University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Duchene, S;Featherstone, L;Haritopoulou-Sinanidou, M;Rambaut, A;Lemey, P;Baele, G

Title:
Temporal signal and the phylodynamic threshold of SARS-CoV-2

Date:
2020-07-01

Citation:
Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P. & Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evolution, 6 (2), https://doi.org/10.1093/ve/veaa061.

Persistent Link:
https://hdl.handle.net/11343/252940

License:
CC BY

VIRUS EVOLUTION

*Virus Evolution*, 2020, 6(2): veaa061

doi: 10.1093/ve/veaa061
Rapid Communication

# Temporal signal and the phylodynamic threshold of SARS-CoV-2

Sebastian Duchene,[1,*,†] Leo Featherstone,[1] Melina Haritopoulou-Sinanidou,[1] Andrew Rambaut,[2] Philippe Lemey,[3,‡] and Guy Baele[3]

[1]Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia, 3000, [2]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK and [3]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium

*Corresponding author: E-mail: sduchene@unimelb.edu.au

[†]https://orcid.org/0000-0002-8878-1758

[‡]https://orcid.org/0000-0003-2826-5353

## Abstract

The ongoing SARS-CoV-2 outbreak marks the first time that large amounts of genome sequence data have been generated and made publicly available in near real time. Early analyses of these data revealed low sequence variation, a finding that is consistent with a recently emerging outbreak, but which raises the question of whether such data are sufficiently informative for phylogenetic inferences of evolutionary rates and time scales. The phylodynamic threshold is a key concept that refers to the point in time at which sufficient molecular evolutionary change has accumulated in available genome samples to obtain robust phylodynamic estimates. For example, before the phylodynamic threshold is reached, genomic variation is so low that even large amounts of genome sequences may be insufficient to estimate the virus's evolutionary rate and the time scale of an outbreak. We collected genome sequences of SARS-CoV-2 from public databases at eight different points in time and conducted a range of tests of temporal signal to determine if and when the phylodynamic threshold was reached, and the range of inferences that could be reliably drawn from these data. Our results indicate that by 2 February 2020, estimates of evolutionary rates and time scales had become possible. Analyses of subsequent data sets, that included between 47 and 122 genomes, converged at an evolutionary rate of about $1.1 \times 10^{-3}$ subs/site/year and a time of origin of around late November 2019. Our study provides guidelines to assess the phylodynamic threshold and demonstrates that establishing this threshold constitutes a fundamental step for understanding the power and limitations of early data in outbreak genome surveillance.

**Key words:** 2019 novel coronavirus (SARS-CoV-2); severe acute respiratory syndrome corona virus 2; molecular clock; temporal signal; phylodynamic threshold; phylogenetics.

## 1. Introduction

Pathogen genome sequence data are increasingly recognised as a key asset in outbreak investigations. Phylodynamic analyses of these data can be used to infer the time and location of origin of an outbreak, the viral evolutionary rate, epidemiological dynamics, and demographic patterns (du Plessis and Stadler 2015; Baele, Lemey, and Suchard 2017). These inferences, however, rely on the genome data being sufficiently informative.

© The Author(s) 2020. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

The ongoing novel coronavirus outbreak (SARS-CoV-2) marks the first time that genome sequence data have been generated and shared publicly as soon as the virus started spreading. The time of origin of SARS-CoV-2 is a pressing question at early stages of the outbreak because it impacts our understanding of its spread and emergence. In practice, the sampling times of genomes can be used to calibrate the molecular clock and infer the viral evolutionary rate and the time scale of the outbreak (Korber et al. 2000). The underlying assumption is that molecular evolution occurs at a predictable rate over time and that the sampling window is sufficiently wide as to capture a measurable amount of evolutionary change in the sampled genomes. Under the condition that the sampling window is sufficiently wide and the evolutionary rate sufficiently high, and genome sequences long enough, the data can be treated as having been obtained from a measurably evolving population (Drummond et al. 2003; Biek et al. 2015). If this is not the case, the data are considered to have no temporal signal and any estimates from the molecular clock are therefore spurious (Duchêne et al. 2015; Murray et al. 2016).

The term 'phylodynamic threshold' pertains to the question of whether a virus has had sufficient time to evolve since its origin so as to warrant tip-dating calibration, under the assumption that genome data from early stages of the outbreak are available (Hedge, Lycett, and Rambaut 2013). Therefore, applying statistical tests of temporal signal to genome data as they are collected can reveal when the phylodynamic threshold is reached. Such analyses are essential to determine the limitations of genome data and the range of inferences that can be reliably drawn from them over time.

Phylodynamic analyses often involve estimating population dynamic parameters. Temporal signal in itself does not guarantee that such parameters can be properly estimated, because they often depend on aspects of the sampling process. For example, biased sampling strategies can mislead inference under phylogeographic and epidemiological models (Duchêne et al. 2019; Kalkauskas et al. 2020). Therefore, our definition of the phylodynamic threshold, as we use it here, pertains to temporal signal as a key prerequisite to perform phylodynamic inference.

Root-to-tip regression is typically used as an informal assessment of temporal signal (Rambaut et al. 2016). While not a statistical test, it is however a valuable visual tool of clocklike behaviour and of outlier detection (e.g. due to mislabelling, contamination, or sequencing errors). Root-to-tip regression consists of estimating an unrooted phylogenetic tree with branch lengths in units of substitutions per site and conducting a regression of the distance from the root to each of the tips as a function of their sampling times (Gojobori, Moriyama, and Kimura 1990; Drummond et al. 2003). Under clocklike evolution and with a wide sampling window, the slope corresponds to a crude estimate of the evolutionary rate, the intercept with the time axis represents the time of origin, and the coefficient of determination, $R^2$, may reflect the degree of clocklike behaviour.

Formal approaches to assess temporal signal include date-randomisation tests and Bayesian evaluation of temporal signal (BETS) (Duchêne et al. 2015, 2020; Murray et al. 2016). Date randomisation tests consist of repeating the analysis several times with permuted sampling times to generate a 'null' distribution of evolutionary rate estimates. The data are considered to have temporal signal if the estimate obtained with the correct sampling times does not overlap with those of the randomisations. In contrast, BETS consists of comparing the statistical fit of models that include the correct sampling times, no sampling times, or permuted sampling times. The premise of BETS is that if the data have temporal signal, using the correct sampling times should have the highest statistical fit (Duchêne et al. 2020). For example, if the sampling window over which the genome data have been collected is very short, such that the data have no temporal signal, then the sampling times are not meaningful and a model incorporating the correct sampling times may not have an improved statistical fit over a model that ignores differences in sampling times. In contrast, if the sampling window is wide enough as to capture many substitutions, using the correct sampling times is expected to result in higher model fit than using permuted sampling times or no sampling times. In a Bayesian context, model fit is determined through the marginal likelihood, and a model is preferred over another according to their ratio of marginal likelihoods, known as the Bayes factor (Kass and Raftery 1995). Marginal likelihoods are typically reported on a logarithmic scale, where a log Bayes factors of at least 1 is considered as positive evidence in favour of a model.

Using BETS and root-to-tip regression, we aimed to find the point in time when the phylodynamic threshold for SARS-CoV-2 was reached. This hence constitutes a series of analyses as the pandemic was starting to unfold and when large-scale sequencing efforts still had to be initiated across the affected countries. We extended this inquiry to explore how other epidemiological parameters, such as the population growth rate, became reliably estimable. Our approach consisted of applying these techniques to a growing sample of genomes across eight time points.

## 2. Results

We collected SARS-CoV-2 genome data from the Global Initiative on Sharing All Influenza Data (GISAID) and from GenBank at eight time points from 23 January to 24 February 2020 (Table 1). The data ranged from 31 to 63 days since the first genome was collected (23 December 2019). Thus, each time point represents a 'snapshot' of the genome data available to that date. Our data only included genomic sequences from human samples, with sequence lengths of at least 28,000 nucleotides and, with high coverage as determined in GISAID (see Supplementary Table S1 for accession numbers). To minimise the impact of potential sequencing errors in our alignments, we deleted obvious errors upon visual inspection and compared our phylogenetic trees to those obtained by other groups (virological.org) and those from Nextstrain (Hadfield et al. 2018).

We conducted Bayesian phylogenetic analyses using BEAST v1.10 using two molecular clock models; a strict clock (SC) and an uncorrelated relaxed clock with an underlying lognormal distribution (UCLN). We set an exponential growth coalescent tree prior, which is appropriate for the early stages of an outbreak and which has been recently used to infer the basic reproductive number and growth rate of SARS-CoV-2 (Volz et al. 2020). For our model comparison in BETS, we estimated (log) marginal likelihoods using generalised stepping-stone sampling (Fan et al. 2011; Baele, Lemey, and Suchard 2016).

Our BETS analyses provided evidence against significant temporal signal in the genome data available up to 23 January 2020 ($n = 22$ genomes). In this data set, the highest model fit to the data was found for analyses with permuted sampling times, followed by those with no sampling times (Fig. 1). Evidence in favour of models without sampling times was also very strong with a highest log Bayes factor of 7.5 for models without sampling times relative to those with the correct sampling times. All data sets obtained subsequently, from 2 February with at

**Table 1.** Description of data snapshots of SARS-CoV-2.

| Publication date range (from 10 January 2020) | Number of genomes | Sampling window (from 23 December 2019) | Days since first genome sample |
|---|---|---|---|
| 23 January | 22 | 17 January 2020 | 31 |
| 2 February | 47 | 27 January 2020 | 41 |
| 6 February | 55 | 28 January 2020 | 45 |
| 10 February | 66 | 3 February2020 | 49 |
| 15 February | 90 | 7 February 2020 | 54 |
| 18 February | 95 | 9 February 2020 | 57 |
| 21 February | 109 | 9 February 2020 | 60 |
| 24 February | 122 | 10 February 2020 | 63 |

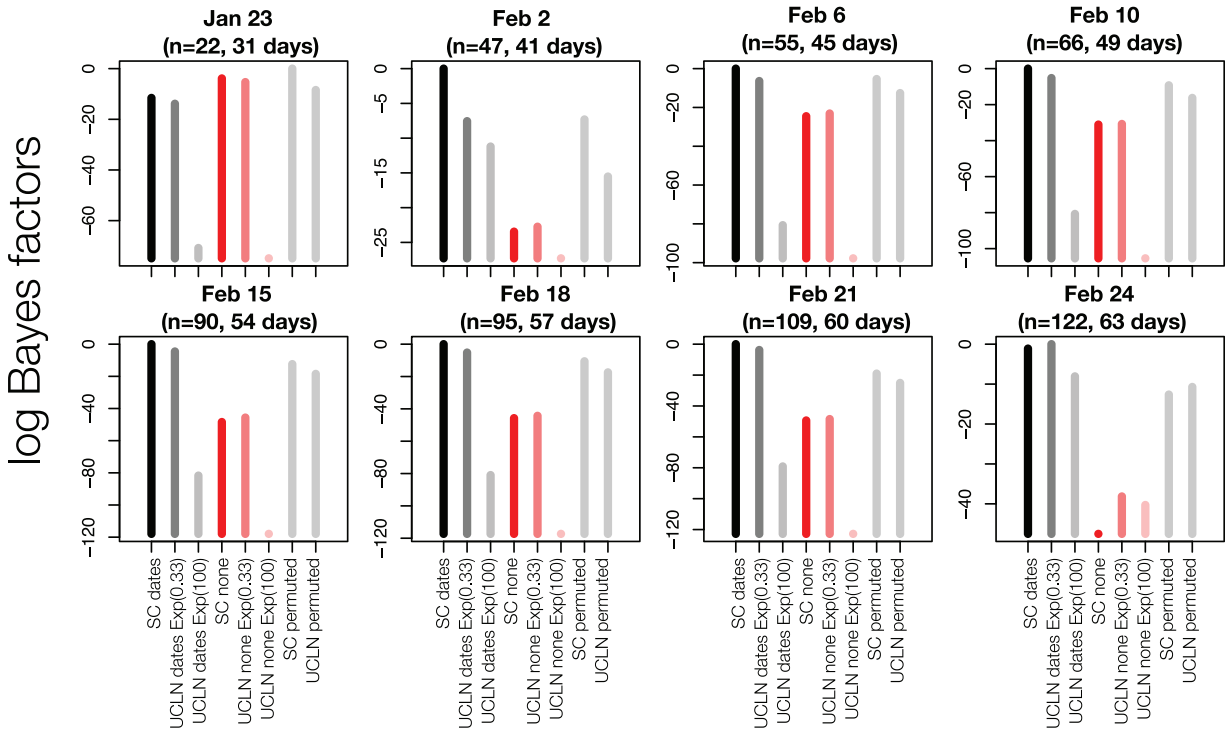



**Figure 1.** BETS results. Each panel corresponds to a snapshot data set collected up to a given month and day in 2020, with a certain number, *n*, of genomes, and the number of days since the first genome sample was collected (23 December 2019). The y-axis represents the log Bayes factors, where the best-performing model has a value of 0. Each bar corresponds to an analysis configuration for BETS, with two possible molecular clock models: the strict (SC) and the uncorrelated relaxed clock with an underlying lognormal distribution (UCLN). For the UCLN, we considered two possible priors on the standard deviation of the lognormal distribution: an exponential distribution with mean 0.33 or with mean 100, labelled as Exp(0.33) and Exp(100), respectively. The sampling times could be configured using the true values (dates), no sampling times (none), or permuted, with these latter two options indicating no temporal signal. For the analyses with permuted sampling times and the UCLN, we used an exponential prior with mean 0.33 for the standard deviation of the lognormal distribution. Black and dark grey bars correspond to analyses with the correct sampling times with the SC or UCLN clock models, respectively. Dark and light red bars are for analyses with no sampling times with these two clock models, and all light grey bars are for analyses with permuted sampling times.

least forty-seven genomes supported the inclusion of the correct sampling times, with log Bayes factors of at least 20 for models with the correct sampling times over those without sampling times. The log Bayes factors for the models with correct sampling times over those with permuted sampling times were at least 5, which is considered as very strong evidence in favour of temporal signal (Kass and Raftery, 1995).

For comparison, we also conducted root-to-tip regressions for the eight snapshot data sets (Fig. 2). The $R^2$ values ranged between 0.11 and 0.2. We did not find an association between $R^2$ and the number of genome samples included. This result may stand in contrast to the expectation that including more independent data should reduce the effect of stochasticity, but the data sets here have an inherently high degree of non-independence. The slopes of the regressions ranged from $6.7 \times 10^{-4}$ to $8.8 \times 10^{-4}$ subs/site/year and the intercept with the X-axis (i.e. the time to the most recent common ancestor) remaining relatively stable at 2019.83 to 2019.86. Although the estimates from the root-to-tip regression are comparable to those previously reported for the virus using explicit phylogenetic methods (Andersen et al. 2020), we note that this approach sometimes produces biased evolutionary rate estimates, possibly due to the fact that internal branches are traversed multiple times (Duchêne et al. 2016). As such, while root-to-tip regression is a valuable tool for visual inspection of the data, it is not a formal molecular clock method nor does it constitute a proper test of temporal signal.

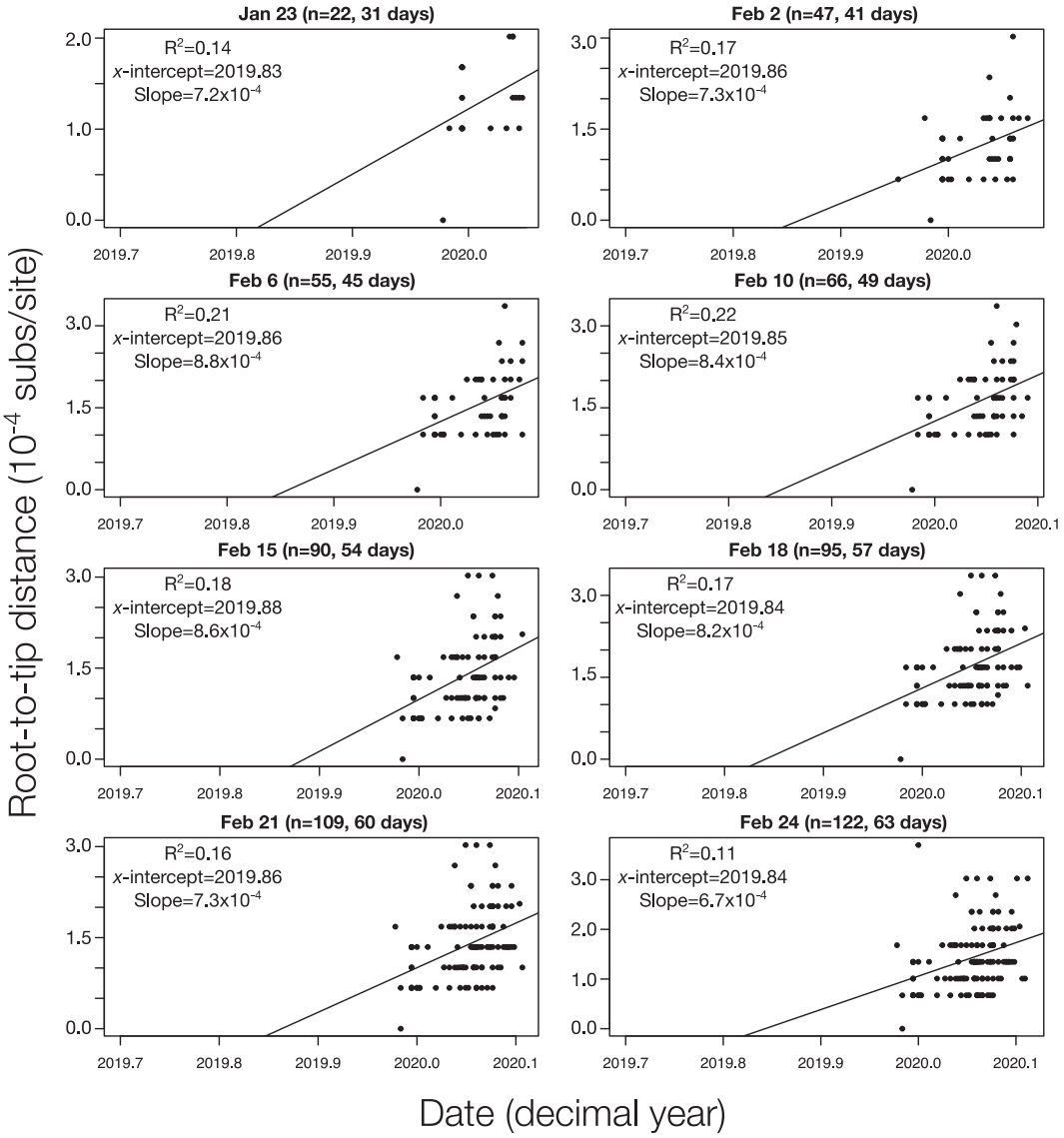



**Figure 2.** Root-to-tip regressions for snapshot data sets. The y-axis corresponds to the root-to-tip distance of phylogenetic trees with branch lengths in units of substitutions per site. The x-axis represents calendar time. Each point corresponds to a tip in the tree. The regression line is the best fitting line using the root position that maximised $R^2$. The $R^2$, the intercept with the x-axis (x-intercept), and slope are shown for each data set, with the latter two representing crude estimates of the evolutionary rate and time of origin, respectively.

Interestingly, all data sets with temporal signal favoured the SC over the UCLN model with the exception of that collected up to 24 February, with 122 genomes, where the log Bayes factor of the UCLN over the SC was 1.81 (Fig. 1). The fact that the SC had high support in data sets collected prior to 24 February probably indicates that they may not be sufficiently informative as to warrant modelling evolutionary rate variation across branches through the UCLN, rather than evidence of strict clocklike behaviour.

A potential reason for why the SC is favoured over the UCLN in many cases is that the default prior on the standard deviation of the lognormal distribution of the UCLN is an exponential distribution with mean 0.33, that has a high density at 0, corresponding to a very low amount of among-lineage rate variation. Intuitively, if the data have low information content, the prior may have a strong influence on the posterior, relative to the data, such that the posterior for this parameter might also be concentrated on 0. In this case, the UCLN may appear overparameterised and the SC would be favoured. We investigated the robustness of model selection to the prior on this parameter by repeating the UCLN analyses with an exponential distribution with mean 100 as the prior for this parameter. Using this less informative prior consistently resulted in a worse model fit across all data sets, and thus did not affect our assessment of temporal signal.

If we restrict our attention to the UCLN with the less informative prior for the 23 January data set, the model that includes sampling times is favoured over that with no sampling times, with a log Bayes factor of 17. If one ignored all other models and priors, this result would indicate the presence of temporal signal. This finding stands in contrast to the SC and UCLN with the more informative prior, which have much higher model fit (Fig. 1; Supplementary Table S2). Consequently, assessing temporal signal using BETS should involve comparing a range of

clock models and careful consideration of the prior on their respective parameters (Duchêne et al. 2020).

We also considered comparisons of prior and posterior distributions to assess the extent to which the data were informative about particular parameters. Our expectation is that the posterior should have a lower variance relative to the prior as more data are included. We considered our estimates of the growth rate ($r$) and scaled population size ($\Phi$) of the exponential coalescent tree prior, the virus's evolutionary rate and the time of origin of the outbreak. An important consideration here is that our method of inspecting the prior consists in running the analyses with no sequence data. Thus, the resulting distributions represent the 'effective', rather than the 'marginal' (i.e. user-specified) prior. The effective prior is the prior conditioned on the number of samples and their ages, the coalescent process and their interaction, whereas the marginal prior is the actual distribution that one sets in the programme. In practice, the effective and marginal prior sometimes differs for parameters that pertain to the tree prior (Warnock et al. 2012; Boskova, Stadler, and Magnus 2018).

Although our marginal priors are identical for all snapshot datasets, we noted that the effective prior differed between data sets for $r$, $\Phi$, and the time of origin (Fig. 3). The posterior from the 23 January snapshot, with twenty-two genomes, was very uncertain for all parameters. For example, the time of origin using the SC ranged from late 2018 to early December 2019. The posterior for $\Phi$ was also more uncertain than its effective prior, which coincides with high uncertainty in the rate and the time of origin.

Our snapshot data sets collected from 2 February, with at least forty-seven genome samples, yielded posterior distributions that were much narrower than their respective effective priors and those of the 23 January snapshot. Our estimates of the evolutionary rate from 2 February converged at a mean of around $1.1 \times 10^{-3}$ substitutions per site per year. The uncertainty in this parameter for the largest data set (24 February, with 122 genomes) using the UCLN clock model is reflected by a 95 per cent credible interval (CI) of between $7.03 \times 10^{-4}$ and $1.5 \times 10^{-3}$ substitutions per site per year. Similarly, the time of origin converged to a mean of late November 2019 and with a 95 per cent CI for the 24 February data set of between late October and mid-December 2019.

Posterior estimates for parameters $r$ and $\Phi$, differed substantially from their effective priors, although to a lesser extent that the evolutionary rate and the time of origin. In particular, the posterior of the time of origin is several times narrower than the prior in all data sets from 2 February, whereas the posterior for $r$ in the largest data set (24 February) is only about two times narrower than its respective effective prior (Fig. 3). Our estimates of $r$ and $\Phi$ did not converge between snapshot data sets, as was the case for the evolutionary rate and time of origin. However, we do not necessarily expect this to happen. For instance, $\Phi$ is proportional to the number of infected individuals at the time of collection of the latest sample (Wallinga and Lipsitch 2007; Boskova, Bonhoeffer, and Stadler 2014), which is expected to increase as the outbreak progresses. Similarly, $r$ is proportional to the reproductive number $R_e$ (i.e. the average number of secondary infections), is expected to decline over time as the number of susceptible individuals decreases and is expected to be affected by spatial structure. In reality, intervention measures such as social distancing and travel restrictions will in most cases result in an earlier decline of this parameter.

## 3. Discussion

The question of whether a viral outbreak has attained the phylodynamic threshold is a highly relevant concept for emerging outbreaks, because it is informative about the amount of sequence data, their temporal spread, and how much evolutionary change has accumulated in the viral genome. The phylodynamic threshold requires a strong assumption about the evolutionary rate based on closely related viruses, and it can be understood as the point in time when sequence data are sufficiently informative about the evolutionary dynamics that

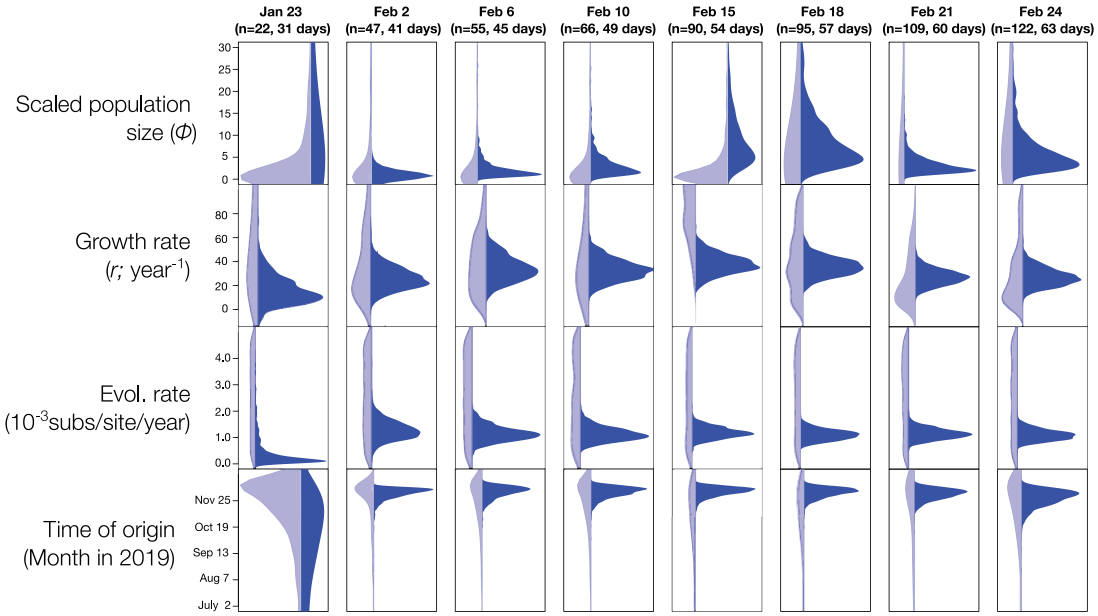



**Figure 3.** Prior and posterior densities for parameters of interest using the molecular clock model with best fit for all snapshot data set (SC for all data sets, except for 24 February, where the UCLN was chosen). The $y$-axis corresponds to parameter values, while the $x$-axis represents the relative density. Light blue densities correspond to the effective prior, while those in dark blue show the posterior.

shape an outbreak, i.e. when the population is measurably evolving. The routine application of tests of temporal signal can effectively answer this question in nearly real-time. Our application of BETS (Duchêne et al., 2019) to data snapshots from the early stages of the outbreak revealed that the phylodynamic threshold of SARS-CoV-2 was reached by about 2 February, when forty-seven genomes were available sampled over 35 days, and 41 days after the first genome was reported.

Our finding that the phylodynamic threshold was attained within about two months of the estimated start of the outbreak demonstrates that Bayesian phylodynamic approaches can capitalise on early collected genome data to make inferences about evolutionary processes, particularly the viral evolutionary rate and the outbreak's time of origin. Our estimates of these two parameters were consistent after the phylodynamic threshold was reached, and also matched previous estimates posted on virological.org and elsewhere (Taiaroa et al. 2020; Volz et al. 2020). Increasing the number of sequences leads to more precise estimates of the evolutionary rate, but we found only marginal improvements in precision after 109 sequences (21 February). The SC was preferred over the UCLN in most data sets. The fact that the UCLN was only supported after 122 sequences were included suggests that the statistical power necessary to support such a relaxed clock model may require more informative data than those available at the early stages in the outbreak. We anticipate that the UCLN will be favoured over the SC in analyses of larger data sets of SARS-CoV-2. We also observed a slight decrease in the precision of evolutionary rate and tMRCA estimates beyond 6 February (Fig. 2 and Fig. 3, respectively), and for clock model support as rate variation increases among these larger samples, a pattern that has been demonstrated previously (Duchêne et al. 2020).

The relative importance of sample size and time span in achieving the phylodynamic threshold is difficult to quantify as this is highly contingent upon the given outbreak. In a recent simulation study, we found that sequence diversity—measured as the number of site patterns—is a key factor in the detection of temporal signal (Duchêne et al. 2020), which could be varied in simulations by modifying the evolutionary rate or increasing sequence length. Ultimately a sufficient amount of evolutionary change will be needed to reach this threshold. However, ongoing work on SARS-CoV-2 has shown that including large amounts of sequence data can sometimes obscure temporal signal due to an increase of among-lineage rate variation, indicating that the phylodynamic threshold does not follow a simple recipe of increasing the number of samples or the sampling window. As a case in point, the Nextstrain workflow uses a fixed evolutionary rate to infer the time scale of SARS-CoV-2, avoiding potential errors in the resulting estimates of evolutionary time scales in downstream analyses (Dellicour et al. 2020).

A key consideration concerning the presence of temporal signal in the data is that this does not necessarily imply that demographic parameters can be reliably estimated using genome sequence data. Comparing the effective prior and posterior is important to assess the information content of the data, but it is not an assessment of the reliability of the estimates. For example, $\Phi$ is generally inversely correlated with the root height, such that if the data have temporal signal, the prior and posterior for this parameter will substantially differ. However, this parameter is proportional to the number of infected individuals at present under the assumption that the number of infections grows exponentially in a deterministic fashion and in the absence of population structure. Clearly, the extent to which the data meet these conditions can affect the interpretation and reliability of such epidemiological parameters. More realistic tree priors may be warranted here, such as those that account for population structure and the sampling process (Scire et al. 2020). In sum, whether the phylodynamic threshold coincides with reliability in estimates of epidemiological parameters depends on the information content in the data, but also on the tree prior and its underlying assumptions.

Ongoing analyses of SARS-CoV-2 will reveal important aspects regarding its evolutionary origin and epidemiological dynamics. On a global scale, the virus is well beyond its phylodynamic threshold, but tests of temporal signal, as applied here, will still be key to understand the time scale of local transmission.

## 4. Methods

We downloaded genome sequence data from GISAID or GenBank, and aligned them using MAFFT (Katoh et al. 2002). We curated the data through comparison with data sets available at virological.org and visual inspections of our alignments (Supplementary Table S1). We only included sequences from humans, that were at least 28,000 nucleotides long, and with high coverage.

### 4.1 Bayesian phylogenetic analyses

We analysed each data snapshot in BEAST (Suchard et al. 2018) using the HKY + $\Gamma$ substitution model. We set a Markov chain Monte Carlo length of $10^7$ steps, sampling every $10^3$ steps. We determined sufficient sampling by verifying that the effective sample size of key parameters was at least 200 using Tracer v1.7 (Rambaut et al. 2018). We assessed temporal signal using BETS (Duchêne et al. 2020). We compared the statistical fit of two molecular clock models, SC and UCLN, and three configurations of sampling times; the correct sampling times, no sampling times, and permuted sampling times, with the latter two corresponding to a lack of temporal signal. For each combination of molecular clock model and sampling times we calculated the (log) marginal likelihood using generalised stepping-stone sampling (Baele, Lemey, and Suchard 2016), for which we employed 200 path steps with a chain length for each power posterior of $10^5$ iterations. We chose priors for all parameters that respected their respective domains, but that were not overly informative, and all of which are proper (i.e. the area under the curve is 1.0; Baele, Lemey, and Suchard 2012) (Table 2). According to BETS, a data set is considered to have temporal signal if (log) Bayes factors support a model with the correct sampling times (Duchêne et al. 2020).

Our comparison of the prior and posterior distributions of key parameters require obtaining the effective, rather than the marginal prior. The effective prior can be obtained by running the analysis in BEAST with no sequence data, which is equivalent to ignoring the sequence likelihood and is done by selecting the option 'sample from prior' in BEAUti, the graphical interface accompanying the BEAST software package (Suchard et al. 2018). All Bayesian phylogenetic analyses were conducted on the SPARTAN high-performance computing service of the University of Melbourne (Meade et al. 2017).

### 4.2 Root-to-tip regression

We estimated phylogenetic trees using maximum-likelihood inference as implemented in IQ-tree v1.6 (Minh et al. 2020), with the optimal substitution model determined by the programme. We used these trees to obtain root-to-tip regressions in

**Table 2.** Prior distributions used for key parameters.

| Parameter | Prior |
|---|---|
| Evolutionary (clock) rate | Continuous time Markov chain (CTMC) |
| Standard deviation of evolutionary rate (UCLN only) | Exponential (mean = 0.33 or mean = 100) |
| Exponential coalescent growth rate | Laplace ($\mu = 0$, scale = 100) |
| Exponential coalescent population size | Lognormal ($\mu = 1.0$, $\sigma = 5$) |

TempEst v1.5 (Rambaut et al. 2016) by selecting the root position that maximised $R^2$.

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

**Conflict of interest:** None declared.

## Acknowledgements

We thank all those who have contributed sequences to the GISAID database (https://www.gisaid.org/). We are grateful to Editor-in-Chief Oliver Pybus and two anonymous reviewers for their helpful comments.

## Funding

S.D. was supported by an Australian Research Council Discovery Early Career Researcher Award (DE190100805) and an Australian National Health and Medical Research Council (APP1157586). A.R. was supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z-ARTIC network). P.L. acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (725422-ReservoirDOCS) and the Research Foundation—Flanders ('Fonds voor Wetenschappelijk Onderzoek—Vlaanderen', G066215N, G0D5117N, and G0B9317N). G.B. acknowledges support from the Interne Fondsen KU Leuven/Internal Funds KU Leuven (C14/18/094)), and the Research Foundation—Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G0E1420N).

## References

Andersen, K. G. et al. (2020) 'The Proximal Origin of SARS-CoV-2', *Nature Medicine*, 26: 450–2.

Baele, G., Lemey, P., and Suchard, M. A. (2016) 'Genealogical Working Distributions for Bayesian Model Testing with Phylogenetic Uncertainty', *Systematic Biology*, 65: 250–64.

—— et al. (2012) 'Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics', *Molecular Biology and Evolution*, 30: 239–43.

—— et al. (2017) 'Emerging Concepts of Data Integration in Pathogen Phylodynamics', *Systematic Biology*, 66: e47–e65.

Biek, R. et al. (2015) 'Measurably Evolving Pathogens in the Genomic Era', *Trends in Ecology & Evolution*, 30: 306–13.

Boskova, V., Bonhoeffer, S., and Stadler, T. (2014) 'Inference of Epidemiological Dynamics Based on Simulated Phylogenies Using Birth-Death and Coalescent Models', *PLOS Computational Biology*, 10: e1003913.

——, Stadler, T., and Magnus, C. (2018) 'The Influence of Phylodynamic Model Specifications on Parameter Estimates of the Zika Virus Epidemic', *Virus Evolution*, 4: vex044.

Dellicour, S. et al. (2020) 'A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 lineages,' bioRxiv 2020.05.05.078758.

Drummond, A. J. et al. (2003) 'Measurably Evolving Populations', *Trends in Ecology & Evolution*, 18: 481–8.

Duchêne, S. et al. (2015) 'The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data', *Molecular Biology and Evolution*, 32: 1895–906.

—— et al. (2016) 'Estimating Evolutionary Rates Using Time-Structured Data: A General Comparison of Phylogenetic Methods', *Bioinformatics*, 32: 3375–9.

—— et al. (2019) 'Inferring Infectious Disease Phylodynamics with Notification Data,' bioRxiv.:596700.

—— et al. (2020) 'Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations,' *Molecular Biology and Evolution*, msaa163.

Fan, Y. et al. (2011) 'Choosing among Partition Models in Bayesian Phylogenetics', *Molecular Biology and Evolution*, 28: 523–32.

Gojobori, T., Moriyama, E. N., and Kimura***, M. (1990) 'Molecular Clock of Viral Evolution, and the Neutral Theory', *Proceedings of the National Academy of Sciences of the United States of America*, 87: 10015–8.

Hadfield, J. et al. (2018) 'Nextstrain: Real-Time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.

Hedge, J., Lycett, S. J., and Rambaut, A. (2013) 'Real-Time Characterization of the Molecular Epidemiology of an Influenza Pandemic', *Biology Letters*, 9: 20130331.

Kalkauskas, A. et al. (2020) Sampling Bias and Model Choice in Continuous Phylogeography: Getting Lost on a Random Walk,' bioRxiv 2020.02.18.954057.

Kass, R. E., and Raftery, A. E. (1995) 'Bayes Factors', *Journal of the American Statistical Association*, 90: 773–95.

Katoh, K. et al. (2002) 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform', *Nucleic Acids Research*, 30: 3059–66.

Korber, B. et al. (2000) 'Timing the Ancestor of the HIV-1 Pandemic Strains', *Science*, 288: 1789–96.

Meade, B. et al. (2017) 'Spartan HPC-Cloud Hybrid: Delivering Performance and Flexibility.' University of Melbourne, 10: 49.

Minh, B. Q. et al. (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37: 1530–4.

Murray, G. G. R. et al. (2016) 'The Effect of Genetic Structure on Molecular Dating and Tests for Temporal Signal', *Methods in Ecology and Evolution*, 7: 80–9.

du Plessis, L., and Stadler, T. (2015) 'Getting to the Root of Epidemic Spread with Phylodynamic Analysis of Genomic Data', *Trends in Microbiology*, 23: 383–6.

Rambaut, A. et al. (2016) 'Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (Formerly Path-O-Gen)', *Virus Evolution*, 2: vew007.

—— et al. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic Biology*, 67: 901–4.

Scire, J. et al. (2020) 'Improved Multi-Type Birth-Death Phylodynamic Inference in BEAST 2,' bioRxiv 2020.01.06.895532.

Suchard, M. A. et al. (2018) 'Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10', *Virus Evolution*, 4: vey016.

Taiaroa, G. et al. (2020) 'Direct RNA Sequencing and Early Evolution of SARS-CoV-2,' bioRxiv 2020.03.05.976167.

Volz, E. M. et al. (2020) 'Report 5: Phylogenetic Analysis of SARS-CoV-2,' *MRC Cent. Glob. Infect. Dis. Anal.*

Wallinga, J., and Lipsitch, M. (2007) 'How Generation Intervals Shape the Relationship between Growth Rates and Reproductive Numbers', *Proceedings of the Royal Society B: Biological Sciences*, 274: 599–604.

Warnock, R. C. M., Yang, Z., and Donoghue, P. C. J. (2012) 'Exploring Uncertainty in the Calibration of the Molecular Clock', *Biology Letters*, 8: 156–9.