June 15, 2022

Assessing the effects of date and sequence data in phylodynamics

Leo A. Featherstone*[,1], Sebastian Duchene[1], Timothy G. Vaughan[2,3]

[1] Peter Doherty Institute for Infection and Immunity, University of Melbourne, Australia.

[2] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.

[3] Swiss Institute of Bioinformatics.

email: leo.featherstone@unimelb.edu.au

Significance Statement

Phylodynamic analysis almost always uses time stamped genome sequences to make epidemiological inference. However the field has lacked a transferable methods for quantifying and comparing the effects of data and genome data. Moreover, the ability to make this inference will be valuable for the future as the scale of pathogen genome sequence increases such that the question of maximising knowledge gain relative to sequencing effort comes into play. We introduce a fully transferable method to quantify the effects of date and sequence data and answer this question on a per-analysis basis.

Abstract

Birth-death based methods are often used to infer epidemiological parameters from pathogen genomes in phylodynamics. However, these methods also base their results on sampling time data in addition to genome sequence data. We introduce a formal method for quantifying the relative impacts of the date and sequence components of the data. We show that either data source can drive inference of the basic reproductive number, R_0, but note that the general approach can be used to investigate other phylodynamic parameters. This framework will allow practitioners to draw conclusions

about which aspect of the data drive inference results, providing a path to a deeper understanding of commonly-used phylodynamic models and to better direct sequencing efforts during future outbreaks.

Introduction

Phylodynamics combines phylogenetic and epidemiological modelling to infer epidemiological dynamics from pathogen genome data (du Plessis and Stadler, 2015; Baele et al., 2018; Volz et al., 2013). Analyses are usually conducted within a Bayesian framework, meaning that the output comprises posterior distributions for parameters of interest, such as the basic reproductive number, R_0 (i.e. the average number of secondary infections in a fully susceptible population). Input data usually consists of time-stamped genome sequences. In the case of birth-death-sampling models (Stadler, 2010), both sequence and date data inform the branching of inferred trees by either temporally clustering lineages or via sequence similarity. Internal nodes are assumed to co-occur with transmission events, such that they provide information about patterns of transmission that sampling time data alone cannot . Sampling times, or date data, are similar to standard epidemiological time series data while sequence data introduce the evolutionary aspect. The widely used birth-death model uses sampling times to infer a sampling rate which is also informative about transmission rates (Boskova et al., 2018; Stadler et al., 2012).

Phylodynamics is experiencing greater use than ever before since the onset of the SARS-CoV-2 pandemic. This commonly includes application to larger and more densely sequenced outbreaks than previously. While the value of pathogen genome data is now well established, an emergent question is whether inclusion of more sequence data after a point is of diminishing returns for some densely sequenced outbreaks. The answer to this question will naturally vary with each dataset and pathogen considered, but a method to quantify the individual effects of date and sequence data presents a transferable way to address it. It would substantially broaden our understanding of the phylodynamic tools that now feature in infectious disease surveillance. It also has the potential to direct sampling efforts to future outbreaks for optimisation of knowledge gain against resource

expenditure.

Earlier work showed that sequence sampling times, referred to here as 'date data', can drive epidemiological inference under the birth-death model (Volz and Frost, 2014; Boskova et al., 2018; Featherstone et al., 2021). However, each stopped short of proposing a transferable way to measure this effect in regular application. The birth-death model is most applicable to the question at hand since it includes a rate of sampling. The coalescent is another a key phylodynamic model, but it typically conditions on sampling dates which therefore precludes from a comparison of date and sequence effects. Some coalescent formulations include a sampling rate (Volz and Frost, 2014), however these are used less often than the birth death or standard coalescent. The coalescent also assumes a low sampling proportion relative to population size such that its typical formulation would be inappropriate for many densely sequenced outbreaks (Boskova et al., 2018), where the question of the effect of large amounts of sequence data is most relevant.

Building upon these earlier results, we introduce a theoretical framework and a new method to quantify the effect of sequence and dates for any parameter under the birth death with continuous sampling. We focus on continuous sampling because it is most relevant to how emerging outbreak data are collected. Our method quantifies and visualises the effect each data source has on the posterior distribution of epidemiological parameters of interest. It also classifies which data source is driving the inference, but crucially also provides a measure of whether a binary classification is meaningful. These observations are a critical addition the phylodynamic toolkit used to inform public health decisions because they clearly quantify the added-knowledge acquired from genomes in a given analysis.

Table 1: Separating data to quantify effects

	Dates Included	Dates Excluded
Sequence Included	Combined effect	Sequence Effects
Sequence Excluded	Date effects	Marginal Prior

New Methods

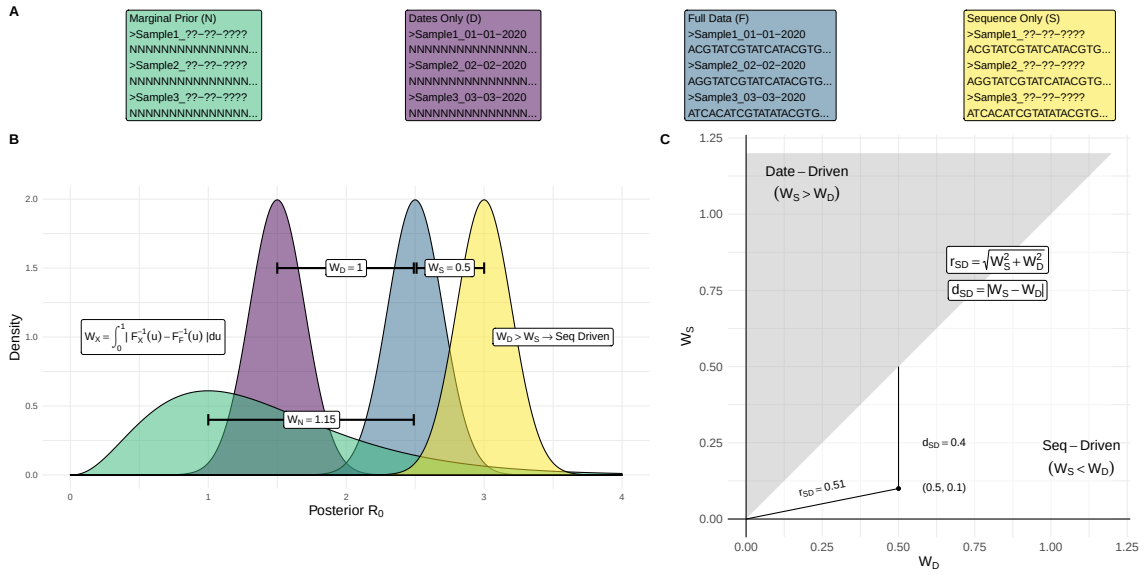

Figure 1: Graphical summary of the process to quantify signal and classify signal drivers. **A**) Coloured boxes give examples of data under each of the 4 treatments with letters in brackets giving shorthand notation for each. From left to right: *Marginal Prior* results from the removal of both date and sequence data. *Dates Only* includes date data while ignoring sequence data through a constant phylogenetic likelihood. This can be represented as converting all sequence characters to 'N' (i.e. alignments of fully missing data). *Full data* represents the usual combination of both sequence and date data. This produces a reference distribution from which the Wasserstein metric to other posteriors is calculated. *Sequence Only* corresponds to the removal and re-estimation of dates while sequence data is retained. **B**) Example posterior output for R_0 with colour corresponding to each treatment in **A**. The Wasserstein metric is calculated as difference in inverse distribution function of each posterior from *Full Data* integrated over 0 to 1. Example values for the Wasserstein metric are given in white boxes. **C**) The plane with x and y axes W_D and W_S and shaded classification regions. r_{SD} is the Euclidean distance from the origin to a point (W_D, W_S), with higher values indicating that one or both of data and sequence data drive differing signals from the reference posterior. d_{SD} is the vertical distance from a point (W_D, W_S) to the line $y = x$, with points closer to this line corresponding to more similar data and sequence data effects such that classification is less meaningful. In the example, distance from the posterior under only sequence data to the full data posterior (W_S) is smallest, leading to classification as 'Seq-Driven'.

Isolating date and sequence data

We conduct four analyses for a given dataset to contrast the effects of complete data, date data, sequence data, and the absence of both (fig 1A). We focus on inferring R_0, with all other parameters fixed, but this new approach is applicable to any parameter under the birth-death with any combination of priors. First, we use complete data to fit a birth-death model and infer the posterior distribution of R_0. This represents the combined effects of dates and sequences. Second, to isolate

the effect of date data, we remove sequence information and retain dates, thus integrating over the prior on tree topology. This is traditionally referred to as 'sampling from the prior', but this term should be avoided in the context of models where the sampling times are treated as data, such as the birth-death. Third, to isolate the effect of sequence data, we keep sequence data and remove dates. This requires estimation of all sampling dates, analogously to how removing sequence data causes integration over topology. We use a novel Markov chain Monte Carlo (MCMC) operator to estimate dates which is implemented in the feast v17 package for BEAST 2 (Bouckaert et al., 2019). Last, and for completeness, we conduct the analysis with both date and sequence data removed. This formally corresponds to the marginal prior conditioned on the number of samples. The resulting Wasserstein metric, W_N, is useful for quantifying whether full data offer information in addition to the prior.

Quantifying data signal

We employ the Wasserstein metric in one dimension to measure a "distance" between each of the sequence posterior, date posterior, or marginal prior, and the posterior derived from the complete data. We write these distances as $W_\bullet$, with $\bullet$ being D, S, or N for the sequence, date, and marginal prior distributions, respectively. For example, the Wasserstein distance W_D from the date posterior to complete data as:

$$W_D = \int_0^1 |F_D^{-1}(u) - F_F^{-1}(u)|\, du,$$

where F_D^{-1} and F_F^{-1} are the inverse empirical distribution functions for the posterior R_0 inferred under date and complete data respectively. The units of $W_\bullet$ are equivalent to the units of the parameter of interest.

As in Fig 1C, we can now consider a plane where the axes are W_D and W_S. We classify the data source with the lowest Wasserstein distance from the complete data posterior as contributing most to the posterior with sequence and date data. In this case the lines $y = x$ marks the classification boundary as in the shading in Fig 1C.

Finally, we can quantify the disagreement in signal between each data source. We define disagreement with respect to the full data posterior, r_{SD} as the magnitude of the vector $(\overrightarrow{W_D, W_S})$ leading to each point in the plane, which is the radius from the origin to the point in other words. Values near zero indicate that the posteriors under data, sequence, and complete data are all near identical and classification of date or sequence driven is less meaningful. Larger values signify that one or both data sources drive differing posteriors and classification is more meaningful. We also define disagreement without respect to full data $d_{SD} = |W_S - W_D|$ as a quantification of disagreement between date and sequence posteriors without respect to full data (Fig S1C). Visually, this corresponds to the vertical distance to the nearest classification boundary ($y = x$) such that smaller values correspond to less meaningful classification. r_{SD} and d_{SD} are similar in that when r_{SD} is near-zero, d_{SD} necessarily is too. d_{SD} also accounts for the case where r_{SD} is high, but both date and sequence data have similarly sized effects. In this case, r_{SD} is higher while d_{SD} is lower and classification of one or another as driving analysis is inaccurate. Tree 8 in Fig 2 presents an example of this.

Results

We simulated 400 alignments to explore the differing signals in date and sequence data using the Wasserstein metric. These derive from 100 simulated outbreaks of 100 cases, sampled with proportion 1 or 0.5, and used to simulate sequences with an evolutionary rate of 10^{-3} or 10^{-5} (subs/site/time). Higher evolutionary rates imply that there are more site patterns and therefore more informative sequence data. We estimated R_0 under each data treatment with all other parameters fixed using a birth-death tree prior. In all analyses, simulated data provided information in addition to the prior ($W_N >\approx 0$, Fig S2). Among the 400 datasets, we observe a mixture of cases where date and sequence data infer similar or dissimilar posterior R_0. This supports the core assumption that date and sequence data can have differing signals concealed in their combination (Fig 2). Classification using the Wasserstein metric results in a mix of date and sequence driven classifications, supporting that our proposed method is sensitive to differences between datasets (Fig 3 A-C). Most datasets

were classified as date-driven (334/400), which is consistent with earlier work showing that dates are highly influential under the birth-death Volz and Frost (2014).

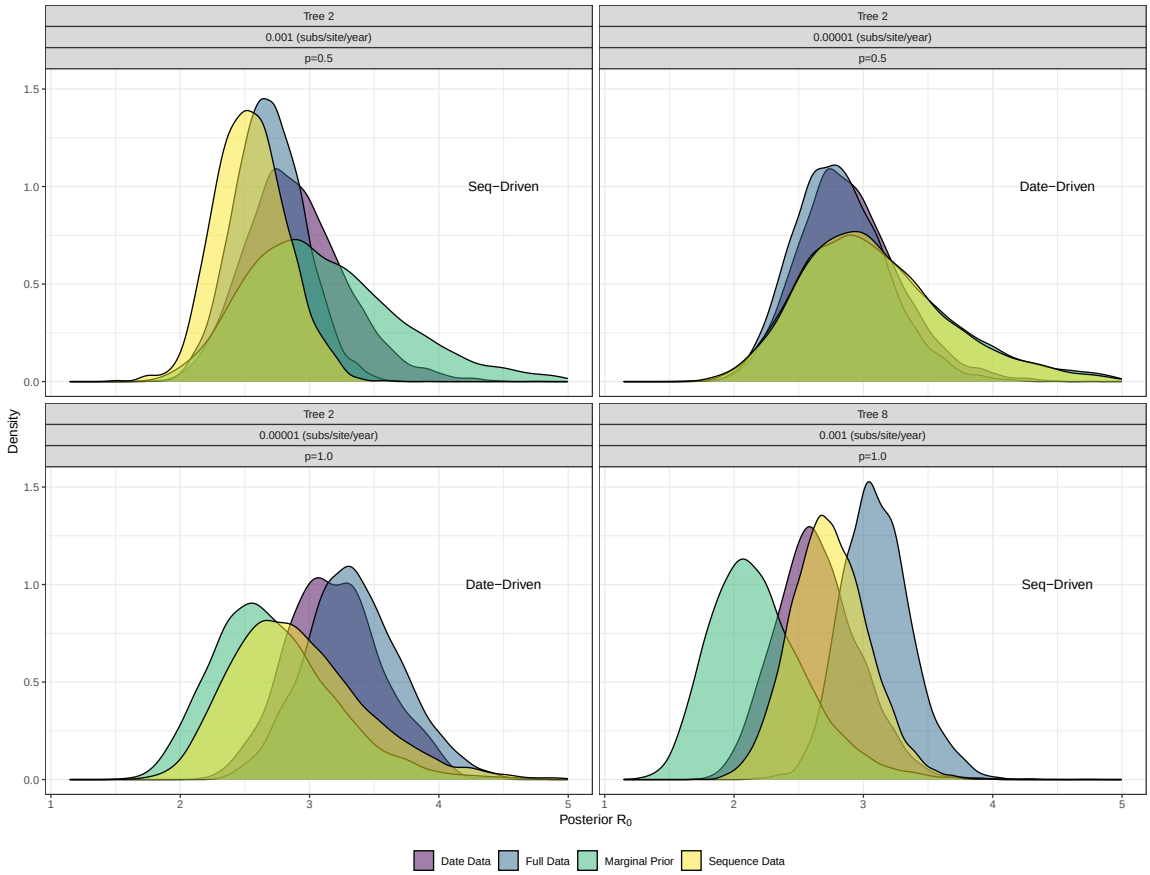

Figure 2: Comparison of full data, date data, sequence data, and no data posteriors for different trees and alignments. Each plot gives an example of posterior output for R_0 under each treatment for different one of the 400 test datasets. Tree 8 gives an example of when both date and sequence posteriors are similar while differing from the full data posterior. Here, r_{SD} is higher, while d_{SD} is lower such that classification using the Waserstein metric is both less meaningful and more likely to reflect noise due to MCMC sampling.

Reliability of Wasserstein Metric

Alignments were simulated under two sampling conditions ($p = 0.5$ and $p = 1$) to test robustness with respect to sampling proportion. We found that sampling proportion does not bias the distribution of the Wasserstein statistic between analyses (Fig 3 D). This supports the assumption that comparing Wasserstein distance between analyses captures different signals concealed in date and sequence data.

We also tested the accuracy of classification using the Wasserstein metric by subsampling and reclassifying each posterior R_0 distribution 100 times. Only 329 of resulting 40000 subsampled posteriors were misclassified across 17 of the 400 datasets (S1). Misclassifiifcation only occurred for datasets where d_{SD} was below $10^{-1.47}$. Smaller d_{SD} values indicate that the effect size of date and sequence data is similar and classification is not therefore not useful.

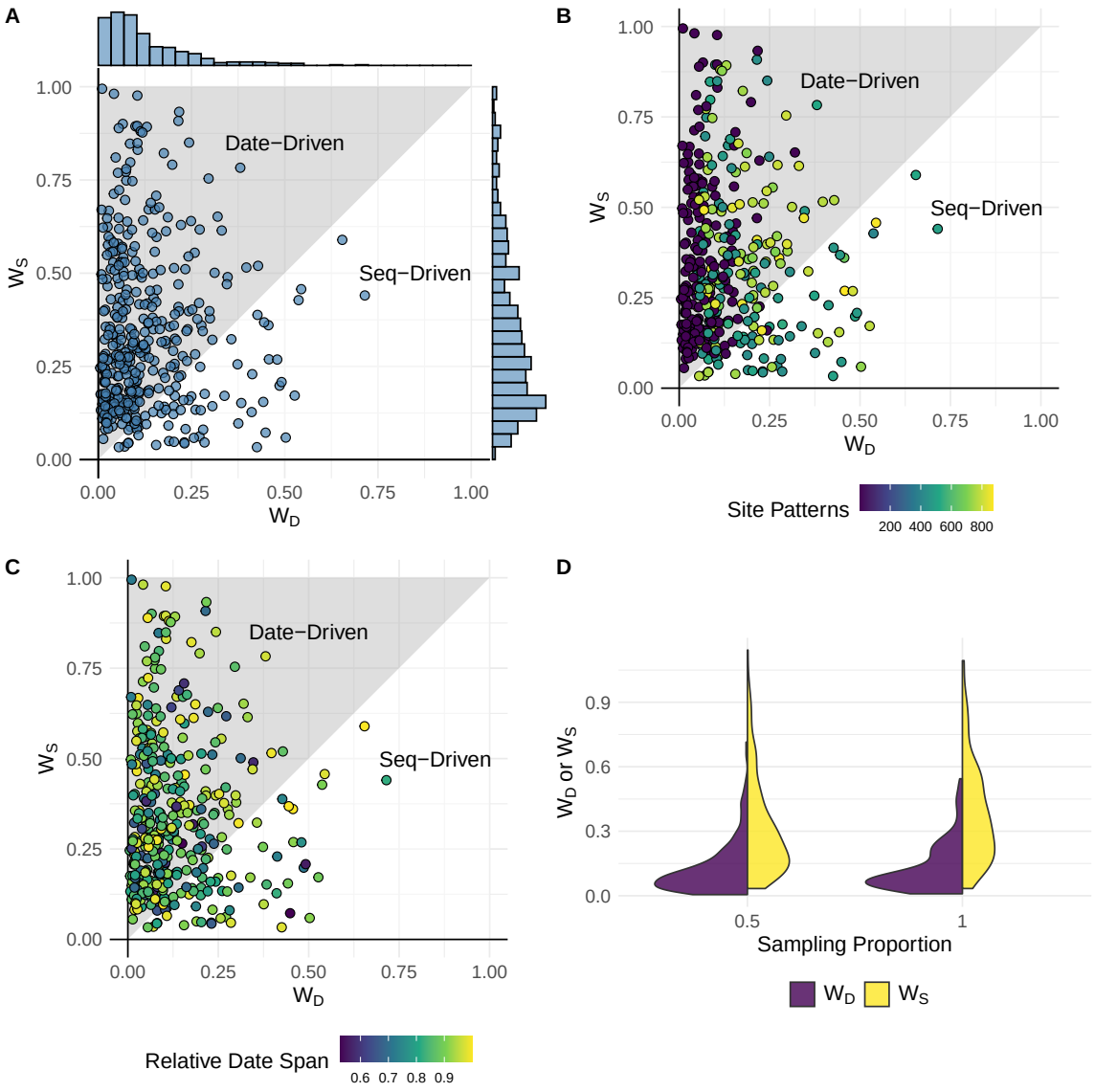

Figure 3: **A-C**) Each point represents (W_D, W_S) for one of the 400 simulated datasets. **A**) The distribution of points with marginal histograms corresponding to the distribution of W_D and W_S respectively. More datasets are classified as date-driven, which is consistent with prior results about the impact of sampling times under the birth-death. **B**) Points coloured by number of site patterns. Lower site patterns tends to co-occur with date-driven classification. **C**) Points coloured by date span with no clear patterns corresponding with classification as for site patterns. **D**) Distribution of W_S and W_D under $p = 0.5$ and $p = 1$. The similarity of distributions across sampling proportion does not drive differences in W_S and W_D.

Observations about the effects of Sequence and Dates

The distributions of W_S is more diffuse than W_D, meaning the sequence data posterior tends to differ more from full data than date data (Fig 3 A). This again aligns with previous results showing

that date data drive inference under the birth-death.

Low sequence diversity, measured here in the number of site patterns, seems to preclude sequence data from driving inference (Fig 3B, Fig S3). This matches the expectation that fewer site patterns results in less sequence information to inform the posterior. On the other hand, the date span does not follow equivalent trend with lower diversity coinciding with analyses being sequence driven. Here the relative date span is the time between the first and last sample, divided by the height of the outbreak tree and thus should be indicative of the information content of the dates. The distribution of relative date-span appears random across classifications, unlike the distribution of site patterns (Fig 3B, S3).

Empirical Results

We analysed data from two SARS-CoV-2 transmission clusters from 2020 in Australia to demonstrate that date-driven and sequence-driven analyses can arise in practice (table 2). The first cluster is classified as sequence-driven, with $r_{SD} = 0.223$ indicating an appreciable difference between the sequence posterior and complete data. $d_{SD} = 0.078$ adds that date data also drives an effect of similar size, offering the interpretation that both date and sequence data are influential in this analysis. The second cluster is classified as date-driven, but with $r_{SD} = 0.009$ and $d_{SD} = 0.008$. The low r_{SD} value indicates a near-negligible difference between date, sequence, and complete data posteriors. Since r_{SD} is low, d_{SD} is necessarily also low. Due to this, classification is effectively meaningless and it can be concluded that both date and sequence data drive a highly congruent signal. Moreover, W_N for both analyses is more than double each of W_S and W_D, which affirms that both sources of data contribute to the posterior deviating from the prior and are therefore informative with respect to the prior in both analyses.

Table 2: Empirical data

	Cluster 1	Cluster 2
n	112	188
Classification	Seq-Driven	Date-Driven
r_{SD}	0.223	0.009
d_{SD}	0.078	0.008
W_D	0.192	0.001
W_S	0.114	0.009
W_N	0.325	0.481

Discussion

The results of our simulation study add clarity to previous work showing that sampling times contribute substantially to phylodynamic inference under the birth death (Volz and Frost, 2014; Featherstone et al., 2021). We demonstrate that lower sequence diversity often precludes sequence data from a comparable effect. We also demonstrate that sequence data are not always secondary in influence and can drive inference of R_0 in some instances, affirming the sensitivity of the birth-death to the signal encoded in sequence data. The tendency for date data to drive inference more than sequence data may be explained by the reduction in uncertainty that each data source offers. Dates impose a hard bound on topology by restricting tree space to a subset of topologies that agree with the chronology of sampling times. Conversely, sequence data inform topology through phylogenetic likelihood, but do not definitively constrain tree space in the same way as date data.

Our method offers novel insight for phylodynamics practitioners because the contribution of sequence data relative to sampling dates is often questioned in birth death analyses of densely sequenced outbreaks. It offers a reproducible approach to answering this question so the influence of genomic data can be commented on. For example, this is useful in a public health reporting context so phylodynamics experts can comment on the drivers inference as new data emerge. As such it offers a tool to initiate research into optimal sampling design for phylodynamic analysis. Any resulting theory can only be speculative at best, given the unpredictable nature of evolution, however this is an important consideration for the future as the scale of pathogen genome sequencing

increases.

Data Archival

All scripts used to simulate and analyse data are available at `https://github.com/LeoFeatherstone/phyloDataSignal.git`. The Feast package, containing our date estimator, is available at `https://github.com/tgvaughan/feast.git`.

Supplementary Methods

Simulation Study

We simulated 100 outbreaks of 100 cases under a birth death process using the `Tree-Sim` R package (Stadler, 2019). The birth rate was set to 2.5, death rate 1, corresponding to $R_0 = 2.5$, and sampling probability $p = 1$, resulting in trees with 100 tips. We then extended this to a set of 200 outbreaks by sampling again with probability $p = 0.5$, resulting in trees of 50 tips. We used a consistent seed such that each outbreak with $p = 0.5$ corresponds to a subsample of another with $p = 1$, allowing us to asses the effect of sampling proportion on inferring $W_\bullet$. For each outbreak, we simulated two sequence alignments of length 20000, which is roughly average for RNA viruses (Sanjuán et al., 2010). We set an HKY model with evolutionary rate set to either 10^{-3} or $10^{-}5 subs/site/time$ using Seq-Gen Rambaut and Grass (1997). Our choice of evolutionary rates allows us to compare the effects higher and lower sequence information with the former corresponding to about 20 substitutions per infection, and 0.2 for the latter. The above resulted in 400 alignments to test in the four treatments described above. We analysed each under a birth-death model with a $Uniform[0, 5]$ prior for R_0 and all other parameters set to the true value for simplicity and to disentangle any impacts of parameter nonidentifiability (Louca et al., 2021).

Empirical Data

We analysed two similar SARS-CoV-2 datasets taken from (Lane et al., 2021). They consisted of 112 and 188 samples respectively. We analysed each dataset under the four conditions above. In each, we placed a $Lognormal(mean = 1, sd = 1.25)$ prior on R_0 and an $Inv - Gamma(\alpha = 5.807, \beta = 346.020)$ prior on the becoming-uninfectious rate (δ) following estimates of the duration of infection ($= \frac{1}{\delta}$) in Lauer et al. (2020). We also fixed the sampling proportion to $p = 0.8$, and placed an $Exp(mean = 0.019)$ prior on the origin, corresponding to a lag of up to one week between the index case the first putative transmission event.

Validating the Wasserstein Metric

We use the transport R package to calculate the Wasserstein metric. We conducted an analysis to ensure that classification using the Wasserstien statistic reflects differences between date and sequence data, rather than noise alone. For each of the 400 test datasets, we subsampled each posterior 100 times with probability 0.5 and reclassified each subsampled dataset (n=40000). Of these, only 329 were misclassified. Of the 400 simulated datasets, those with any degree of misclassified subsamples had substantially smaller differences between Wasserstein distance to the date-only and sequence-only posteriors (d_{SD})) (Figure S1). Misclassification occurred for (d_{SD})) less than roughly $10^{-1.5}$, with complete reliability above this level. Differences below $10^{-1.5}$ correspond to a level of difference between data and sequence only posteriors where classification is of little significance. Classification is wholly reliably above this threshold, where classification is more warranted.

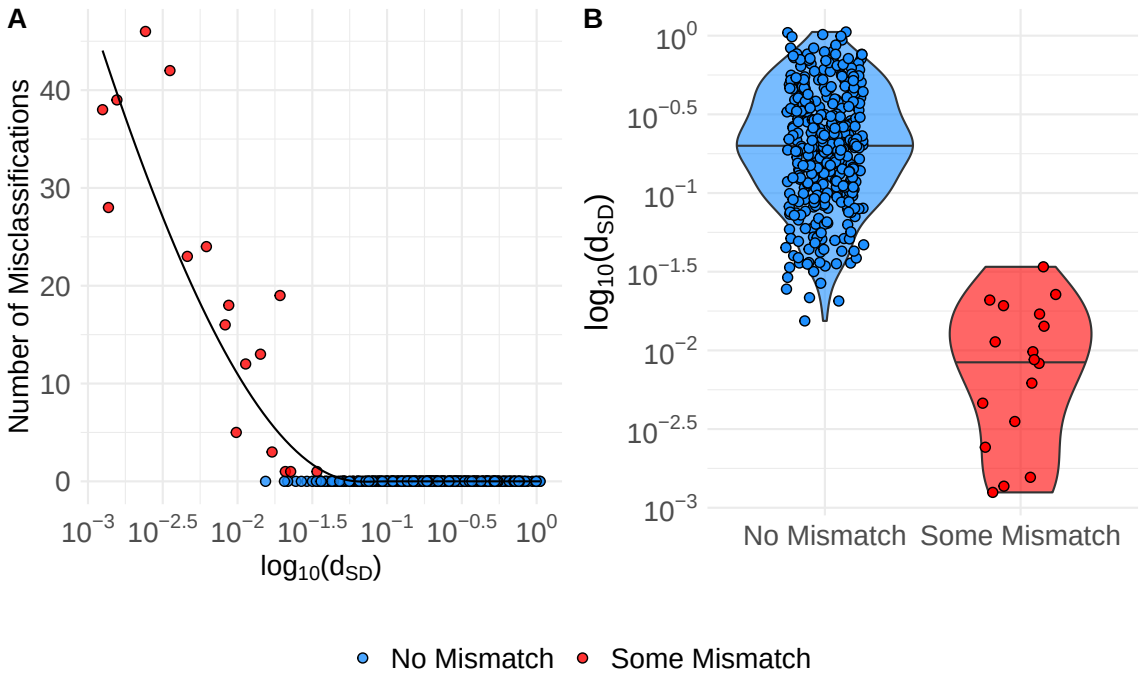

Figure S1: A) Each point presents the number of misclassification in subsampling posterior R_0 for each of the 400 simulated datasets. X-axis is the log transformed difference between W_S and W_D. B) Violin plots with jittered points of the difference between W_S and W_D for simulated alignments where there was wither some one no misclassification. Both A and B support that misclassication only occurs where the difference between W_S and W_D is negligible, and classification is not meaningful in the first instance.

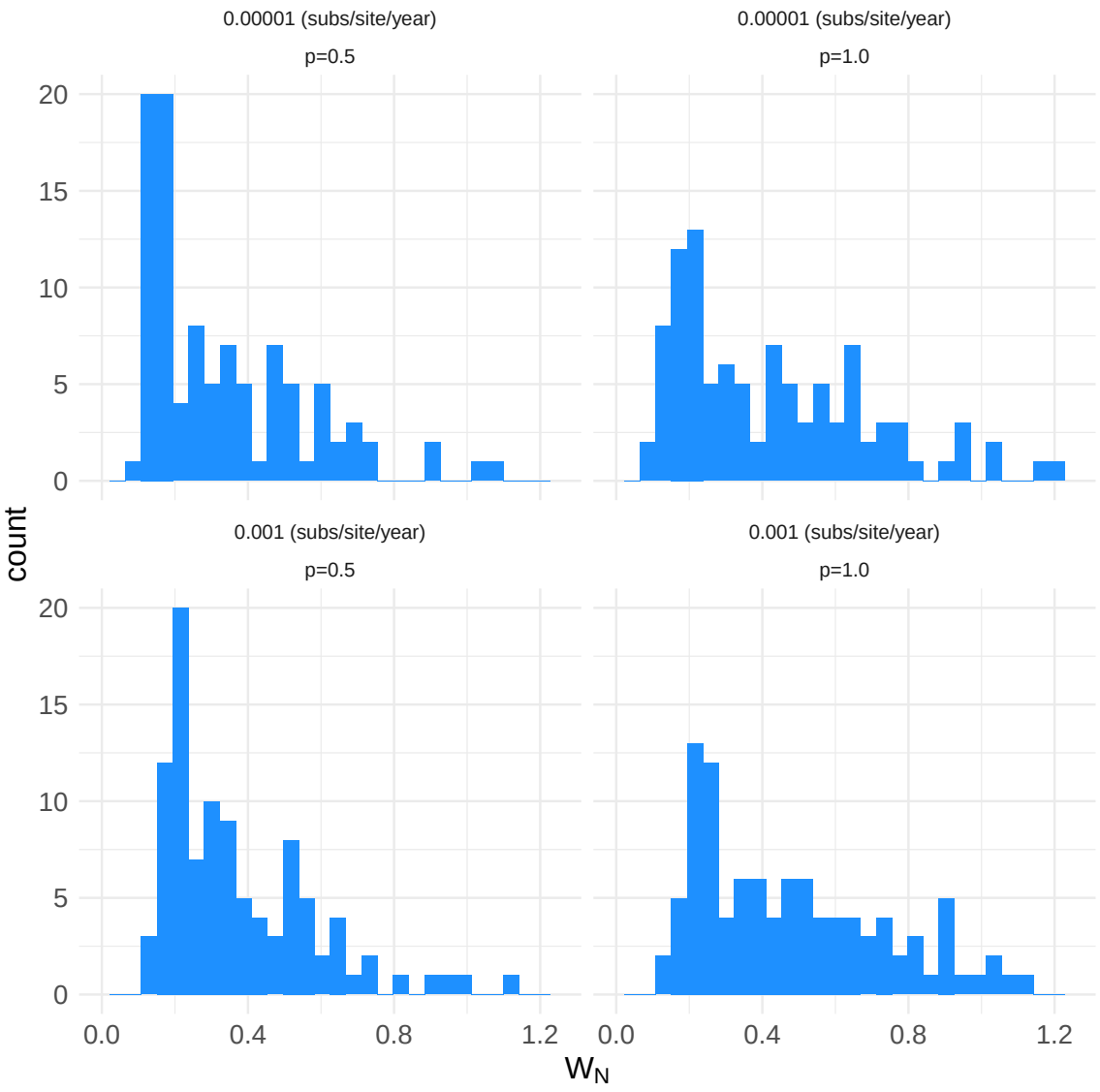

Figure S2: Histogram of W_N for each simulated dataset, separated by evolutionary rate and sampling proportion. W_N ranges from 0.1 to 1.2, such that simulated data provide additional information beyond that of the prior in each analysis.

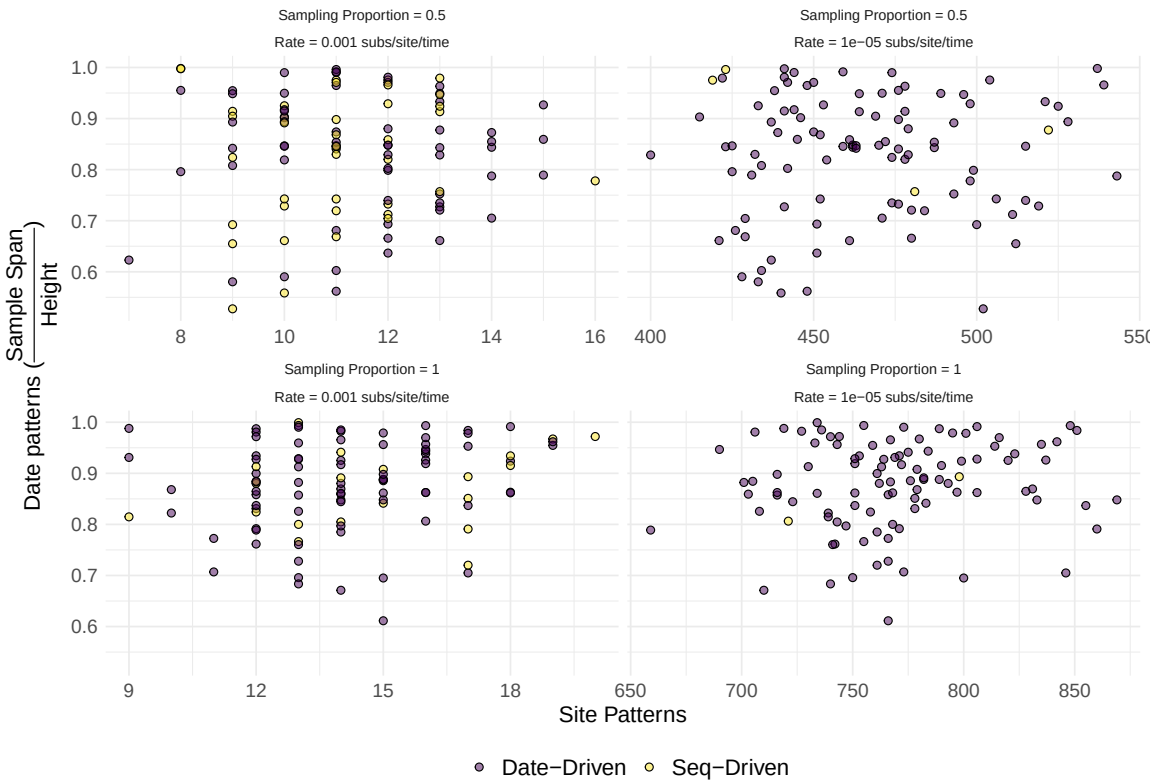

Figure S3: Date patterns ($= \frac{\text{Sampling Span}}{\text{Height}}$) against site patterns for each simulated dataset. Plots are separated by evolutionary rate and sampling proportion such that there are 400 points coloured by Wasserstein Classification across the entire figure. Higher evolutionary rates increase the proportion of Sequence driven datasets, but within each rate there is no clear pattern in date patterns, site patterns, or sampling rate driving classification.

Acknowledgements

LAF is grateful to Swissnex for awarding a Student Research Scholarship to foster this research. SD and LAF were received funding from the Australian Research Council (DE190100805), the Australian Medical Research Futures Fund (MRF9200006), and the National Health and Medical Research Council (APP1157586)

References

G. Baele, S. Dellicour, M. A. Suchard, P. Lemey, and B. Vrancken. Recent advances in computational phylodynamics. *Current opinion in virology*, 31:24–32, 2018.

V. Boskova, T. Stadler, and C. Magnus. The influence of phylodynamic model specifications on parameter estimates of the zika virus epidemic. *Virus evolution*, 4(1):vex044, 2018.

R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. D. Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. d. Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4):e1006650, Apr. 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006650. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006650`. Publisher: Public Library of Science.

L. du Plessis and T. Stadler. Getting to the root of epidemic spread with phylodynamic analysis of genomic data. *Trends in Microbiology*, 23(7):383–386, 2015.

L. A. Featherstone, F. Di Giallonardo, E. C. Holmes, T. G. Vaughan, and S. Duchêne. Infectious disease phylodynamics with occurrence data. *Methods in Ecology and Evolution*, 12(8): 1498–1507, 2021. doi: https://doi.org/10.1111/2041-210X.13620. URL `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13620`.

C. R. Lane, N. L. Sherry, A. F. Porter, S. Duchene, K. Horan, P. Andersson, M. Wilmot, A. Turner, S. Dougall, S. A. Johnson, et al. Genomics-informed responses in the elimination of covid-19 in victoria, australia: an observational, genomic epidemiological study. *The Lancet Public Health*, 6 (8):e547–e556, 2021.

S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, A. S. Azman, N. G. Reich, and J. Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172 (9):577–582, 2020. doi: 10.7326/M20-0504. URL `https://doi.org/10.7326/M20-0504`. PMID: 32150748.

S. Louca, A. McLaughlin, A. MacPherson, J. B. Joy, and M. W. Pennell. Fundamental identifiability limits in molecular epidemiology. *Molecular Biology and Evolution*, 38(9):4010–4024, 2021.

A. Rambaut and N. C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/13.3.235. URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/13.3.235`.

R. Sanjuán, M. R. Nebot, N. Chirico, L. M. Mansky, and R. Belshaw. Viral mutation rates. *Journal of virology*, 84(19):9733–9748, 2010.

T. Stadler. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267(3): 396–404, Dec. 2010. ISSN 00225193. doi: 10.1016/j.jtbi.2010.09.010. URL `https://linkinghub.elsevier.com/retrieve/pii/S0022519310004765`.

T. Stadler. *TreeSim: Simulating Phylogenetic Trees*, 2019. URL `https://CRAN.R-project.org/package=TreeSim`. R package version 2.4.

T. Stadler, R. Kouyos, V. von Wyl, S. Yerly, J. Böni, P. Bürgisser, T. Klimkait, B. Joos, P. Rieder, D. Xie, et al. Estimating the basic reproductive number from viral sequence data. *Molecular biology and evolution*, 29(1):347–357, 2012.

E. M. Volz and S. D. W. Frost. Sampling through time and phylodynamic inference with coalescent and birth–death models. *Journal of the Royal Society Interface*, 11(101), Dec. 2014. ISSN 1742-5689. doi: 10.1098/rsif.2014.0945. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4223917/`.

E. M. Volz, K. Koelle, and T. Bedford. Viral phylodynamics. *PLoS computational biology*, 9(3): e1002947, 2013.