Understanding the effects of date rounding in phylodynamics

Leo A. Featherstone[*,1], [Order TBA], Sebastian Duchene[†,1]

August 30, 2023

[1] Peter Doherty Institute for Infection and Immunity, University of Melbourne, Australia.
email: leo.featherstone@unimelb.edu.au

Abstract

Do at end, need to be Public-Healthy-y

Introduction

value of phylodynamics: Insight where epi falls short. Important role in the ebola, sars-cov-2. Transitioned to more at the forefront with finding VOCs etc. (cite volz)

Pathogen genomics plays an increasingly important role in our understanding of infectious outbreaks in recent decades. It offers insight across many scales of transmission, from the pandemic and epidemimec scales, such as for SARS-CoV-2 and Ebola virususes, to more localised transmission or bacterial pathogens (Lancet, 2021). Phylodynamic analysis has concurrently beed adopted as a key method to make temporal and spatial inference from pathogen genomes, particularly since the 2013-2016 West African Ebola outbreak (Mbala-Kingebeni *et al.*, 2019). It is most useful where temporal and spatial records of transmission are sparse, because it allows genomc information to be converted into each.

Phylodynamic methods draw a range of inferences from pathogen sequence data, including epidemioloigcal rates, spatial dynamics, and the time and location of origin (Attwood *et al.*, 2022, du Plessis and Stadler, 2015, Featherstone *et al.*, 2022). The basis of all such inferences is that epidemiological spread leaves a trace in the form of substitutions or other molecular information that can be used to reconstruct transmission processes. Pathogen populaions that meet this assumption known as 'measurably evolving populations' (Biek *et al.*, 2015, Drummond *et al.*, 2003). In accordance, phylodynamics always relies on a combination of genome sequence and associated sampling times to infer parameters of transmission in a temporally explicit way.

Ideal phylodynamic datasets should include precise sampling dates alongside genome sequences (Black *et al.*, 2020), but sampling times necessary carry over information about times of symptom onset, sampling, or otherwise accessing healthcare for individual patients. This can pose an unacceptable risk for patient confidentiality where sampling times can be used to identify individual patients. In some cases, sampling times or dates of admission that are available for purchase have allowed identification for a majority of patients on record (Shean and Greninger, 2018, Sweeney, 2013). In acknlowdgement of this risk, Shean and Greninger (2018) suggest that Expert Determination govern whether sampling times be released alingside genome sequences, and the resolution to which thay are disclosed (day, month, year). Essentially, This process involves an expert determining whether information is safe to release on a case-by-case basis.

From a phylodynamic point of view, sampling times with reduced resolution are usable. We consider sampling times with specificty less than to the day as having reduced date resolution. There are many examples of phylodynamic analysis conducted for a diverse array of viral and bacterial pathogens with reduced date resolution. These include viral pathogens such as Rabies Virus, Enterovirus, SARS-CoV-2, Dengue virus, and bacterial pathogens such as *Klebsiella pneumoniae*, *Streptococcus pneumoniae*, and *Mycobacterium tuberculosis* (**???????**). However, theoretical and applied results show sampling times can substantially direct inference (Featherstone *et al.*, 2021, 2023, Volz and Frost, 2014), raising the concern of whether reduced date resolution biases inference.

In this work, we seek to characterise when and where biases arise due to reduced date resolution. We also propose a simple form of encryption where date resolution is required, but poses an unacceptable risk to patient confidentiality. We show that redued date induces bias expecially where temporal resolution lost coincides with the averge time required for a mutation to arise in a given pathogen. We visualise the relationship between date resolution and average mutation time in Fig 1.

something about what workarounds exist already? Intervals satch sequence information? - Commonest workaround is to round dates: - examples Ideally empirical sequence data sets should include precise sampling time information for all samples, with the day, month, and year of collection (Black *et al.*, 2020). Nevertheless sequence sampling dates are often considered part of the associated metadata and may be unavailable or imprecise for different reasons, including patient or organisation confidentiality or a lack of a consistent platform for storing these data (Raza and Luheshi, 2016). The amount of genome sequences for SARS-CoV-2 in the GISAID database is about 15.8M (Shu and McCauley, 2017), with about 2.4% (382K) having 'incomplete' date information, where no sampling

time may be reported, or its precision may only include the year (verified early August 2023).

Including these sequences with imprecise sampling times in phylodynamic analyses requires the researcher to assume that they have been sampled at an arbitrary day. Selecting the arbitrary day can be motivated by convenience, for instance with all samples from 2020 being assigned 1st January 2020 or 15 June 2020, or by sampling a random day within 2020 using a statistical distribution. In any case, this practice introduces a degree of error. Indeed,sequences sampled 11 months apart may be assigned the exact same day.

In this work, we seek to chaaracterise where and whn bias arises, insign for how to deal with sensitive dates. The central hypothesis

Phylodynamic models that estimate epidemiological parameters such as the effective reproductive number R_e exploit substitutions and sequence sampling times (Featherstone *et al.*, 2023). As a consequence, pathogens for which the timescale of transmission coincides with the timescale over which they acquire substitutions are particularly well suited for phylodynamic analyses. As a case in point, H1N1 influenza virus accumulates substitutions at a rate of about 4 $\times 10^{-3}$ subs/site/year (Hedge *et al.*, 2013). Because its genome length is around 13,158 bp, we would expect 0.06 substitutions over the course of an infection ($\sim$ 4 days) and one substitution to appear every 90 days. Clearly, providing sampling times with a precision of a year would remove valuable information about the molecular and epidemiological dynamics, which occur over a timescale shorter than the unit of rounding and thus potentially introduce bias to phylodynamic estimates. In contrast, the evolutionary rate for TB has been estimated to be of the order of 10^{-8} subs/site/year (Menardo *et al.*, 2019), resulting in about 0.043 substitutions per year (for a genome length of 4.3 Mbp), one substitution every 23 years, and an expected 0.34 substitutions over the course of an infection (8 years) (Kühnert *et al.*, 2018). For this bacterium, providing the sampling year or month may be sufficiently precise to correctly inform phylodynamic analyses.

What we did, what we showed, encryption case. Here, we investigated the impact of different degrees of precision in sampling times on phylodyanmic estimates of key parameters, including R_e, the molecular clock rate, and the time of origin of the outbreak. We considered a range of pathogens, H1N1 influenza, SARS-CoV-2, *Shigella sonnei*, and TB. These organisms have undergone substantial genome surveillance and have different infectious periods and molecular evolutionary dynamics. To quantify the impact of date precision in phylodynamics, we conducted extensive simulations, where we are able to assess precision and accuracy in estimates of key parameters.

Importantly, although phylodyanmic analyses benefit from using precise sampling times, they are

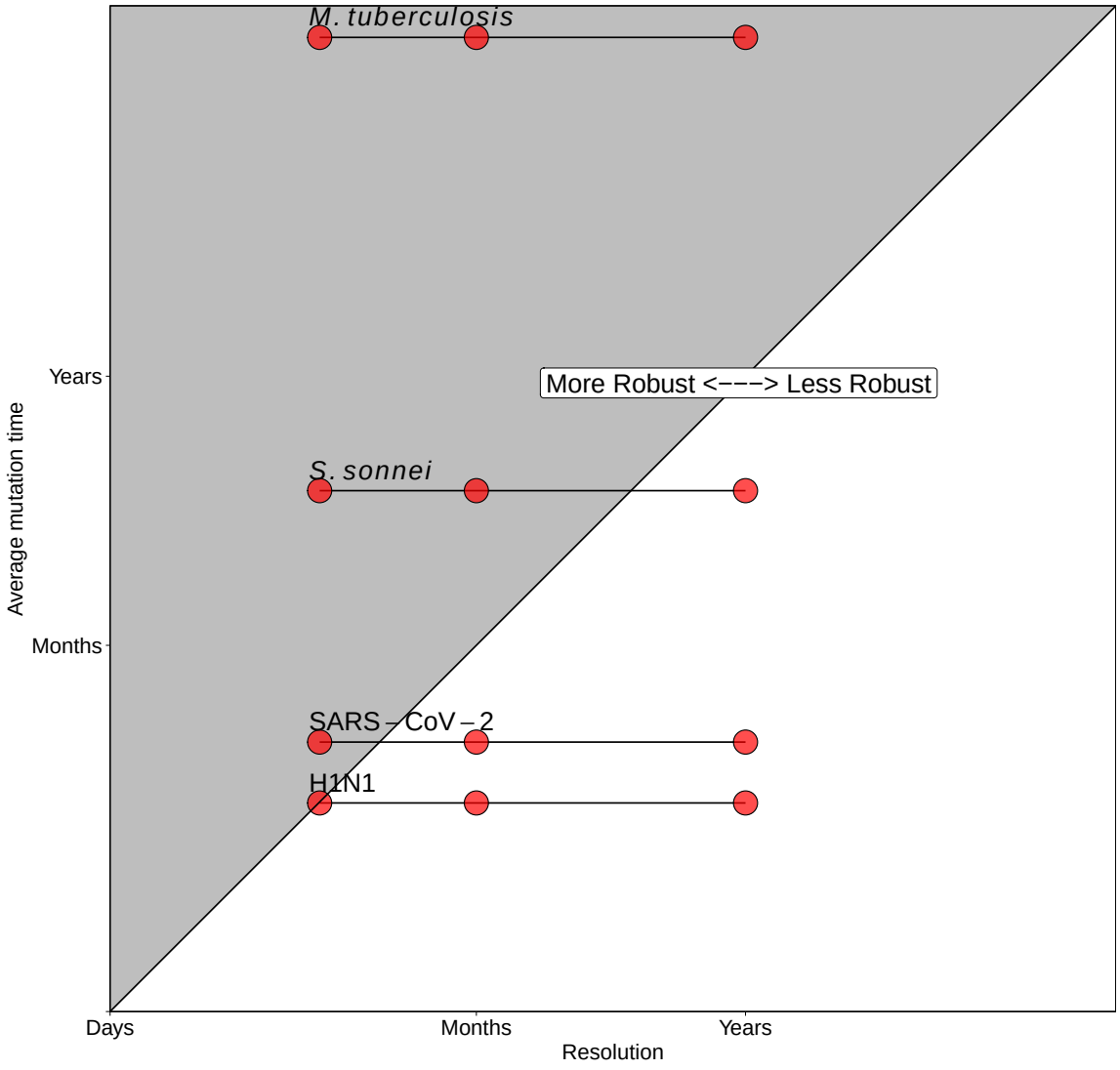

Figure 1: Add Hypothesis...

agnostic to the actual calendar date. An exponentially growing population sampled at a constant rate in the year 1997 will have the same distribution of sequence sampling times and coalescent events as one from the year 2020. Thus, a straight-forward approach to encrypt sampling dates is to provide their the number of days between sampling times, and not the actual calendar dates. - Not sure if this should be mentioned here.

Methods

Overview

Our study is based around 4 empirical datasets of H1N1 influenza virus, SARS-CoV-2, *Shigella sonnei*, and *Mycobacterium tuberculosis* and a corresponding simulation study. For both the empirical and simulated datasets, we performed phylodynamic analysis with sampling dates rounded to the day, month, year, and measure the resulting bias critical parameters - R_0 / R_e and the age of the outbreak (origin hereafter). For example, two samples from 2000/05/14 and 2000/05/02 would become 2000/05/01, if rounded to the month.

In this context R_0 refers to the *basic* reproductive number and R_e is the *effective* reproductive number. These parameters correspond to the average number of secondary infections at the start of an outbreak (i.e. assuming a fully susceptible population; R_0) or thereafter (R_e) (reviewed by (du Plessis and Stadler, 2015, Featherstone *et al.*, 2022, Kühnert *et al.*, 2011)). We consider the origin parameter as the age of the root node and not the length of the root branch to facilitate comparison between models, in contrast to Stadler Stadler *et al.* (2012), where the origin parameter includes the age of the root branch.

The two viral datasets consist of samples from the 2009 H1N1 pandemic (n=161) from Hedge *et al.* (2013), and a cluster of early SARS-CoV-2 cases from Australia in 2020 (n = 112) (Lane *et al.*, 2021). The bacterial datasets consist of Australian *S. sonnei* samples from an outbreak studied by Ingle *et al.* (2019), and 36 *M. tuberculosis* samples from a 25 year outbreak studied by Kühnert *et al.* (2018). These data were chosen because they encompass a diversity of epidemiological dynamics and scales with variable rates of substitution.

Simulation Study

We simulated outbreaks as Birth-Death sampling processes using the Master package in BEAST v2.6.6 (Bouckaert *et al.*, 2019, Vaughan and Drummond, 2013). These simulations consisted of 100 replicates over 4 parameter sets the represent values for each of the empirical datasets. All parameter sets include a proportion of cases sequenced (p), duration (T), and a "becoming uninfectious" rate (δ = reciprocal of the duration of infection). For simulations corresponding the viral datasets, transmission is modelled via R_0, the average number of secondary infections. For those corresponding to the bacterial datasets, we allow the effective reproductive numbers to change after an interval of time, R_{e_1} and R_{e_2}, with a change time at $0.5T$. This resulted in a total of 400 outbreak

datasets which we then used to simulate sequence data under a Jukes-Cantor model using Seq-Gen v1.3.4 (Rambaut and Grass, 1997). Substitution rates, genome lengths, and the above outbreak parameters are summarised in tables 1 and 2.

Table 1: Parameter sets outbreaks corresponding to each empircal dataset.

Microbe	$\delta(yrs)^{-1}$	R_0	R_{e_1}	R_{e_2}	p	T(yrs)	Source
H1N1	91.31	1.3	-	-	0.015	0.25	Hedge *et al.* (2013)
SARS-CoV-2	36.56	2.5	-	-	0.80	0.16	Lane *et al.* (2021)
Shigella sonnei	52.18	-	1.5	1.01	0.40	0.50	Ingle *et al.* (2019)
M. tuberculosis	0.125	-	2.0	1.10	0.08	25.0	Kühnert *et al.* (2018)

Table 2: Substitution rates and genome length for sequence simulation.

Microbe	Substitution Rate (subs/site/yr)	Genome Length	Time/Sub/Genome (yrs)
H1N1	4×10^{-3}	13158	0.0190
SARS-CoV-2	1×10^{-3}	29903	0.0334
S. sonnei	9×10^{-7}	4825265	0.3454
M. tuberculosis	1×10^{-7}	4300000	23.256

Empirical Data

We conducted Bayesian phylodynamic analyses were conducted using a Birth-Death skyline tree prior in BEAST v2.6.6 (Bouckaert *et al.*, 2019). We sampled from the posterior distribution using Markov chain Monte Carlo (MCMC), with length of 5×10^8 steps, with the initial 10% discarded as burnin. To determine sufficient sampling from the stationary distribution we verified that the effective sample size (ESS) of key parameters was at least 200.

H1N1

The H1N1 data consist of 161 samples from North America during the 2009 H1N1 influenza virus pandemic, analysed by Hedge *et al.* (2013). This dataset provides an example of a rapidly evolving pathogen sparsely sampled over a longer epidemiological timescale.

We placed a Lognormal($\mu = 0, \sigma = 1$) prior on R_0, $\beta(1,1)$ prior on p, and fixed the becoming-uninfectious ($\delta = 91$), corresponding to a 4 day duration of infection. We also placed an improper ($U(0,\infty)$) prior on the origin and a $U(10^{-4}, 10^{-2})$ prior on the substitution rate. This prior corresponds to analysis of these data in Featherstone *et al.* (2023).

SARS-CoV-2

The SARS-CoV-2 data are 112 samples from a densely sequenced transmission cluster in Victoria, Australia in 2020, first analysed by Lane *et al.* (2021). These data are similar to the H1N1 datasets in presenting a quickly evolving viral pathogen, but contrast in that virtually all cases in the cluster were sequenced.

Prior configurations are identical to those used in Featherstone *et al.* (2023) to analyse the same data. Briefly, we placed a

Lognormal(mean $= 1, \text{sd} = 1.25$) prior on R_0 and an Inv-Gamma($\alpha = 5.807, \beta = 346.020$) prior on the becoming-uninfectious rate (δ). The sampling proportion was fixed to $p = 0.8$ since every known Victorian SARS-CoV-2 case was sequenced at this stage of the pandemic, with a roughly 20% sequencing failure rate. We also placed an Exp(mean $= 0.019$) prior on the origin, corresponding to a lag of up to one week between the index case and the first putative transmission event. The substitution rate was fixed a 10^{-3} following (Duchene *et al.*, 2020).

Shigella sonnei

The *S. Sonnei* dataset originates from Ingle *et al.* (2019) and consists of a single nucleotide polymorphism (SNP) alignment of 146 sequenced isolates from infected men who have sex with men in Australia. These data provide an example of densely sequenced bacterial outbreak.

To accommodate changing transmission dynamics, we included two intervals for R_e with a Lognormal($\mu = 0, \sigma = 1$) prior on each. We also placed a $\beta(1, 1)$ prior on the sampling proportion, a $U(0, 1000)$ prior on the origin, and fixed the becoming un-infectious rate at $\delta = 73.05$ corresponding to a 5 day duration of infection.

To generate the SNP alignment, we (Enter Danielle...)

Mycobacterium tuberculosis

The *M. tuberculosis* dataset consists of 36 sequenced isolates taken from a retrospectively recognised outbreak in California, USA, and originating in the Wat Tham Krabok refugee camp in Thailand. We applied the same similar prior configuration to Kühnert *et al.* (2018), with the exception of including 2 intervals for R_e and fitting a strict molecular clock with a $\Gamma(\alpha = 0.001, \beta = 1000.0)$ prior.

Results

Simulation study

Broadly, the bias in posterior mean reproductive number increases with decreasing date resolution. This effect is most pronounced for the viral simulation conditions, where the rounding units of one month or one year is greater than the amount of time expected for one mutation to arise. In this case, date rounding condenses divergent sequences in time, driving a signal for higher rates of evolution and transmission. Conversely, the effect is less pronounced in the bacterial conditions where the date resolution lost is a smaller fraction of the effective mutation time, such as for the *M. tuberculosis* and *S.sonnei* data sets. In these cases, sequences are less divergent such that temporal clustering does not inflate posterior evolutionary rate. Moreover, the sampling timespans for these datasets are longer (table 1), meaning that clustering to month or year leads to a less pronounced inflation of the reproductive number as samples still remain temporally distributed.

Corresponding with the above trend in evolutionary rate, the mean posterior origin time have an upwards bias, representing a signal for a shorter outbreak duration (Fig S1). This is the result of a well understood axis among phylodynamic models where higher rates of evolution suggest shorter periods of evolution (Featherstone *et al.*, 2023). In the epidemiological view, this translates into placing more weight on lower values for the duration of the outbreak.

The H1N1 influenza virus simulation conditions demonstrate strongest relationship between high estimates of evolutionary rates and shorter outbreak duration. It can be thought of as the simulation conditions with the highest divergence among sequences relative to simulation time, owing to a combination of a higher mutation and transmission rate alongside a lower mutation rate (table 1). For the date rounding to the year, we see extremely high values of R_0 and substitution, with means of around 10^8 and 10^6 subs/site/year, respectively. Such values, although implausible, demonstrate a key point that bias in posterior estimates compounds with decreasing date resolution. The effect is nonlinear, but also exacerbated by more divergent sequences, which would otherwise make for an idea phylodynamic dataset (Featherstone *et al.*, 2023). Rounding to the month demonstrates intermediate effects with erroneously high bias. The SARS-Cov-2 simulation condition presents a similar trend, albeit with a less extreme degree of bias.

The two bacterial simulation conditions demonstrate the same trends in R_e, the substitution rate, and the origin. The *S. sonnei* dataset shows intermediate effects with minimal bias when moving to month resolution and larger effects at year levels for all the above parameters. This is expected

given its effective mutation time is somewhere between the order of months and years (table 1). This effect is also markedly increased for R_{e_2} in comparison to R_{e_1} at the year level, suggesting that bias also increases where more distinct samples appear to arise at the same time (we expect more samples in the second window of the *S. sonnei* simulations).

The *M. tuberculosis* simulation conditions effectively act as a control conditions, since it appears inter to date rounding. Again this is expected, because this dataset reflects both longer simulation time, with temporal clustering less likely to inflate R_e, but also the effective mutation time is longer than 1 year. As such, even rounding to a year is unlikely to drive a signal for increased evolutionary rate or a more recent origin time.

Empirical Results

	organism	resolution	meanR0	R0HPD	meanRe1	Re1HPD	meanRe2	Re2HP
1	H1N1	Day	1.083	[1.05, 1.11]		[NA, NA]		[NA, N
2	H1N1	Month	1.144	[1.11, 1.17]		[NA, NA]		[NA, N
3	H1N1	Year	1.154×10^8	[8.98×10^7, 1.45e+08]		[NA, NA]		[NA, N
4	SARS-CoV-2	Day	1.207	[0.919, 1.57]		[NA, NA]		[NA, N
5	SARS-CoV-2	Month	5.972	[3.84, 9.21]		[NA, NA]		[NA, N
6	SARS-CoV-2	Year	18.689	[10.5, 29.8]		[NA, NA]		[NA, N
7	Shigella	Day		[NA, NA]	1.072	[1.03, 1.11]	0.982	[0.968,
8	Shigella	Month		[NA, NA]	1.073	[1.03, 1.11]	0.983	[0.969,
9	Shigella	Year		[NA, NA]	1.174	[1.13, 1.22]	0.949	[0.933,
10	TB	Day		[NA, NA]	2.492	[0.688, 4.88]	1.292	[0.704,
11	TB	Month		[NA, NA]	2.789	[0.576, 5.15]	1.390	[0.735,
12	TB	Year		[NA, NA]	2.751	[0.5, 5.27]	1.484	[0.774,

Broadly, analyses of the empirical datasets reproduce the trends of bias in reproductive number, substitution rate, and time of origin from the simulation study (figures 4,5). That is, the reproductive number increases with decreasing date resolution along with an increase in the substitution rate and corresponding decrease in the origin. There are a few exceptions to this trend that we consider below and which we attribute to the difference between simulated and empirical sampling time distributions.

H1N1 influenza virus

Posterior R_0 increases with decreasing date resolution in a comparable way to the simulation study. However, the posterior substitution rate and origin time estimates remain essentially the same for day and month resolution (mean values of 1.083, 1.44 and 10^{-2}, 10^{-2} respectively)(table), before

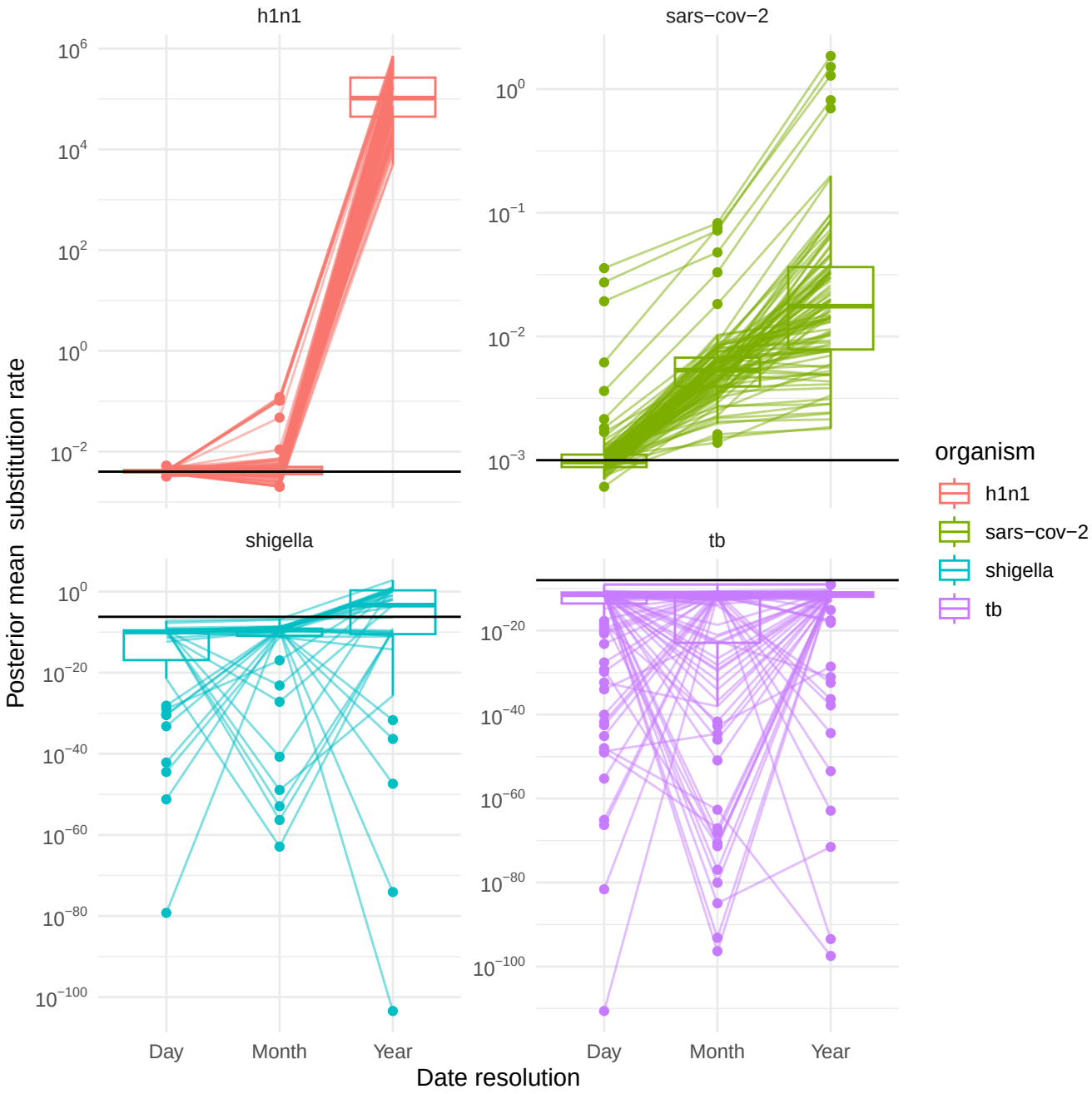

Figure 2: Mean posterior evolutionary rate for each simulation condition over decreasing date resolution. Lines connect individual simulated datasets across analyses with decreasing date resolution and horizontal black lines mark the true evolutionary rate. Mean posterior evolutionary rate increases where date rounding clusters more divergent sequences, such as in the case of the viral datasets. The effect is less pronounced for the slower evolving simulation conditions - (*S. sonnei* and *M. tuberculosis*).

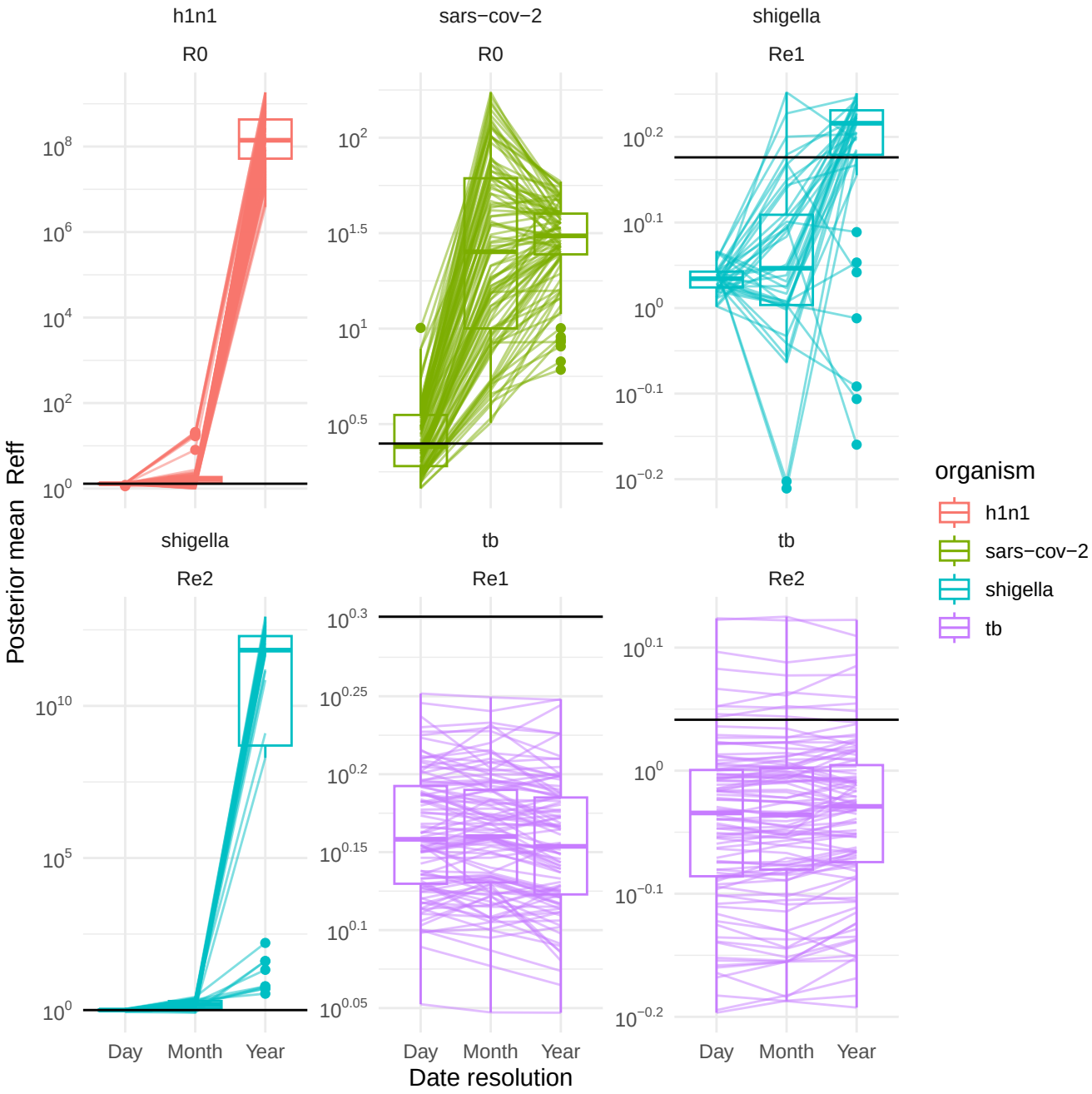

Figure 3: Bias in R_0 or R_e over decreasing date resolution for simulated data. Lines connect posterior mean reproductive number for individual simulated datasets analysed under decreasing date resolution under each simulation condition. Horizontal black lines show the true value. In general, the reproductive number biases upwares with decreasing date resolution, with the most dimished effects where the date resolution is a smaller fraction of average time required for a mutation (*S. sonnei* and *M. tuberculosis*).

moving upwards at year resolution as expected from the simulation study. This finding can be explained by the sampling time distribution, since the earlier samples occur later in their month (change fig3 to date axis proper), such that rounding them down to the first of the month effectively expands the timespan of sampling and thus driving signal for a lower evolutionary rate and older origin time.

SARS-CoV-2

The SARS-CoV-2 datasets behaves as expected with respect to the posterior R_0. In particular, rounding to the month results in an unlikely, but plausible value of $R_0 = 5.972$ (table). Rounding to the year inflates R_0 further as expected.

The posterior substitution rate of SARS-CoV-2 remains essentially stable when rounding to the month or year, with a mean value of $10^{-3.5}$ (subs/site/time) for both (table). In addition, the origin time at month resolution becomes older after rounding the sampling times. These findings stand in contrast to the expectation of rounding sampling times leading to an overestimate of the substitution rate and a corresponding underestimation of the time of origin. We again attribute these differences to the distribution of the empirical sampling times, which are not as consistently distributed as they are for the simulated outbreaks. There appears to be one early sample (Fig 4 A) that likely drives the signal for an older outbreak when rounding to the month because it is pushed back in time. At the same time, the substitution rate increases with decreasing date resolution as expected, likely due to the clustering of the the rest of the samples after the earliest.

S. sonnei

For R_{e_1}, the *S. sonnei* dataset matched the simulation study, with month rounding having a minimal effect, but year rounding inducing an upwards bias (mean values of 1.072, 1.073 respectively, figure 4). R_{e_2} departs from expectation. The estimate for this parameter decreases when rounding to the year. We speculate that this occurs because it compensates for elevated R_{e_1} earlier in the outbreak. This is supported by a markedly lower origin value (mean of 4.004, table), such that the outbreak appears as an intensified early burst. The substitution rate remains stable across date resolutions, which is expected given the overall low substitution in this data set (around 10^{-6} subs/site/year).

M. tuberculosis

The *M. tuberculosis* data recapitulate the outcome of the simulation study. Posterior origin times and evolutionary rate remain consistent across decreasing date resolution at 20 years and 10^{-7} (subs/site/time) respectively. We observe minimal upwards bias in posterior R_{e_1} and R_{e_2}, and the expectation that $R_{e_1} > R_{e_s}$ is met, coinciding with an earlier burst of transmission in agreement with Kühnert *et al.* (2018). This reaffirms that if the effective mutation time sufficiently large compared to the date resolution lost, then date rounding has a lesser effect.

Discussion

The results of the simulation study can by summarised as showing that date rounding inflates estimates of evolutionary and epidemiological rates by temporally clustering differentiated genome sequences. This factor manifested as upwards bias of the effective reproductive number, substitution rate, and an underestimate of the age of the outbreak. The extent of the bias increased with more diverged sequences and decreased date resolution. In other words, it increases with the assertion that more evolution occurred in less time. This is why bias increased for simulation conditions with the highest amount of mutation per unit time - the H1N1 influenza virus conditions followed SARS-CoV-2 and *S. sonnei*. Our *M. tuberculosis* simulations, in their inertness to date rounding, also support this explanation because they were unlikely to generate any mutations over the month or even year mutation timescales.

Our empirical analyses broadly recapitulated the results of the simulation study, but also introduced notable exceptions which emphasised the unpredictability of the magnitude and direction of estimation bias when rounding dates. For example, the posterior substitution rate of the H1N1 influenza virus dataset did not display an upward bias when rounding to the month. The SARS-CoV evolutionary rate did not increase when moving from month to year rounding, and the posterior R_{e_2} decreased when rounding to the year for the *S. sonnei* dataset. In each case, we attribute these differences to the way in which the distribution of empirical sampling times differed from the consistency our our simulations. This meant that date rounding did not always result in temporal clustering of divergent sequences. In the example of the SARS-CoV-2 dataset where samples originated from the end of one month and start of the next, rounding down to the start of each month serves to spread out the samples over time, overriding the effect of clustering samples from the same month.

Taken together, the results form the simulation study and empirical data show that although date rounding biases empidemiological estimates in a theoretically predictable direction, the intensity of the bias is difficult to predict and varies with the parameter space the data notionally inhabit. Moreover, features of real-world sampling such as fine-scale clustering of sampling times over longer sampling efforts can unpredictably dampen or reverse expected bias due to date rounding. Put succinctly, date rounding induces bias unpredicable bias due to the interaction of theoretical aspects of phylodynamic models real-world data features.

We conclude that accurate sampling time information is essential where phylodynamic insight is needed to understand infectious disease epidemiology and evolution. There does not appear to be an clear way to adjust for the bias caused otherwise. However, as acknowledged from the beginning of this article, it may impose an unacceptable level of risk to patient confidentiality to release highly precise isolate sampling times, as can theoretically be used to identify individual patients. To circumvent this and deliver timely phylodynamic results, we finish by proposing an extremely simple form of encryption that may lower the level of risk in sharing sampling time to the day to acceptable levels.

The simplest encryption of dates

The functional component of phylodynamic data is the *difference* between sequences and dates, rather than their absolute values. After all, our methods are comparative within a sample. Thus we can prioritise exact information and protect patient identity at the same time. We propose that data deposited in online databases include dates that are all shifted in time by an unknown seed number, and reinterpret results by factoring this in. For example, if the sampling times of a dataset of 3 samples are 2000, 2001 and 2002, then we may randomly draw a seed of 1000 with which to shift and dates deposit online 2000, 2001 and 2002 $\rightarrow$ 3000, 3001 and 3002. Then results can be reinterpreted with regard to the random seed. If, for example the estimated time of onset was 3 years before the most recent sample, then those receiving the data will not be able to place this in time, while those on the data generation end can interpret this correctly (estimated time of onset = 2002-3 = 1999). In the same vein, transmission parameters such as R_e can be understood to pertain to the true sampling time.

References

Attwood, S. W. *et al.* (2022). Phylogenetic and phylodynamic approaches to understanding and combating the early sars-cov-2 pandemic. *Nature Reviews Genetics*, **23**(9), 547–562.

Biek, R. *et al.* (2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*, **30**(6), 306–313. Publisher: Elsevier.

Black, A. *et al.* (2020). Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature medicine*, **26**(6), 832–841.

Bouckaert, R. *et al.* (2019). BEAST 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, **15**(4), e1006650. Publisher: Public Library of Science.

Drummond, A. J. *et al.* (2003). Measurably evolving populations. *Trends in ecology & evolution*, **18**(9), 481–488.

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

Duchene, S. *et al.* (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution*, **6**(2), veaa061.

Featherstone, L. A. *et al.* (2021). Infectious disease phylodynamics with occurrence data. *Methods in Ecology and Evolution*, **12**(8), 1498–1507. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13620.

Featherstone, L. A. *et al.* (2022). Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications. *Virus Evolution*, **8**(1), veac045.

Featherstone, L. A. *et al.* (2023). Decoding the Fundamental Drivers of Phylodynamic Inference. *Molecular Biology and Evolution*, **40**(6), msad132.

Hedge, J. *et al.* (2013). Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biology Letters*, **9**(5), 20130331.

Ingle, D. J. *et al.* (2019). Co-circulation of Multidrug-resistant Shigella Among Men Who Have Sex With Men in Australia. *Clinical Infectious Diseases*, **69**(9), 1535–1544.

Kühnert, D. *et al.* (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, genetics and evolution*, **11**(8), 1825–1841.

Kühnert, D. *et al.* (2018). Tuberculosis outbreak investigation using phylodynamic analysis. *Epidemics*, **25**, 47–53.

Lancet, T. (2021). Genomic sequencing in pandemics. *Lancet (London, England)*, **397**(10273), 445.

Lane, C. R. *et al.* (2021). 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.

Mbala-Kingebeni, P. *et al.* (2019). Medical countermeasures during the 2018 ebola virus disease outbreak in the north kivu and ituri provinces of the democratic republic of the congo: a rapid genomic assessment. *The Lancet infectious diseases*, **19**(6), 648–657.

Menardo, F. *et al.* (2019). The molecular clock of mycobacterium tuberculosis. *PLoS pathogens*, **15**(9), e1008067.

Rambaut, A. and Grass, N. C. (1997). Seq-gen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, **13**(3), 235–238.

Raza, S. and Luheshi, L. (2016). Big data or bust: realizing the microbial genomics revolution. *Microbial Genomics*, **2**(2).

Shean, R. C. and Greninger, A. L. (2018). Private collection: high correlation of sample collection and patient admission date in clinical microbiological testing complicates sharing of phylodynamic metadata. *Virus Evolution*, **4**(1), vey005.

Shu, Y. and McCauley, J. (2017). Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, **22**(13), 30494.

Stadler, T. *et al.* (2012). Estimating the basic reproductive number from viral sequence data. *Molecular biology and evolution*, **29**(1), 347–357.

Sweeney, L. (2013). Matching Known Patients to Health Records in Washington State Data. *SSRN Electronic Journal*.

Vaughan, T. G. and Drummond, A. J. (2013). A stochastic simulator of birth–death master equations with application to phylodynamics. *Molecular Biology and Evolution*, **30**(6), 1480–1493.

Volz, E. M. and Frost, S. D. W. (2014). Sampling through time and phylodynamic inference with coalescent and birth-death models. *Journal of the Royal Society, Interface*, **11**(101), 20140945.

Supplementary Material

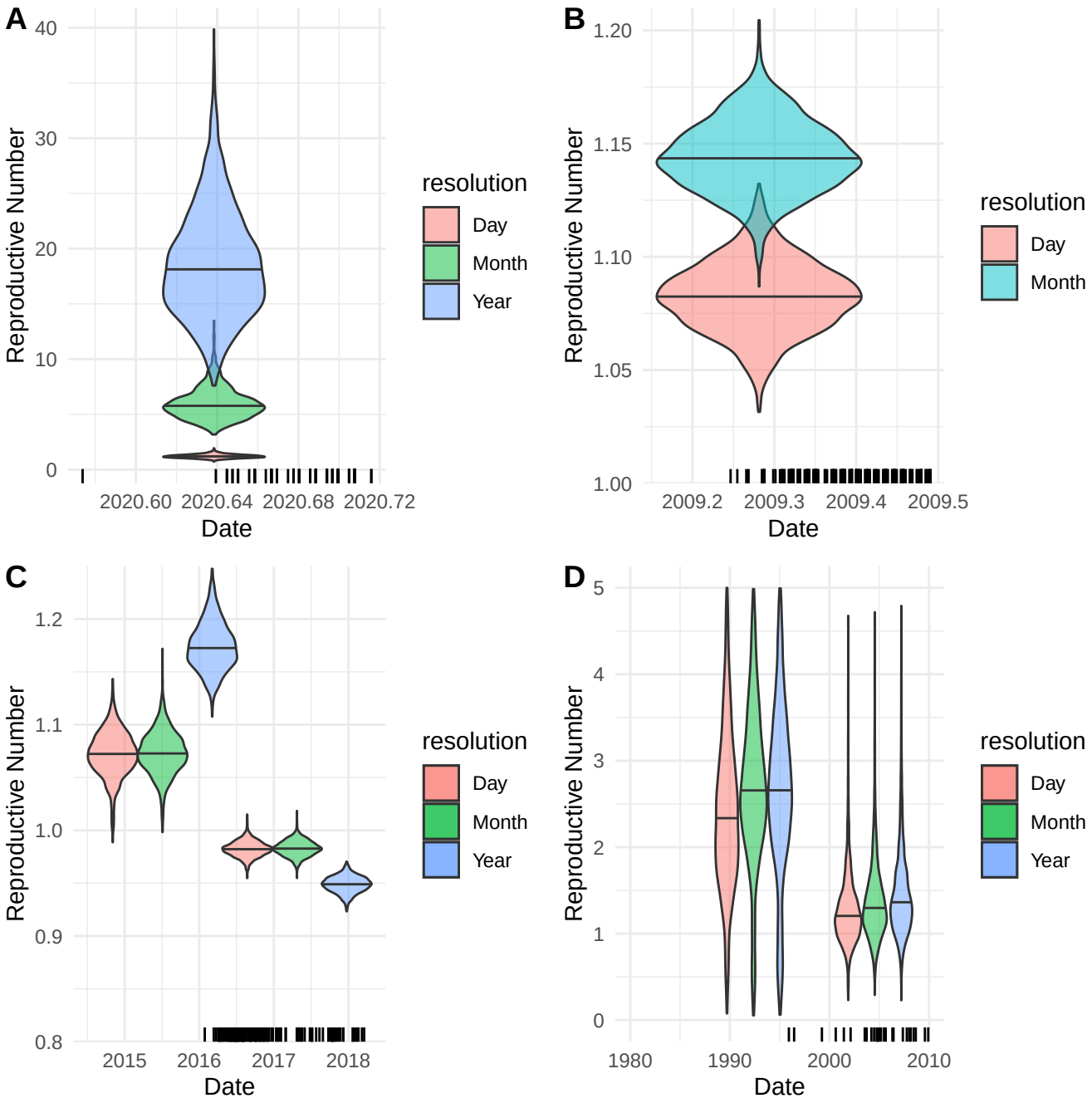

Figure 4: Posterior reproductive number and origin for each empirical dataset coloured by level of date resolution. Posterior origin times are represented as rescaled posterior frequencies along the Date axis and posterior reproductive numbers are given in violin plots on the vertical axis. For the H1N1 and SARS-CoV-2 datasets, posterior R_0 across date resolution is overlayed and overlaps minimally. For the *S. sonnei* and *M. tuberculosis*, posterior R_{e_1} and R_{e_2} (left to right) are displayed in adjacent groups. The change time between them is itself variable as half of the origin time. Samplng times are given as black mark son each date axis.

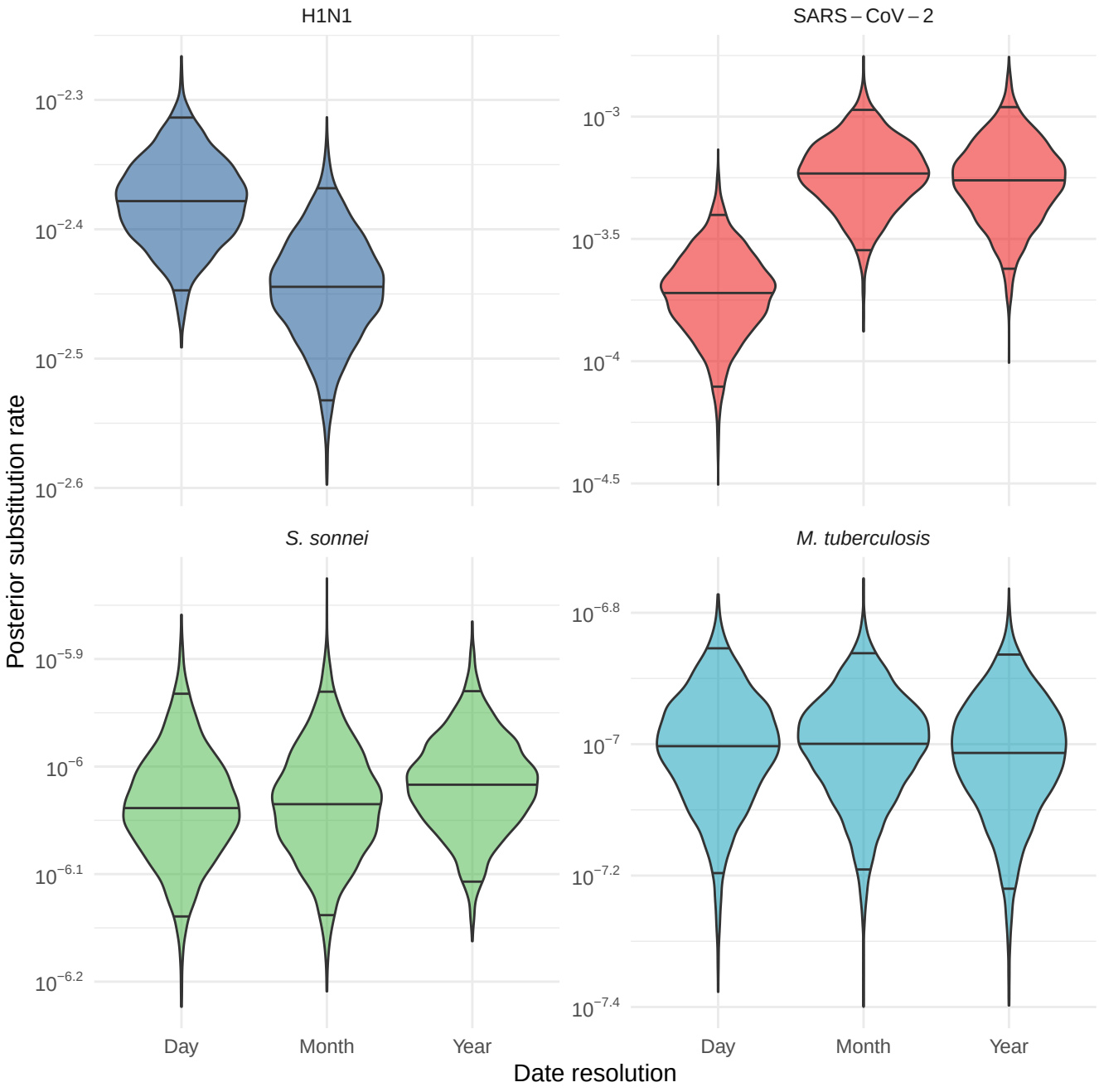

Figure 5: Posterior substitution rate for each empirical dataset across analyses with decreasing date resolution.

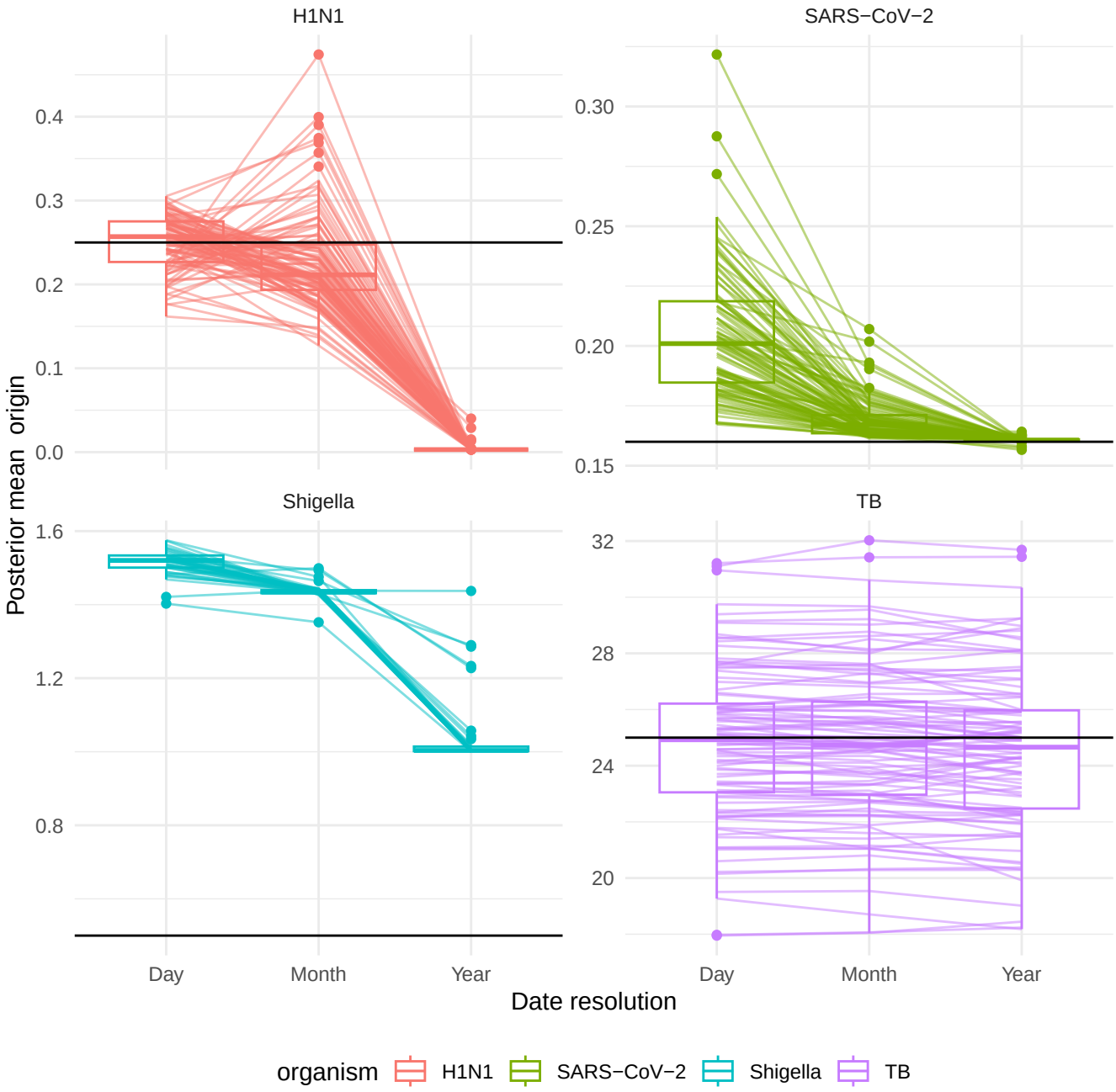

Figure S1: Meas posterior origin for each simulation condition over decreasing date resolution. Lines connect individual simulated datasets across analyses with decreasing date resolution and horizontal black lines mark the true evolutionary rate. Mean posterior origin decreases where date roundeding clusters more divergent sequences, such as in the case of the viral datasets. The effect is less pronounced for the slower evolving simulation conditions - (*S. sonnei* and *M. tuberculosis*).

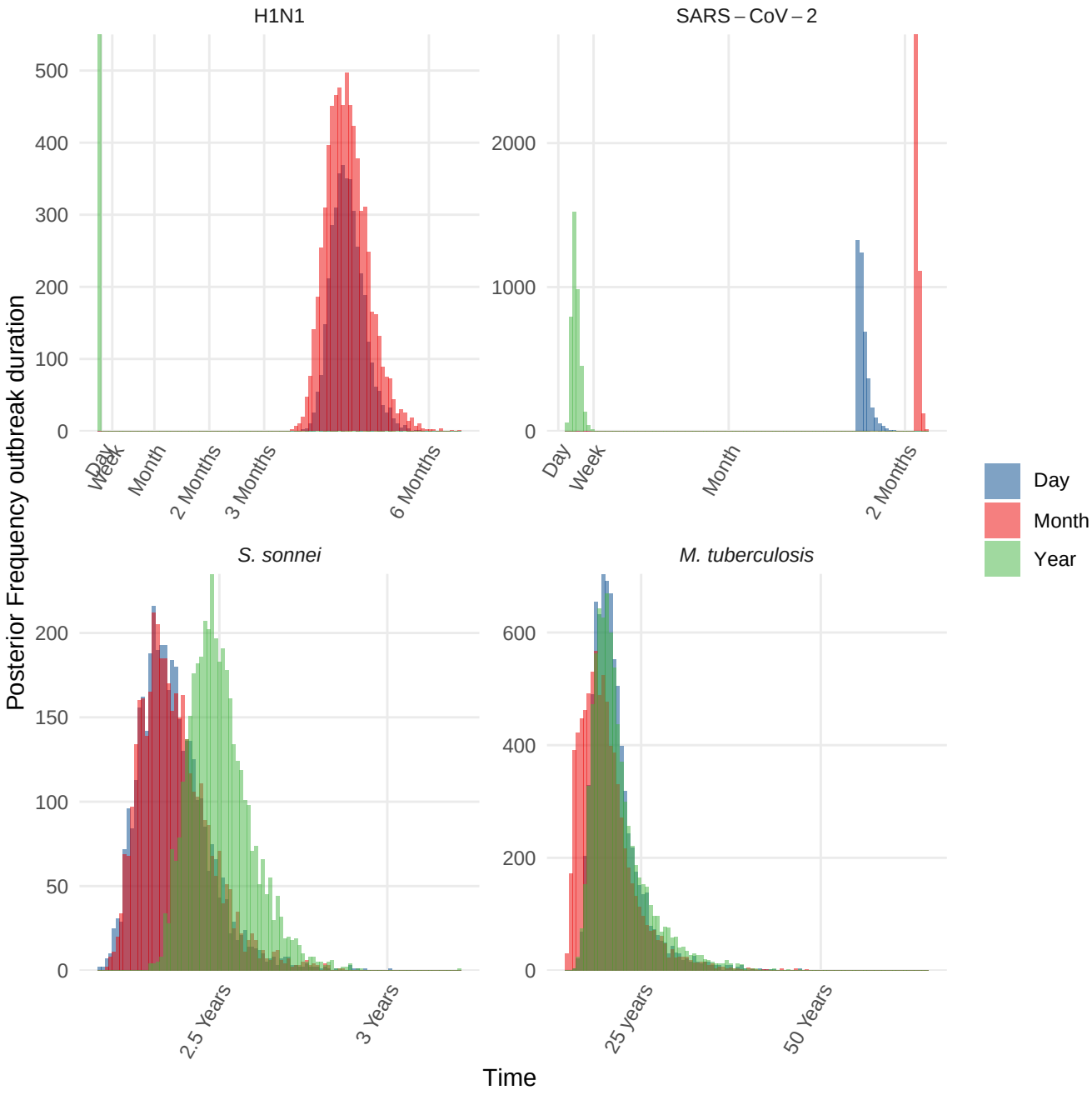

Figure S2: Posterior origin time across date resolution and simulation conditions.