	EPI_ISL_ 450405
EPI_ISL_ 450404	EPI_ISL_ 610175	EPI_ISL_ 497824	EPI_ISL_ 450408
EPI_ISL_ 450405	EPI_ISL_ 610177	EPI_ISL_ 497826	EPI_ISL_ 450409
EPI_ISL_ 450408		EPI_ISL_ 497827	EPI_ISL_ 450410
EPI_ISL_ 450409		EPI_ISL_ 497831	EPI_ISL_ 450411
EPI_ISL_ 450410		EPI_ISL_ 497832	EPI_ISL_ 450412
EPI_ISL_ 450411		EPI_ISL_ 497846	EPI_ISL_ 476802
EPI_ISL_ 450412		EPI_ISL_ 497847	EPI_ISL_ 476804

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 476801		EPI_ISL_ 497848	EPI_ISL_ 497769
EPI_ISL_ 476802		EPI_ISL_ 497856	EPI_ISL_ 497771
EPI_ISL_ 476803		EPI_ISL_ 497860	EPI_ISL_ 497773
EPI_ISL_ 476804		EPI_ISL_ 497865	EPI_ISL_ 497783
EPI_ISL_ 497769		EPI_ISL_ 539820	EPI_ISL_ 497784
EPI_ISL_ 497771		EPI_ISL_ 539850	EPI_ISL_ 497791
EPI_ISL_ 497773		EPI_ISL_ 539851	EPI_ISL_ 497797
EPI_ISL_ 497775		EPI_ISL_ 610165	EPI_ISL_ 497811
EPI_ISL_ 497783		EPI_ISL_ 610166	EPI_ISL_ 497812
EPI_ISL_ 497784		EPI_ISL_ 610167	EPI_ISL_ 497818
EPI_ISL_ 497786		EPI_ISL_ 610168	EPI_ISL_ 497819
EPI_ISL_ 497791		EPI_ISL_ 610169	EPI_ISL_ 497823
EPI_ISL_ 497796		EPI_ISL_ 610171	EPI_ISL_ 497824
EPI_ISL_ 497797		EPI_ISL_ 610173	EPI_ISL_ 497827
EPI_ISL_ 497798		EPI_ISL_ 610174	EPI_ISL_ 497831
EPI_ISL_ 497799		EPI_ISL_ 610175	EPI_ISL_ 497833
EPI_ISL_ 497806		EPI_ISL_ 610177	EPI_ISL_ 497848
EPI_ISL_ 497808			EPI_ISL_ 497850
EPI_ISL_ 497810			EPI_ISL_ 497856
EPI_ISL_ 497811			EPI_ISL_ 497860
EPI_ISL_ 497812			EPI_ISL_ 497864
EPI_ISL_ 497813			EPI_ISL_ 497865
EPI_ISL_ 497818			EPI_ISL_ 539850
EPI_ISL_ 497819			EPI_ISL_ 539851
EPI_ISL_ 497820			EPI_ISL_ 610165
EPI_ISL_ 497821			EPI_ISL_ 610166
EPI_ISL_ 497823			EPI_ISL_ 610172
EPI_ISL_ 497824			EPI_ISL_ 610177

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 497826			
EPI_ISL_ 497827			
EPI_ISL_ 497831			
EPI_ISL_ 497832			
EPI_ISL_ 497833			
EPI_ISL_ 497840			
EPI_ISL_ 497845			
EPI_ISL_ 497846			
EPI_ISL_ 497847			
EPI_ISL_ 497848			
EPI_ISL_ 497850			
EPI_ISL_ 497856			
EPI_ISL_ 497860			
EPI_ISL_ 497864			
EPI_ISL_ 497865			
EPI_ISL_ 497870			
EPI_ISL_ 516798			
EPI_ISL_ 539820			
EPI_ISL_ 539850			
EPI_ISL_ 539851			
EPI_ISL_ 610165			
EPI_ISL_ 610166			
EPI_ISL_ 610167			
EPI_ISL_ 610168			
EPI_ISL_ 610169			
EPI_ISL_ 610170			
EPI_ISL_ 610171			
EPI_ISL_ 610172			

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 610173			
EPI_ISL_ 610174			
EPI_ISL_ 610175			
EPI_ISL_ 610177			

Supplementary Table 4: Accession ID of each Amazonas State, Brazil sequence for each sampling strategy used within this study

Unsampled	Proportional	Uniform	Reciprocal-proportional
EPI_ISL_ 1034306	EPI_ISL_ 1034304	EPI_ISL_ 1034304	EPI_ISL_ 1034306
EPI_ISL_ 1060876	EPI_ISL_ 1034306	EPI_ISL_ 1034306	EPI_ISL_ 1060913
EPI_ISL_ 1060877	EPI_ISL_ 1060877	EPI_ISL_ 1060877	EPI_ISL_ 1060914
EPI_ISL_ 1060881	EPI_ISL_ 1060881	EPI_ISL_ 1060881	EPI_ISL_ 1068149
EPI_ISL_ 1060888	EPI_ISL_ 1060897	EPI_ISL_ 1060888	EPI_ISL_ 1068150
EPI_ISL_ 1060889	EPI_ISL_ 1060900	EPI_ISL_ 1060889	EPI_ISL_ 1068156
EPI_ISL_ 1060894	EPI_ISL_ 1060902	EPI_ISL_ 1060897	EPI_ISL_ 1068198
EPI_ISL_ 1060897	EPI_ISL_ 1060904	EPI_ISL_ 1060900	EPI_ISL_ 1068258
EPI_ISL_ 1060900	EPI_ISL_ 1060906	EPI_ISL_ 1060912	EPI_ISL_ 1068260
EPI_ISL_ 1060902	EPI_ISL_ 1060912	EPI_ISL_ 1060913	EPI_ISL_ 1068262
EPI_ISL_ 1060904	EPI_ISL_ 1060913	EPI_ISL_ 1060956	EPI_ISL_ 1068263
EPI_ISL_ 1060906	EPI_ISL_ 1060914	EPI_ISL_ 1061026	EPI_ISL_ 1068264
EPI_ISL_ 1060911	EPI_ISL_ 1060918	EPI_ISL_ 1068111	EPI_ISL_ 1068278
EPI_ISL_ 1060912	EPI_ISL_ 1060956	EPI_ISL_ 1068149	EPI_ISL_ 1068286
EPI_ISL_ 1060913	EPI_ISL_ 1061026	EPI_ISL_ 1068150	EPI_ISL_ 1068288
EPI_ISL_ 1060914	EPI_ISL_ 1068110	EPI_ISL_ 1068154	EPI_ISL_ 1166615
EPI_ISL_ 1060918	EPI_ISL_ 1068111	EPI_ISL_ 1068158	EPI_ISL_ 1213190
EPI_ISL_ 1060956	EPI_ISL_ 1068112	EPI_ISL_ 1068160	EPI_ISL_ 1261690
EPI_ISL_ 1061026	EPI_ISL_ 1068114	EPI_ISL_ 1068169	EPI_ISL_ 1261694
EPI_ISL_ 1068110	EPI_ISL_ 1068149	EPI_ISL_ 1068198	EPI_ISL_ 2777236
EPI_ISL_ 1068111	EPI_ISL_ 1068150	EPI_ISL_ 1068222	EPI_ISL_ 2777320
EPI_ISL_ 1068112	EPI_ISL_ 1068151	EPI_ISL_ 1068225	EPI_ISL_ 2777363
EPI_ISL_ 1068114	EPI_ISL_ 1068154	EPI_ISL_ 1068226	EPI_ISL_ 2777375

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 1068149	EPI_ISL_ 1068156	EPI_ISL_ 1068243	EPI_ISL_ 2777376
EPI_ISL_ 1068150	EPI_ISL_ 1068158	EPI_ISL_ 1068248	EPI_ISL_ 2777384
EPI_ISL_ 1068151	EPI_ISL_ 1068160	EPI_ISL_ 1068249	EPI_ISL_ 2777388
EPI_ISL_ 1068154	EPI_ISL_ 1068169	EPI_ISL_ 1068260	EPI_ISL_ 2777397
EPI_ISL_ 1068156	EPI_ISL_ 1068198	EPI_ISL_ 1068261	EPI_ISL_ 2777399
EPI_ISL_ 1068158	EPI_ISL_ 1068221	EPI_ISL_ 1068262	EPI_ISL_ 2777401
EPI_ISL_ 1068160	EPI_ISL_ 1068222	EPI_ISL_ 1068263	EPI_ISL_ 2777403
EPI_ISL_ 1068169	EPI_ISL_ 1068225	EPI_ISL_ 1068264	EPI_ISL_ 2777404
EPI_ISL_ 1068198	EPI_ISL_ 1068248	EPI_ISL_ 1068266	EPI_ISL_ 2777409
EPI_ISL_ 1068221	EPI_ISL_ 1068249	EPI_ISL_ 1068268	EPI_ISL_ 2777410
EPI_ISL_ 1068222	EPI_ISL_ 1068258	EPI_ISL_ 1068269	EPI_ISL_ 2777414
EPI_ISL_ 1068225	EPI_ISL_ 1068260	EPI_ISL_ 1068270	EPI_ISL_ 2777415
EPI_ISL_ 1068226	EPI_ISL_ 1068261	EPI_ISL_ 1068271	EPI_ISL_ 2777465
EPI_ISL_ 1068243	EPI_ISL_ 1068262	EPI_ISL_ 1068272	EPI_ISL_ 2777466
EPI_ISL_ 1068248	EPI_ISL_ 1068263	EPI_ISL_ 1068273	EPI_ISL_ 2777467
EPI_ISL_ 1068249	EPI_ISL_ 1068264	EPI_ISL_ 1068274	EPI_ISL_ 2777469
EPI_ISL_ 1068258	EPI_ISL_ 1068266	EPI_ISL_ 1068279	EPI_ISL_ 2777470
EPI_ISL_ 1068260	EPI_ISL_ 1068268	EPI_ISL_ 1068282	EPI_ISL_ 2777472
EPI_ISL_ 1068261	EPI_ISL_ 1068269	EPI_ISL_ 1068283	EPI_ISL_ 2777473
EPI_ISL_ 1068262	EPI_ISL_ 1068270	EPI_ISL_ 1068284	EPI_ISL_ 2777474
EPI_ISL_ 1068263	EPI_ISL_ 1068271	EPI_ISL_ 1068285	EPI_ISL_ 2777475
EPI_ISL_ 1068264	EPI_ISL_ 1068272	EPI_ISL_ 1068286	EPI_ISL_ 2777482
EPI_ISL_ 1068266	EPI_ISL_ 1068273	EPI_ISL_ 1068287	EPI_ISL_ 2777483
EPI_ISL_ 1068268	EPI_ISL_ 1068274	EPI_ISL_ 1068288	EPI_ISL_ 2777485
EPI_ISL_ 1068269	EPI_ISL_ 1068275	EPI_ISL_ 1068290	EPI_ISL_ 2777503
EPI_ISL_ 1068270	EPI_ISL_ 1068276	EPI_ISL_ 1068291	EPI_ISL_ 2777508
EPI_ISL_ 1068271	EPI_ISL_ 1068278	EPI_ISL_ 1068292	EPI_ISL_ 2777509
EPI_ISL_ 1068272	EPI_ISL_ 1068279	EPI_ISL_ 1166615	EPI_ISL_ 2777516
EPI_ISL_ 1068273	EPI_ISL_ 1068280	EPI_ISL_ 1213190	EPI_ISL_ 2777599
EPI_ISL_ 1068274	EPI_ISL_ 1068281	EPI_ISL_ 1213204	EPI_ISL_ 2777698
EPI_ISL_ 1068275	EPI_ISL_ 1068282	EPI_ISL_ 1261683	EPI_ISL_ 2777986
EPI_ISL_ 1068276	EPI_ISL_ 1068283	EPI_ISL_ 1261685	EPI_ISL_ 2777987
EPI_ISL_ 1068278	EPI_ISL_ 1068284	EPI_ISL_ 1261690	EPI_ISL_ 2777993

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 1068279	EPI_ISL_ 1068285	EPI_ISL_ 1261694	EPI_ISL_ 2777999
EPI_ISL_ 1068280	EPI_ISL_ 1068286	EPI_ISL_ 2777236	EPI_ISL_ 2778002
EPI_ISL_ 1068281	EPI_ISL_ 1068287	EPI_ISL_ 2777248	EPI_ISL_ 2778004
EPI_ISL_ 1068282	EPI_ISL_ 1068288	EPI_ISL_ 2777249	EPI_ISL_ 2778005
EPI_ISL_ 1068283	EPI_ISL_ 1068289	EPI_ISL_ 2777250	EPI_ISL_ 833138
EPI_ISL_ 1068284	EPI_ISL_ 1068290	EPI_ISL_ 2777320	EPI_ISL_ 833140
EPI_ISL_ 1068285	EPI_ISL_ 1068291	EPI_ISL_ 2777363	EPI_ISL_ 906071
EPI_ISL_ 1068286	EPI_ISL_ 1068292	EPI_ISL_ 2777364	EPI_ISL_ 918505
EPI_ISL_ 1068287	EPI_ISL_ 1166615	EPI_ISL_ 2777373	EPI_ISL_ 918506
EPI_ISL_ 1068288	EPI_ISL_ 1213190	EPI_ISL_ 2777374	EPI_ISL_ 918508
EPI_ISL_ 1068289	EPI_ISL_ 1213204	EPI_ISL_ 2777375	EPI_ISL_ 918509
EPI_ISL_ 1068290	EPI_ISL_ 1261683	EPI_ISL_ 2777376	
EPI_ISL_ 1068291	EPI_ISL_ 1261685	EPI_ISL_ 2777377	
EPI_ISL_ 1068292	EPI_ISL_ 1261690	EPI_ISL_ 2777378	
EPI_ISL_ 1166615	EPI_ISL_ 1261694	EPI_ISL_ 2777380	
EPI_ISL_ 1213190	EPI_ISL_ 2777236	EPI_ISL_ 2777383	
EPI_ISL_ 1213204	EPI_ISL_ 2777238	EPI_ISL_ 2777384	
EPI_ISL_ 1261683	EPI_ISL_ 2777248	EPI_ISL_ 2777385	
EPI_ISL_ 1261685	EPI_ISL_ 2777249	EPI_ISL_ 2777388	
EPI_ISL_ 1261690	EPI_ISL_ 2777250	EPI_ISL_ 2777397	
EPI_ISL_ 1261694	EPI_ISL_ 2777251	EPI_ISL_ 2777398	
EPI_ISL_ 2777236	EPI_ISL_ 2777320	EPI_ISL_ 2777399	
EPI_ISL_ 2777238	EPI_ISL_ 2777363	EPI_ISL_ 2777400	
EPI_ISL_ 2777248	EPI_ISL_ 2777364	EPI_ISL_ 2777401	
EPI_ISL_ 2777249	EPI_ISL_ 2777373	EPI_ISL_ 2777402	
EPI_ISL_ 2777250	EPI_ISL_ 2777374	EPI_ISL_ 2777403	
EPI_ISL_ 2777251	EPI_ISL_ 2777375	EPI_ISL_ 2777404	
EPI_ISL_ 2777320	EPI_ISL_ 2777376	EPI_ISL_ 2777405	
EPI_ISL_ 2777363	EPI_ISL_ 2777377	EPI_ISL_ 2777406	
EPI_ISL_ 2777364	EPI_ISL_ 2777378	EPI_ISL_ 2777407	
EPI_ISL_ 2777373	EPI_ISL_ 2777380	EPI_ISL_ 2777408	
EPI_ISL_ 2777374	EPI_ISL_ 2777382	EPI_ISL_ 2777410	
EPI_ISL_ 2777375	EPI_ISL_ 2777383	EPI_ISL_ 2777412	

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 2777376	EPI_ISL_ 2777384	EPI_ISL_ 2777413	
EPI_ISL_ 2777377	EPI_ISL_ 2777385	EPI_ISL_ 2777414	
EPI_ISL_ 2777378	EPI_ISL_ 2777388	EPI_ISL_ 2777415	
EPI_ISL_ 2777380	EPI_ISL_ 2777397	EPI_ISL_ 2777417	
EPI_ISL_ 2777382	EPI_ISL_ 2777398	EPI_ISL_ 2777418	
EPI_ISL_ 2777383	EPI_ISL_ 2777399	EPI_ISL_ 2777419	
EPI_ISL_ 2777384	EPI_ISL_ 2777400	EPI_ISL_ 2777454	
EPI_ISL_ 2777385	EPI_ISL_ 2777401	EPI_ISL_ 2777461	
EPI_ISL_ 2777388	EPI_ISL_ 2777402	EPI_ISL_ 2777462	
EPI_ISL_ 2777397	EPI_ISL_ 2777403	EPI_ISL_ 2777465	
EPI_ISL_ 2777398	EPI_ISL_ 2777404	EPI_ISL_ 2777466	
EPI_ISL_ 2777399	EPI_ISL_ 2777405	EPI_ISL_ 2777467	
EPI_ISL_ 2777400	EPI_ISL_ 2777406	EPI_ISL_ 2777469	
EPI_ISL_ 2777401	EPI_ISL_ 2777407	EPI_ISL_ 2777470	
EPI_ISL_ 2777402	EPI_ISL_ 2777408	EPI_ISL_ 2777472	
EPI_ISL_ 2777403	EPI_ISL_ 2777409	EPI_ISL_ 2777473	
EPI_ISL_ 2777404	EPI_ISL_ 2777410	EPI_ISL_ 2777474	
EPI_ISL_ 2777405	EPI_ISL_ 2777412	EPI_ISL_ 2777475	
EPI_ISL_ 2777406	EPI_ISL_ 2777413	EPI_ISL_ 2777477	
EPI_ISL_ 2777407	EPI_ISL_ 2777414	EPI_ISL_ 2777478	
EPI_ISL_ 2777408	EPI_ISL_ 2777415	EPI_ISL_ 2777479	
EPI_ISL_ 2777409	EPI_ISL_ 2777416	EPI_ISL_ 2777481	
EPI_ISL_ 2777410	EPI_ISL_ 2777417	EPI_ISL_ 2777482	
EPI_ISL_ 2777412	EPI_ISL_ 2777418	EPI_ISL_ 2777483	
EPI_ISL_ 2777413	EPI_ISL_ 2777419	EPI_ISL_ 2777485	
EPI_ISL_ 2777414	EPI_ISL_ 2777420	EPI_ISL_ 2777495	
EPI_ISL_ 2777415	EPI_ISL_ 2777454	EPI_ISL_ 2777498	
EPI_ISL_ 2777416	EPI_ISL_ 2777460	EPI_ISL_ 2777503	
EPI_ISL_ 2777417	EPI_ISL_ 2777461	EPI_ISL_ 2777507	
EPI_ISL_ 2777418	EPI_ISL_ 2777462	EPI_ISL_ 2777508	
EPI_ISL_ 2777419	EPI_ISL_ 2777464	EPI_ISL_ 2777539	
EPI_ISL_ 2777420	EPI_ISL_ 2777466	EPI_ISL_ 2777599	
EPI_ISL_ 2777454	EPI_ISL_ 2777467	EPI_ISL_ 2777698	

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 2777460	EPI_ISL_ 2777468	EPI_ISL_ 2777700	
EPI_ISL_ 2777461	EPI_ISL_ 2777469	EPI_ISL_ 2777701	
EPI_ISL_ 2777462	EPI_ISL_ 2777470	EPI_ISL_ 2777740	
EPI_ISL_ 2777464	EPI_ISL_ 2777472	EPI_ISL_ 2777986	
EPI_ISL_ 2777465	EPI_ISL_ 2777473	EPI_ISL_ 2777987	
EPI_ISL_ 2777466	EPI_ISL_ 2777475	EPI_ISL_ 2777993	
EPI_ISL_ 2777467	EPI_ISL_ 2777477	EPI_ISL_ 2777995	
EPI_ISL_ 2777468	EPI_ISL_ 2777478	EPI_ISL_ 2777996	
EPI_ISL_ 2777469	EPI_ISL_ 2777481	EPI_ISL_ 2777997	
EPI_ISL_ 2777470	EPI_ISL_ 2777482	EPI_ISL_ 2777998	
EPI_ISL_ 2777471	EPI_ISL_ 2777495	EPI_ISL_ 2777999	
EPI_ISL_ 2777472	EPI_ISL_ 2777498	EPI_ISL_ 2778000	
EPI_ISL_ 2777473	EPI_ISL_ 2777499	EPI_ISL_ 2778002	
EPI_ISL_ 2777474	EPI_ISL_ 2777503	EPI_ISL_ 2778005	
EPI_ISL_ 2777475	EPI_ISL_ 2777508	EPI_ISL_ 811149	
EPI_ISL_ 2777477	EPI_ISL_ 2777516	EPI_ISL_ 833136	
EPI_ISL_ 2777478	EPI_ISL_ 2777539	EPI_ISL_ 833139	
EPI_ISL_ 2777479	EPI_ISL_ 2777698	EPI_ISL_ 833140	
EPI_ISL_ 2777481	EPI_ISL_ 2777701	EPI_ISL_ 906071	
EPI_ISL_ 2777482	EPI_ISL_ 2777740	EPI_ISL_ 906077	
EPI_ISL_ 2777483	EPI_ISL_ 2777986	EPI_ISL_ 906081	
EPI_ISL_ 2777484	EPI_ISL_ 2777987	EPI_ISL_ 918500	
EPI_ISL_ 2777485	EPI_ISL_ 2777995	EPI_ISL_ 918502	
EPI_ISL_ 2777495	EPI_ISL_ 2777996	EPI_ISL_ 918503	
EPI_ISL_ 2777498	EPI_ISL_ 2777997	EPI_ISL_ 918506	
EPI_ISL_ 2777499	EPI_ISL_ 2777998	EPI_ISL_ 918508	
EPI_ISL_ 2777503	EPI_ISL_ 2778002	EPI_ISL_ 918509	
EPI_ISL_ 2777507	EPI_ISL_ 2778005	EPI_ISL_ 918511	
EPI_ISL_ 2777508	EPI_ISL_ 811149		
EPI_ISL_ 2777509	EPI_ISL_ 833136		
EPI_ISL_ 2777516	EPI_ISL_ 833138		
EPI_ISL_ 2777539	EPI_ISL_ 833139		
EPI_ISL_ 2777599	EPI_ISL_ 833140		

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 2777698	EPI_ISL_ 906071		
EPI_ISL_ 2777700	EPI_ISL_ 906080		
EPI_ISL_ 2777701	EPI_ISL_ 906081		
EPI_ISL_ 2777740	EPI_ISL_ 918500		
EPI_ISL_ 2777986	EPI_ISL_ 918501		
EPI_ISL_ 2777987	EPI_ISL_ 918502		
EPI_ISL_ 2777993	EPI_ISL_ 918503		
EPI_ISL_ 2777995	EPI_ISL_ 918505		
EPI_ISL_ 2777996	EPI_ISL_ 918506		
EPI_ISL_ 2777997	EPI_ISL_ 918507		
EPI_ISL_ 2777998	EPI_ISL_ 918508		
EPI_ISL_ 2777999	EPI_ISL_ 918510		
EPI_ISL_ 2778000	EPI_ISL_ 918511		
EPI_ISL_ 2778002			
EPI_ISL_ 2778004			
EPI_ISL_ 2778005			
EPI_ISL_ 811149			
EPI_ISL_ 833136			
EPI_ISL_ 833138			
EPI_ISL_ 833139			
EPI_ISL_ 833140			
EPI_ISL_ 906071			
EPI_ISL_ 906075			
EPI_ISL_ 906076			
EPI_ISL_ 906077			
EPI_ISL_ 906080			
EPI_ISL_ 906081			
EPI_ISL_ 918499			
EPI_ISL_ 918500			
EPI_ISL_ 918501			
EPI_ISL_ 918502			
EPI_ISL_ 918503			
EPI_ISL_ 918504			

medRxiv preprint doi: https://doi.org/10.1101/2022.02.04.22270165; this version posted March 16, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

EPI_ISL_ 918505			
EPI_ISL_ 918506			
EPI_ISL_ 918507			
EPI_ISL_ 918508			
EPI_ISL_ 918509			
EPI_ISL_ 918510			
EPI_ISL_ 918511			