Distribution Agreement

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this dissertation. I retain all ownership rights to the copyright of the dissertation. I also retain the right to use in future works (such as articles or books) all or part of this dissertation.

_____ _____

Yeongseon Park Date

Inferring transmission dynamics through patterns of genetic variation

By

Yeongseon Park
Doctor of Philosophy

Population Biology, Ecology and Evolution

Katharina Koelle, Ph.D.
Advisor

Anne Piantadosi, M.D., Ph.D.
Committee Member

Daniel Weissman, Ph.D.
Committee Member

Max Lau, Ph.D.
Committee Member

Rustom Antia, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Inferring transmission dynamics through patterns of genetic variation

By

Yeongseon Park
B.S., Ewha Womans University, 2016
M.S., Ewha Womans University, 2018

Advisor: Katharina Koelle, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Population Biology, Ecology and Evolution
2025

Abstract

Inferring transmission dynamics through patterns of genetic variation
By Yeongseon Park

For a rapidly evolving population, the pattern of genetic variation is shaped by evolutionary processes and population dynamics, providing a window to understand the underlying dynamics. In the context of infectious disease dynamics, a rapidly increasing number of pathogen genome sequences complements case-based inferences, allowing a better understanding of transmission dynamics. For more robust and reliable genome-based inference, this thesis attempts to further understand genome-based approaches, especially during the early spread of newly emerged viruses. In particular, this thesis focuses on considerations and challenges of phylodynamic inferences during the early spread of newly emerged viruses or variants. I first propose a novel approach to circumvent the phylogenetic uncertainty due to the low level of genetic variation during early spread. Then, I examine the misspecification of generation interval distribution for the early exponential growth phase. Next, I investigate the non-randomness in the dataset from the over-representation of epidemiological clusters, which could intensify when there are fewer sequences available, such as in early outbreaks. Lastly, I revisit the relationship between transmission trees and phylogenies by reviewing the inference approaches that infer transmission trees from phylogenetic trees. Together, this thesis aims to advance our understanding of important considerations in phylodynamic analyses and provides insights for improved implementation and interpretation of these methodologies.

Inferring transmission dynamics through patterns of genetic variation

By

Yeongseon Park
B.S., Ewha Womans University, 2016
M.S., Ewha Womans University, 2018

Advisor: Katharina Koelle, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Population Biology, Ecology and Evolution
2025

Acknowledgments

In Korea, there is a saying that it takes a whole village to raise a child. Likewise, there has been a whole community behind this dissertation, supporting me in countless ways throughout this journey. First, I would like to thank my advisor, Dr. Katia Koelle. I am incredibly grateful for the patience and trust she gave me throughout this process, which gave me the space to explore my ideas and supported my journey toward independence. Her guidance taught me how models can capture reality without unnecessary complexity. In addition, she encouraged and supported me in presenting my work at many conferences, giving me ample opportunities to connect with other researchers in the field and to learn the importance of sharing my work.

I am also deeply grateful to my committee members — Dr. Anne Piantadosi, Dr. Daniel Weissman, Dr. Max Lau, and Dr. Rustom Antia — for offering invaluable insights throughout this process, each from their own area of expertise. Beyond academic guidance, they also supported me in many other ways, including career advice and personal mentorship. Their support broadened my perspective in ways I will always carry with me. I would also like to thank my previous advisor, Dr. Yuseob Kim, who led me to this path. He introduced me to how biological systems can be studied through models and mathematics, and under his guidance, I learned how to look beyond mathematical expressions to understand their deeper meaning. Without him, I might have gone down a very different path.

I must also thank the Koelle Lab members — in particular, Amber Coats, Elizabeth Somerson, Ananya Saha, Mike Martin, Brent Allman, Diana Vera Cruz, Molly Gallagher, Jeremy Harris, Lisa Bono, and Dave VanInsberghe. I always felt welcome in the Koelle Lab, and they have been a great team to talk with about not only work but also life and the joys and challenges of graduate school.

I would also like to thank the PBEE community. I was especially lucky to have two cohorts — my admission cohort and my qualifying exam cohort. We supported each

other during the COVID-19 pandemic and got through the hard times of preparing for the qualifying exam together. I also thank EKSEA members for being both mentors and friends throughout this journey.

Finally, I would like to thank my parents, who have given me endless support throughout this journey. They have always encouraged me to follow my heart while providing a home where I could return and recharge. I also thank my sister, who has been the person closest to me both physically and emotionally during this time. Without her, this would not have been possible — or at least, it would have been so much harder. My little brother has also been a reliable source of laughter, and I would like to thank him too. There are others I could not mention here who have supported me and cheered me on, near or far. I would like to take this opportunity to express my heartfelt thanks to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The fate of genetic variation introduced into a population through mutation is shaped by various evolutionary processes. These processes leave signatures on patterns of genetic variation across the genome sequences of the population's individuals. By analyzing these patterns of genetic variation, we can infer the evolutionary processes that shaped the population. This pattern of genetic variation can further be used to reconstruct the evolutionary relationship between samples as a phylogeny. In a phylogeny, each sample is located at the tip of the tree, and common ancestors are represented at the inner nodes. The branch length indicates the evolutionary distances between individuals, calculated based on sequence evolution models.

For rapidly evolving populations, including RNA viruses, the time scale of these evolutionary processes can overlap with their population dynamic processes. This leads to potential interactions that leave detectable signatures in the pattern of genetic variation, which enables inference of their evolutionary and population dynamics history. If the accumulation of mutations can be significantly fast, it is considered as "measurably evolving." To address the temporal structure in the sampled sequences, samples from these measurably evolving populations are often analyzed using phylogeny. The field of phylodynamics aims to identify the processes responsible for generating

observed or reconstructed phylogenies, which represent the evolutionary relationships between sampled genome sequences.

If a population is measurably evolving and sampling times are known, we can calculate the rate of molecular evolution based on the 'molecular clock.' The molecular evolution rate of this clock enables the conversion of branch lengths in the reconstructed phylogeny into calendar time based on the sampling time. The resulting time-resolved tree provides insights into the timing of events, such as the branching of a lineage. In the context of infectious disease, these branching events and tip dates correspond to infection events and sampling times of infected individuals.

In the epidemiological context, a time-resolved phylogeny serves as a proxy for a partially observed transmission tree. Population dynamic models describe the mechanisms that generate transmission trees, and phylodynamic inferences rely on these models to estimate parameters of interest. Phylodynamic inference primarily relies on two major classes of population dynamic models: coalescent and birth-death models. Under coalescent models, trees are characterized as the coalescent events of existing lineages backward in time (Stadler et al., 2024). On the other hand, birth-death models characterize trees through the birth and death of lineages forward in time. Using these population dynamic models, phylodynamic approaches infer the parameters that govern the underlying dynamics of the sampled population.

In the context of infectious diseases, phylodynamics has been used to infer the transmission dynamics of infectious diseases. It has been used to infer key quantities in public health, including the reproduction number, which is the expected number of secondary infections from a single infected individual and epidemic growth rate. This makes genome sequences a valuable source of information that complements traditional case data for public health-related decision-making, such as planning and evaluating the nonpharmaceutical intervention or the effectiveness of the vaccination. As such, the importance of more robust and reliable genome-based inference is growing.

In this regard, this thesis aimed to further understanding of genome-based approaches for more robust and reliable genome-based inference. In particular, this thesis will focus on considerations and challenges of phylodynamic inferences during the early spread of newly emerged viruses or variants.

In the context of infectious diseases, phylodynamics has been widely used to infer transmission dynamics and estimate key public health quantities, such as the reproduction number, which is the expected number of secondary infections caused by a single infected individual, and the epidemic growth rate. This makes viral genome sequences a valuable source of information that complements traditional case data for public health decision-making, including the planning and evaluation of nonpharmaceutical interventions and vaccination strategies. As the role of genomic data is gaining importance, robust and reliable genome-based inference becomes increasingly important. In this context, this thesis contributes to advancing genome-based phylodynamic inference by examining current methods and proposing a new approach to address key challenges during early viral spread. Specifically, it focuses on the considerations and challenges involved in conducting phylodynamic inference during the early spread of newly emerged viruses or variants.

Chapter 2 focuses on the low level of genetic variation in the samples. Since the reconstruction of phylogenies relies on genetic variation in the sampled genome sequences, insufficient genetic diversity can lead to uncertainty in phylogenetic reconstruction (Boskova et al., 2018), manifesting as unresolved relationships between samples (Maddison, 1989). This concern is particularly relevant for pathogens with relatively low mutation rates, such as SARS-CoV-2 (Markov et al., 2023). Consequently, especially during early spread, integrating phylogenetic uncertainty becomes crucial. This integration can be achieved through Bayesian approaches that explore the tree space rather than considering a single given tree (Boskova et al., 2018; Park et al., 2023). However, such integration becomes computationally intensive under substantial

uncertainty (Park et al., 2023). Moreover, even when phylogenetic uncertainty is integrated, low levels of genetic variation may still yield unreliable phylodynamic estimates (Lam and Duchene, 2021). The impact of low genetic diversity is more pronounced in inferences based on coalescent models, as they do not utilize sequence data sampling time as an additional information source (Boskova et al., 2018). Furthermore, the reconstruction of phylogenies relies on models describing sequence evolution, and these additional parameters must be estimated alongside epidemiological parameters from limited sequence data (Park et al., 2023). To circumvent the need to reconstruct a phylogenetic tree, **Chapter 2** presents a novel tree-free approach to infer epidemiological parameters using the time-series of the number of segregating sites.

In **Chapter 3**, systematic biases introduced by violation of model assumptions are investigated. As in any inference approach, violation of assumptions can introduce biases in phylodynamic analyses. A well-known example is the misspecification of the sampling process in the birth-death model (Volz and Frost, 2014). However, even when it is not explicitly stated in the model, there are assumptions that are shaped by model components. These implicit assumptions may also introduce bias when misspecified. **Chapter 3** focuses on the implicit assumptions in tree models regarding the generation interval distribution and investigates the impact of generation interval misspecification on phylodynamic analyses.

Next, **Chapter 4** focuses on the non-randomness in sequence dataset in phylodynamic analyses. The importance of addressing sampling effort in phylodynamic inferences was pointed out early by Frost et al. (2015). Although representative sampling is ideal, real-world sampling varies spatially and temporally. Additionally, sequence data may include epidemiologically clustered sequences. In particular, focusing on the dataset with epidemiologically clustered sequences, **Chapter 4** examines the bias introduced by the clustered sequences and evaluates summary statistics that could be used to identify the non-randomness in sequence datasets.

Finally, **Chapter 5** revisits the relationship between the phylogeny and transmission tree. Often, the phylodynamic inferences consider phylogeny to be a proxy for the partially observed transmission tree. Yet, these two structures are conceptually distinct, as recognized in earlier work. Focusing on the scale of 'who-infected-whom,' **Chapter 5** reviews historical approaches to inferring transmission histories from phylogenies. It then analyzes existing methods based on their underlying assumptions, data requirements, and methodologies. The chapter concludes by suggesting that the choice of approach should be guided by available data characteristics and that systematic comparisons are needed to better inform users.

Chapter 1 References

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), Jan. 2018. ISSN 2057-1577. doi: 10.1093/ve/vex044. URL http://dx.doi.org/10.1093/ve/vex044.

S. D. Frost, O. G. Pybus, J. R. Gog, C. Viboud, S. Bonhoeffer, and T. Bedford. Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92, Mar. 2015. ISSN 17554365. doi: 10.1016/j.epidem.2014.09.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1755436514000437.

A. Lam and S. Duchene. The Impacts of Low Diversity Sequence Data on Phylodynamic Inference during an Emerging Epidemic. *Viruses*, 13(1):79, Jan. 2021. ISSN 1999-4915. doi: 10.3390/v13010079. URL http://dx.doi.org/10.3390/v13010079.

W. Maddison. Reconstructing Character Evolution On Polytomous Cladograms. *Cladistics*, 5(4):365–377, Dec. 1989. ISSN 1096-0031. doi: 10.1111/j.1096-0031.1989.tb00569.x. URL http://dx.doi.org/10.1111/j.1096-0031.1989.tb00569.x.

P. V. Markov, M. Ghafari, M. Beer, K. Lythgoe, P. Simmonds, N. I. Stilianakis, and A. Katzourakis. The evolution of SARS-CoV-2. *Nature Reviews Microbiology*, 21 (6):361–379, Apr. 2023. ISSN 1740-1534. doi: 10.1038/s41579-023-00878-2. URL http://dx.doi.org/10.1038/s41579-023-00878-2.

Y. Park, M. A. Martin, and K. Koelle. Epidemiological inference for emerging viruses using segregating sites. *Nature Communications*, 14(1):3105, 2023.

T. Stadler, C. Magnus, T. G. Vaughan, J. Barido-Sottani, V. Bošková, J. S. Huisman, and J. Pečerska. *Decoding genomes: from sequences to phylodynamics*. publisher not identified, United States, first edition edition, 2024. ISBN 9798324659998.

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):20140945, Dec. 2014. ISSN 1742-5662. doi: 10.1098/rsif.2014.0945. URL `http://dx.doi.org/10.1098/rsif.2014.0945`.

Chapter 2

Epidemiological inference for emerging viruses using segregating sites

The following *Article* was published in *Nature Communications* in May 2023. Our goal was to develop a novel tree-free inference approach to estimate key epidemiological parameters from viral sequence data. We show that the trajectory of a number of segregating sites, which summarizes the level of genetic variation over time, is informative of the underlying epidemiological dynamics. We then developed an inference approach to infer the basic reproduction number and the timing of the index case based on the trajectory of the segregating sites. We verified the approach using simulated datasets and applied it to the SARS-CoV-2 sequences sampled from France from late 2019 to early 2020.

2.1 Contribution to the published work

Conceptualization, Methodology, Software, Validation, Formal analysis (Figure 1-3A, 4-5A, 7-9, Figure S1-S7, S8A-B, S10-S12), Investigation (Figure 1-3A, 4-5A, 7-9, Figure

S1-S7, S8A-B, S10-S12), Writing, Visualization (Figure 1-3A, 4-5A, 7-9, Figure S1-S7, S8A-B, S10-S12).

2.2 Published manuscript

nature communications

Article

https://doi.org/10.1038/s41467-023-38809-7

Epidemiological inference for emerging viruses using segregating sites

Received: 8 September 2022

Accepted: 16 May 2023

Published online: 29 May 2023

Check for updates

Yeongseon Park [1], Michael A. Martin [1,4] & Katia Koelle [2,3]

Epidemiological models are commonly fit to case and pathogen sequence data to estimate parameters and to infer unobserved disease dynamics. Here, we present an inference approach based on sequence data that is well suited for model fitting early on during the expansion of a viral lineage. Our approach relies on a trajectory of segregating sites to infer epidemiological parameters within a Sequential Monte Carlo framework. Using simulated data, we first show that our approach accurately recovers key epidemiological quantities under a single-introduction scenario. We then apply our approach to SARS-CoV-2 sequence data from France, estimating a basic reproduction number of approximately 2.3-2.7 under an epidemiological model that allows for multiple introductions. Our approach presented here indicates that inference approaches that rely on simple population genetic summary statistics can be informative of epidemiological parameters and can be used for reconstructing infectious disease dynamics during the early expansion of a viral lineage.

Phylodynamic inference methods use pathogen sequence data to estimate epidemiological quantities such as the basic reproduction number and to reconstruct epidemiological patterns of incidence and prevalence. These inference methods have been applied to sequence data across a broad range of RNA viruses, including HIV[1–4], Ebola virus[5–7], dengue viruses[8], influenza viruses[9], and most recently severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[10–12]. Most commonly, phylodynamic inference methods rely on underlying coalescent models or birth-death models. Coalescent-based approaches have been generalized to accommodate time-varying population sizes and structured epidemiological models, for example, susceptible-exposed-infected-recovered (SEIR) models and models with spatial subdivision[6,13]. Birth-death approaches[14,15], where a birth in the context of infectious diseases corresponds to a new infection and death corresponds to a recovery from infection, carry advantages such as capturing the role of demographic stochasticity in disease dynamics, which may be particularly important in emerging diseases that start with low infection numbers[16]. Birth-death approaches have also been expanded to incorporate the complex nature of infectious disease dynamics including structured populations[17]. Both coalescent-based and birth-death phylodynamic inference approaches rely on time-resolved phylogenies and have been incorporated into the phylogenetics software packages BEAST1[18] and BEAST2[19] to allow for joint estimation of epidemiological parameters and dynamics while integrating over phylogenetic uncertainty[6,20]. Integrating over phylogenetic uncertainty is crucial when applying these methods to viral sequence data that are sampled over a short period of time and contain only low levels of genetic diversity. However, integrating over phylogenetic uncertainty can be computationally intensive. Moreover, phylodynamic approaches that use reconstructed trees for inference require estimation of parameters associated with models of sequence evolution, along with parameters that are of more immediate epidemiological interest.

Here, we present an alternative sequence-based statistical inference method that may be particularly useful when viral sequences are sampled over short time periods and when phylogenetic uncertainty present in time-resolved viral phylogenies is considerable. Instead of relying on viral phylogenies to infer epidemiological parameters or to reconstruct patterns of viral spread, the "tree-free" method we propose here fits epidemiological models to time series of the number of segregating sites (that is, the number of polymorphic sites) present in a sampled viral population. The approach we propose here allows for

[1]Graduate Program in Population Biology, Ecology, and Evolution, Emory University, Atlanta, GA 30322, USA. [2]Department of Biology, Emory University, Atlanta, GA 30322, USA. [3]Emory Center of Excellence for Influenza Research and Response (CEIRR), Atlanta, GA, USA. [4]Present address: Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. e-mail: katia.koelle@emory.edu

structured infectious disease models to be considered in a straightforward "plug-and-play" manner. It also incorporates the effect that demographic noise has on epidemiological dynamics. Below, we first describe how segregating site trajectories are calculated using sequence data and how they are impacted by sampling effort, rates of viral spread, and transmission heterogeneity. We then describe our proposed statistical inference method and apply it to simulated data to demonstrate the ability of this method to infer epidemiological parameters and to reconstruct unobserved epidemiological dynamics. Finally, we apply our segregating sites method to SARS-CoV-2 sequence data from France, arriving at quantitatively similar parameter estimates to those arrived at using epidemiological data.

Results

Segregating site trajectories are informative of epidemiological dynamics

The number of segregating sites present in a set of sampled viral sequences is defined as the number of nucleotide sites at which genetic variation is present in the sample set. To determine whether the number of segregating sites that are observed over time in a viral population may be informative of underlying epidemiological dynamics, we forward-simulated a classic susceptible-exposed-infected-recovered (SEIR) epidemiological model, augmented with viral evolution, under various sampling efforts and parameterizations (Fig. 1; Methods). Simulations of this augmented SEIR model, initialized with a single infected individual, first indicate that segregating site trajectories are sensitive to sampling effort, as expected (Fig. 1a, b). More specifically, we considered three different sampling strategies, each with sequences binned in consecutive, nonoverlapping 4-day time windows to calculate segregating site trajectories. These three sampling strategies consisted of a strategy with full sampling effort (all sequences per 4-day time window), one with dense sampling effort (40 sequences per 4-day time window) and one with sparse sampling effort (20 sequences per 4-day time window). With all three of these sampling efforts, the number of segregating sites first increases as the epidemic grows, with mutations accumulating in the virus population. Following the peak of the epidemic, the number of segregating sites starts to decline as viral sublineages die out, reducing the amount of genetic variation present in the viral population. A comparison between full, dense, and sparse sampling efforts indicates that lowering sampling effort results in a lower number of observed segregating sites during any time window. This is because at lower sampling effort, less of the genetic variation present in a viral population over a given time window is likely to be sampled. The patterns shown here across sampling strategies are robust to the time window length used for the calculation of segregating site trajectories (Figure S1).

To assess whether segregating site trajectories could be used for statistical inference, we first considered whether these trajectories differed between epidemics governed by different basic reproduction numbers (R_0 values). Figure 1c shows simulations of the SEIR model under two parameterizations of the basic reproduction number: an R_0 of 1.6, corresponding to the simulation shown in Fig. 1a, and a higher R_0 of 2.0 (implemented via a higher transmission rate β). The epidemic with the higher R_0 expanded more rapidly (Fig. 1c) and, under the same sampling effort, resulted in a more rapid increase in the number of segregating sites (Fig. 1d). This indicates that segregating site trajectories can be informative of R_0 early on in an epidemic.

We next considered the effect of transmission heterogeneity on segregating site trajectories. Many viral pathogens are characterized by 'superspreading' dynamics, where a relatively small proportion of infected individuals are responsible for a large proportion of secondary infections[21]. The extent of transmission heterogeneity is often gauged relative to the 20/80 rule (where the most infectious 20% of infected individuals are responsible for 80% of the secondary cases[22]).

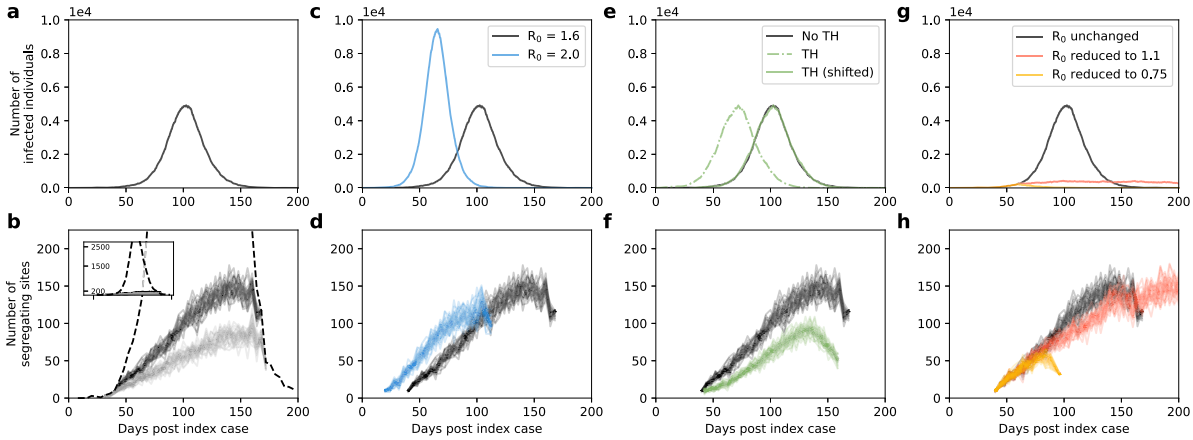

Fig. 1 | Segregating site trajectories under simulated epidemiological dynamics. a Dynamics of infected individuals (I) under an SEIR model simulated with an R_0 of 1.6. **b** Segregating site trajectories under full (black dashed line), dense (black lines), and sparse (gray lines) sampling efforts. Dense and sparse sampling correspond to 40 and 20 sequences sampled per time window, respectively. **c** Simulated infected dynamics under the SEIR model with an R_0 of 2.0 (blue line) compared to those of the $R_0 = 1.6$ simulation (black line). **d** Segregating site trajectories for the two simulations shown in panel **c**. **e** Simulated infected dynamics under the SEIR model with transmission heterogeneity (green, dashed line) compared to those of the $R_0 = 1.6$ simulation (black line) without transmission heterogeneity. Transmission heterogeneity was included by setting the parameter p_h to 0.06. For ease of comparing segregating site trajectories, the transmission heterogeneity simulation was shifted later in time (green, solid line). **f** Segregating site trajectories for the shifted transmission heterogeneity simulation (green lines) and the original simulation (black lines). **g** Simulated infected dynamics under the SEIR model with changing R_0. In the simulations shown in red and yellow, when the number of infected individuals reached 400, R_0 was decreased to 1.1 and 0.75, respectively. The simulation in black has R_0 remaining at 1.6. **h** Segregating site trajectories for the three simulations shown in panel g. Dense sampling effort was used to generate all segregating site trajectories shown in panels **d, f**, and **h**. 30 randomly-sampled segregating site trajectories are shown for each sampling effort in panel b and for each epidemiological scenario in panels **d**, **f,** and **h.** In all model simulations, $\gamma_E = 1/2$ days^{-1}, $\gamma_I = 1/3$ days^{-1}, population size $N = 10^5$, and the per genome, per transmission mutation rate $\mu = 0.2$. Initial conditions are $S(t_0) = N\text{-}1$, $E(t_0) = 0$, $I(t_0) = 1$, and $R(t_0) = 0$. For the transmission heterogeneity simulation (panel **e**), $I_h(t_0) = 1$ and $I_l(t_0) = 0$ was used instead of $I(t_0) = 1$. A time step of $\tau = 0.1$ days was used in the Gillespie τ -leap algorithm.

Some pathogens like SARS-CoV-2 exhibit extreme levels of superspreading, with as low as 10-15% of infected individuals responsible for 80% of secondary cases[10,23–25]. Because transmission heterogeneity is known to impact patterns of viral genetic diversity[26], we simulated the above SEIR model with transmission heterogeneity to ascertain its effects on segregating site trajectories (Methods). Because transmission heterogeneity has a negligible impact on epidemiological dynamics once the number of infected individuals is large[27], epidemiological dynamics with and without transmission heterogeneity should be quantitatively similar to one another, with transmission heterogeneity simply expected to shorten the timing of epidemic onset in simulations with successful invasion[21]. Our simulations, parameterized with extreme transmission heterogeneity of 6/80, confirm this pattern (Fig. 1e). To compare segregating site trajectories between these simulations, we therefore shifted the simulation with transmission heterogeneity later in time such that the two simulated epidemics peaked at similar times (Fig. 1e). Comparisons of segregating site trajectories between these simulations indicated that transmission heterogeneity decreased the number of segregating sites during every time window (Fig. 1f). As expected, lower levels of transmission heterogeneity result in less substantial decreases in the number of segregating sites (Figure S2). Together, these results indicate that transmission heterogeneity needs to be taken into consideration when estimating epidemiological parameters using segregating site trajectories.

Finally, we wanted to assess whether changes in R_0 over the course of an epidemic would leave signatures in segregating site trajectories. We considered this scenario because phylodynamic inference has often been used to quantify the effect of public health interventions on R_0, most recently in the context of SARS-CoV-2[10,11]. We thus implemented simulations with R_0 starting at 1.6 and then either remaining at 1.6 or reduced to either 1.1 or 0.75 when the number of infected individuals reached 400 (Fig. 1g). The segregating site trajectories for these three simulations indicate that reductions in R_0 over the course of an epidemic leave signatures in this summary statistic of viral diversity (Fig. 1h). The signatures left in the segregating site trajectories reflect the epidemiological dynamics that result from the reductions in R_0. Reducing R_0 to 1.1 results in a slower increase in the number of cases and a delayed, as well as broader, epidemic peak; as such, the number of segregating sites increases more slowly and the decline in the number of segregating sites is not apparent over the time period shown. Reducing R_0 to 0.75 results in an immediate decline in cases, with an observed drop in the number of segregating sites due to the stochastic loss of viral sublineages. Similar magnitude reductions in R_0 that were implemented later on in the simulated epidemic yielded fainter signatures of this effect in the segregating site trajectories (Figure S3).

Epidemiological inference using segregating site trajectories

To examine the extent to which inference based on segregating sites can be used for epidemiological parameter estimation, we generated a mock segregating site trajectory by forward simulating an SEIR model with an R_0 of 1.6. From this simulation, we randomly sampled 500 viral sequences (corresponding to approximately 0.78% of infections being sampled) and binned these sequences into 4-day time windows based on their sampling times (Fig. 2a). Figure 2b shows the segregating site trajectory from these binned sequences. From this trajectory, we first attempted to estimate only R_0 under the assumption that the timing of the index case t_0 is known (Methods). We estimated an R_0 value of 1.58 (95% confidence interval of 1.37 to 1.81; Fig. 2c), demonstrating that our segregating sites inference approach applied to this simulated dataset is able to recover the true R_0 value of 1.6. Lower levels of sampling effort (100 viral sequences) resulted in an R_0 estimate to 1.65 and a broader 95% confidence interval (1.30 to 2.06; Figure S4). Instead of random sampling of sequences, adopting a more uniformly distributed sampling strategy acted to reduce the uncertainty in the R_0 estimate (Figure S5). In Figure S6, we present results for the same set of sequences as those used in Fig. 2, with the sequence data binned instead in time windows of 1 day, 2 days, 6 days, and 10 days, rather than in a time window of 4 days. These results show that R_0 estimates are not biased by the use of different time window lengths.

Because the timing of the index case t_0 (in cases with a single introduction) is almost certainly not known for an emerging epidemic, we further attempted to estimate both R_0 and t_0 using the segregating site trajectory shown in Fig. 2b. We considered a range of R_0 values between 1.0 to 2.5 and a broad range of t_0 starting 50 days prior to the true start date of 0 and ending at the date of the first sampled sequence. We divided this parameter space into fine-resolution parameter combinations (R_0 intervals of 0.1 and t_0 intervals of 2 days) and ran 20 SMC simulations for every parameter combination. In Fig. 3a, we plot the mean value of the 20 SMC log-likelihoods for every parameter combination in the considered parameter space. Examination of this plot indicates that there is a log-likelihood ridge that runs between early t_0/low R_0 parameter sides, indicating that inference using

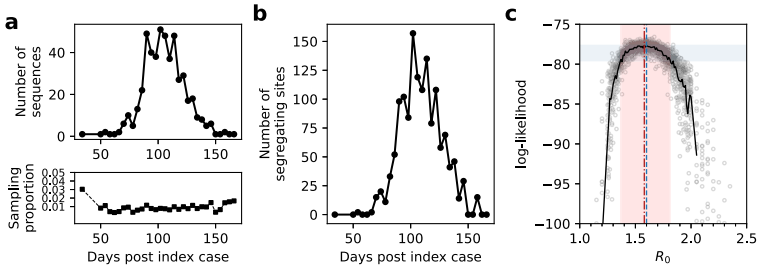

Fig. 2 | Epidemiological inference on a simulated trajectory of segregating sites. a, top The number of sampled sequences over time, binned by 4-day time windows. Sampling was done in proportion to the number of individuals recovering in a time window. In all, 500 sequences were sampled over the course of the simulated epidemic. **a**, bottom The proportion of sampled individuals in each time window, obtained by dividing the number of sampled individuals by the number of individuals who recovered during a time window. **b** Simulated segregating site trajectory from the sampled sequences, by time window. **c** Estimation of R_0 using Sequential Monte Carlo (SMC). Points show log-likelihood values from different SMC simulations. R_0 values between 1.0 and 1.25 and between 2.0 and 2.5 were considered with a step size of 0.1. R_0 values between 1.25 and 2.0 were considered with a step size of 0.01. Solid black curve shows the mean of 20 data points for each R_0 value. The vertical red dashed line shows the maximum likelihood estimate (MLE) of R_0. The red band shows the 95% confidence interval of R_0. The vertical blue line shows the true value of $R_0 = 1.6$. The MLE and 95% CI were obtained using the mean log-likelihood values. The 95% CI band included the set of R_0 values with log-likelihoods that fell within 1.92 units of the highest mean log-likelihood value, based on a chi-squared distribution with 1 degree of freedom. Model parameters for the simulated data set are: $R_0 = 1.6$, $\gamma_E = {}^1/2$ days^{-1}, $\gamma_I = 1/3$ days^{-1}, population size $N = 10^5$, $t_0 = 0$, and the per genome, per transmission mutation rate $\mu = 0.2$. Initial conditions are $S(t_0) = N\text{-}1$, $E(t_0) = 0$, $I(t_0) = 1$, and $R(t_0) = 0$. A time step of $\tau = 0.1$ days was used in the Gillespie τ -leap algorithm.

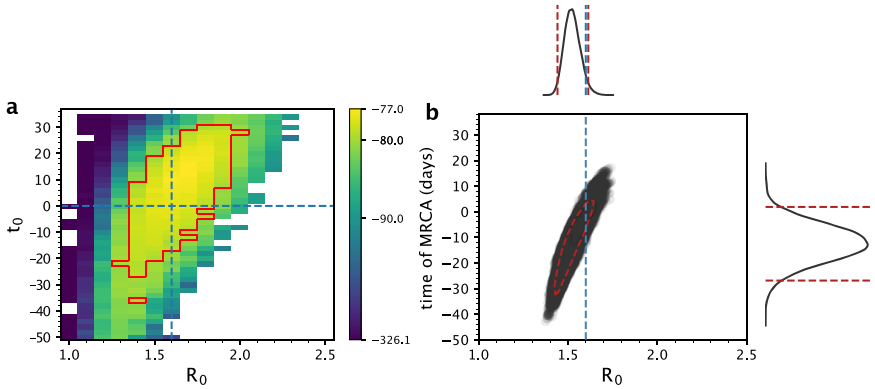

Fig. 3 | Joint estimation of the basic reproduction number (R_0) and the timing of the index case (t_0) using simulated data, and comparison against PhyDyn. **a** The log-likelihood surface based on the segregating site trajectory shown in Fig. 2b is shown over a range of R_0 and t_0 parameter combinations. The log-likelihood value shown in each cell is the mean log-likelihood value calculated from 20 SMC simulations. Blank cells yielded mean log-likelihood values of negative infinity. The red boundary shows the set of (R_0, t_0) values that fall within the 95% confidence region. Parameter combinations within the red boundary have mean log-likelihood values that fall within 2.996 units of the highest mean log-likelihood value, based on a chi-squared distribution with 2 degrees of freedom. **b** Joint density plot for R_0 and the time of the most recent common ancestor (tMRCA), as estimated using PhyDyn[6] on the same set of 500 sampled sequences. Dashed red line in the joint density plot shows the 95% HPD interval of the joint density.

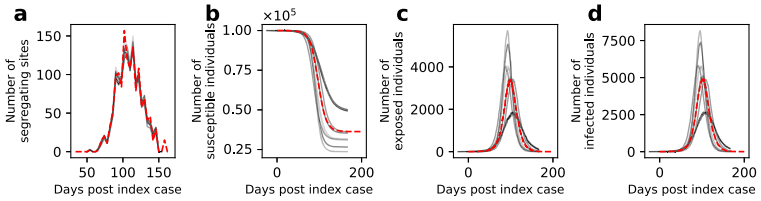

Fig. 4 | Reconstruction of unobserved state variables. **a** Simulated trajectory of the number of segregating sites (dashed red), alongside reconstructed trajectories of the number of segregating sites (gray). **b** Simulated dynamics of susceptible individuals (dashed red), alongside reconstructed dynamics of susceptible individuals (gray). **c** Simulated dynamics of exposed individuals (dashed red), alongside reconstructed dynamics of exposed individuals (gray). **d** Simulated dynamics of infected individuals (dashed red), alongside reconstructed dynamics of infected individuals (gray). Reconstructed state variables were obtained by running the particle filter using R_0 and t_0 parameter values randomly sampled from within the 95% CI region, with a further condition that the log-likelihood from the run exceeded the 95% CI region log-likelihood cutoff shown in Fig. 3a. To show that resampling of particles during the SMC performs effectively, we show in Figure S7 the dynamics of these unobserved state variables in particles that are sampled at different time points during the SMC procedure that may be lost by the end of the simulation as a result of resampling.

segregating site trajectories can in principle estimate both t_0 and R_0. The parameter combination with the highest mean log-likelihood was $R_0 = 1.7$ and $t_0 = 16$ days, with the true parameter combination of $R_0 = 1.6$ and $t_0 = 0$ days falling within the 95% confidence region of the estimated parameters. Our results therefore indicate that joint estimation of these parameters is thus possible in cases where a single introduction is responsible for igniting local circulation. Using our estimates of R_0 and t_0, we reconstructed the dynamics of the segregating sites (Fig. 4a) and unobserved state variables: the number of susceptible, exposed, and infected individuals over time (Fig. 4b-d). These reconstructed state variables captured the true epidemiological dynamics, demonstrating that our segregating sites approach can be used to infer epidemiological variables that generally go unobserved.

As mentioned in the Introduction, there are existing phylodynamic inference approaches available that can estimate epidemiological model parameters using viral phylogenies that have been reconstructed from sequence data. Of particular note is the coalescent-based inference approach developed by Volz[13] that has been implemented as PhyDyn[6] in BEAST2. To compare our results using the segregating sites approach to results using PhyDyn, we generated mock viral nucleotide sequences from our set of 500 sampled sequences (Methods) and used these nucleotide sequences as input into PhyDyn. Assuming the same epidemiological model structure and using uninformative priors, PhyDyn was similarly able to recover the true R_0 value of 1.6 used in the forward simulation (Fig. 3b; 95% credible interval = 1.44 to 1.61). Because PhyDyn infers epidemiological parameters using a tree-based method, the program does not estimate the time of the index case t_0. Instead, it estimates the time of the most recent common ancestor (tMRCA) of the viral phylogeny. The credible interval of PhyDyn's tMRCA estimate spanned from −26.89 to 1.87 days post the true time of the index case ($t_0 = 0$). Times of a most recent common ancestor, however, are generally later (and never earlier) than the time of the index case. This is because some viral lineages likely go unsampled and the pruning of these unsampled lineages results in a tMRCA that can be considerably later than the time of the index case t_0[28]. As such, interpretation of the PhyDyn results would almost certainly result in timing the index case t_0 as less than 0 (too early), given 1.87 days as the top end of the tMRCA credible interval. This potentially early estimate of t_0 may be due to the "push-of-the-past" effect[29], which results from the assumption of deterministic dynamics in the inference process when the underlying population dynamics are stochastic (and conditioned on the persistence of a lineage). This "push-of-the-past" effect is usually reflected in an overestimate of the growth rate (or an overestimate in R_0) in coalescent-based inference approaches that are applied to datasets with small population sizes during their exponential growth phase[16]. Here, because R_0 controls not only the rate of increase in the number of infected individuals at the start of the simulated epidemic but also

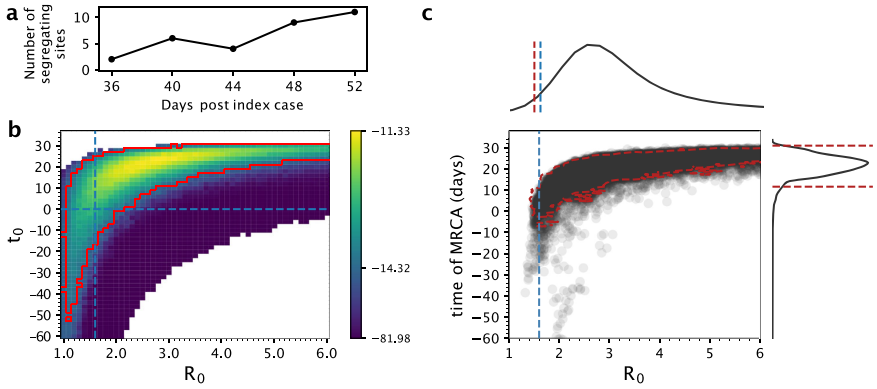

Fig. 5 | Joint estimation of the basic reproduction number (R_0) and the timing of the index case (t_0) using early samples from the simulation, with comparison against PhyDyn. a Simulated trajectory of the number of segregating sites using early sequences. Sequences were binned into 4-day windows, with 10 individuals sampled from each time window. **b** The log-likelihood surface based on a segregating site trajectory shown in panel (**a**). As in Fig. 3a, the log-likelihood value shown in each cell is the mean log-likelihood value calculated from 20 SMC simulations and the 95% CI boundary shown in red contains sets of parameter combinations that fall within 2.966 log-likelihood units of the maximum log-likelihood. Blank cells had mean log-likelihood values of negative infinity. (**c**) Joint density plot for R_0 and the time of the most recent common ancestor (tMRCA), as estimated using PhyDyn[6] on the same set of 50 sampled sequences. Dashed red line in the joint density plot shows the 95% HPD interval of the joint density. For R_0, only the lower bound of the 95% HPD is shown as the upper bound is above 6. In panels **a** through **c**, simulations were parameterized with a per genome, per transmission mutation rate of μ = 0.2.

the time at which the simulated epidemic starts to decline, the "push-of-the-past" effect may instead be reflected in a tMRCA estimate that likely occurs too early. Because our inference approach implements stochastic population dynamics, it appropriately accounts for the push-of-the-past effect, as do phylodynamic inference approaches that incorporate stochastic population dynamics (e.g., birth-death models).

Because the impetus for developing the segregating sites inference approach was based on the extent of phylogenetic uncertainty present early on in an epidemic, we re-applied the inference approach to sequences sampled early on during the simulated epidemic, with time window bins ending on days 36, 40, 44, 48, and 52 (Fig. 5a). During each of these five-time windows, we sampled 10 sequences, resulting in a total of 50 sampled sequences. Our results on this subset of simulated data indicate that R_0 and t_0 could again be jointly estimated, although the confidence intervals for R_0 and t_0 were both considerably broader, as expected with a much shorter time series (Fig. 5b). Similarly, on this same subset of data, PhyDyn's 95% credible intervals were considerably broader (95% credible interval for R_0 = 1.48 to 10.80). For this particular time series, both the segregating sites approach and PhyDyn tended to overestimate the true value of R_0 = 1.6 (Figs. 5b, 5c). For PhyDyn, the "push-of-the-past" effect[29] may have contributed to the overestimation of R_0.

To determine whether there might be an upwards bias in the estimation of R_0 using the segregating sites approach, we simulated an additional short dataset under the same epidemiological model structure and model parameterization, with the exception of the mutation rate μ, which we increased from 0.2 to 0.4. To calculate the segregating sites trajectory, we sampled from this simulation as we did for Fig. 5a–c, with 10 sequences sampled in each of the five time windows (Figure S8a). The maximum likelihood estimates of R_0 using our segregating sites approach did not overestimate the true R_0 of 1.6 in this dataset, although the time of the index case was again estimated to be slightly later than the true value of t_0 = 0 (Figure S8b). Compared to the results on the μ = 0.2 short dataset (Fig. 5b), the 95% confidence region spanned over a similar extent of parameter space. PhyDyn also did not overestimate R_0 on this μ = 0.4 short dataset (Figure S8c). Moreover, its 95% credible interval was considerably smaller than on the μ = 0.2 short dataset. This result makes sense: at higher mutation rates, phylogenetic uncertainty is reduced and tree-based inference approaches are expected to improve. In contrast, a low-dimensional summary statistic, such as the number of segregating sites cannot take advantage of the higher-dimensional structure present in the sequence data.

Epidemiological inference using SARS-CoV-2 sequences from France

We applied the segregating sites inference approach to a set of SARS-CoV-2 sequences sampled from France between January 23, 2020, and March 17, 2020 (the date on which a country-wide lockdown began). We decided to apply our approach to this set of sequences for several reasons. First, many of the 479 available full-genome sequences from France over this time period appear to be genetically very similar to one another[30], indicating that one major lineage took off in France (or at least, that most sampled sequences derived from one major lineage). This lineage would be the focus of our analysis. Second, an in-depth epidemiological analysis previously inferred R_0 for France prior to the March 17 lockdown measures that were implemented[31]. That analysis fit a compartmental infectious disease model to epidemiological data that included case, hospitalization, and death data. Because our segregating sites inference approach can accommodate epidemiological model structures of arbitrary complexity, we could adopt the same model structure as in this previous analysis. We could also set the epidemiological parameters that were assumed fixed in this previous analysis to their same values. By controlling for model structure and the set of model parameters assumed as given, we could ask to what extent sequence data corroborate the R_0 estimates arrived at from detailed fits to epidemiological data.

To apply our segregating sites approach to the viral sequences from France, we first identified the subset of the 479 sequences that constituted a single, large lineage. To keep with the "tree-free" emphasis of our approach, we identified this subset of sequences (n = 432) without inferring a phylogeny (Methods). Using phylogenetic inference, however, we confirmed that our subset of sequences constituted a single clade, with sequences from France falling outside of this clade being excluded (Figure S9). To generate a segregating site trajectory from these sequences, we defined 4-day time windows such that the last time window ended on March 17, 2020. Figure 6a shows the number of sequences falling into each time window. Figure 6b shows the segregating site trajectory calculated from these sequences.

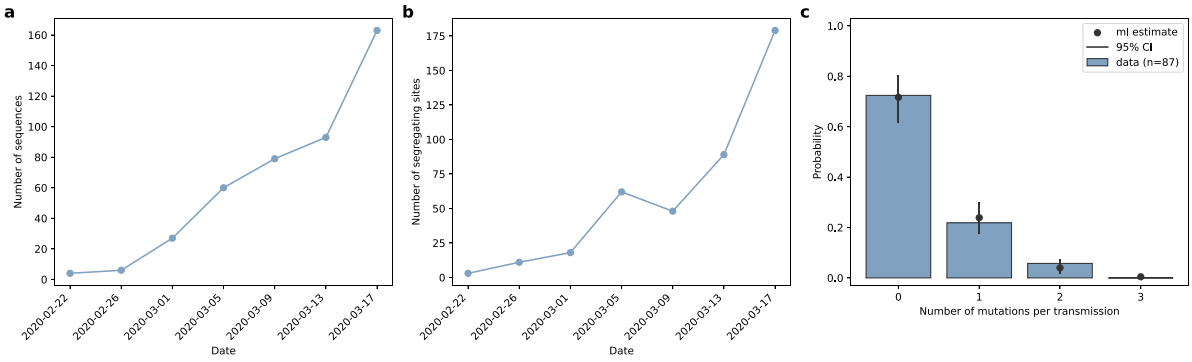

Fig. 6 | Sequences and parameters used for epidemiological inference based on SARS-CoV-2 sequences from France. a The number of sequences sampled over time, calculated using a 4-day time window. **b** The segregating site trajectory calculated from the binned sequences shown in panel (**a**). **c** Estimation of the per-genome, per-transmission mutation rate μ. The histogram shows the fraction of 87 analyzed transmission pairs with consensus sequences that differ from one another by the number of mutations shown on the x-axis. The mean number of mutations per transmission is μ = 0.33 (95% CI = 0.22–0.48). Black dots represent the probability of observing 0, 1, 2, and 3 mutations assuming a Poisson distribution with a mean of 0.33. Vertical black error bars span the probability of observing 0, 1, 2, and 3 mutations assuming Poisson distributions with mean values of 0.22 and 0.48.

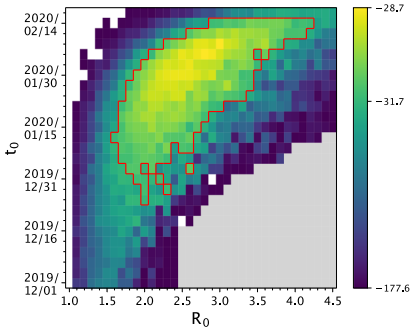

Fig. 7 | Joint estimation of the basic reproduction number R_0 and the time of the index case t_0 for the France SARS-CoV-2 data. The joint log-likelihood surface based on the estimated segregating site trajectory for the France data. Each cell shows the mean log-likelihood value based on 10 SMC simulations. Blank cells indicate mean log-likelihood values of negative infinity. Gray cells indicate where log-likelihood values were not evaluated. The red lines denote the set of parameter values that fall within the 95% confidence interval. A few 'islands' of parameter combinations that fall either outside or inside the 95% CI are apparent and are due to the variation in the log-likelihood values obtained from the SMC simulations.

We parameterized the model with a per genome, per transmission mutation rate μ using consensus sequence data from established SARS-CoV-2 transmission pairs that were available in the literature[32–35] (Methods). Specifically, for each of the 87 transmission pairs we had access to, we calculated the nucleotide distance between the consensus sequence of the donor sample and that of the recipient sample and fit a Poisson distribution to these data (Fig. 6c). Using this approach, we estimated a μ value of 0.33, corresponding approximately to one mutation occurring every 3 transmission events.

Similar to the approach we undertook with our simulated data, we first attempted to jointly estimate R_0 and the timing of the index case t_0 for this segregating site trajectory. We considered a broad parameter space over which to calculate log-likelihood values. Specifically, we considered R_0 values between 1.0 and 4.5 and t_0 values of between December 1st, 2019 and February 14th, 2020. We ran 10 SMC simulations and calculated the mean log-likelihood for each parameter combination (Fig. 7). We estimated R_0 to be 3.0 (95% confidence interval = 1.6 to 4.2), consistent with the R_0 estimate of 2.9 (95% confidence interval = 2.81 to 3.01) arrived at through epidemiological time series analysis[31]. We estimated t_0 to be February 8th, 2020 (95% confidence interval = December 25, 2019, to February 14, 2020).

We decided to further consider an alternative model that allowed for multiple introductions of the focal lineage into France (Methods). This decision was based on evolutionary analyses that have shown that regional SARS-CoV-2 epidemics in Europe (as well as in the United States) were initiated through multiple introductions rather than only a single one[36]. Instead of attempting to jointly estimate R_0 and t_0, we attempted to jointly estimate R_0 and a parameter η using the segregating site trajectory. The parameter η quantifies the extent to which transmission between France and regions outside of France is reduced relative to transmission occurring within France. This model further required specification of the time at which the basal genotype evolved outside of France, which we refer to as t_e. We considered a broad parameter space over which to calculate log-likelihood values (R_0 values between 1.0 and 4.0 and η values between 10^{-8} and 10^{-1}) and three different t_e values: December 24, 2019, January 1, 2020, and January 8, 2020 (Methods). At each of these t_e values, we ran 10 SMC simulations and calculated the mean log-likelihood for each parameter combination (Fig. 8a–c). We estimated R_0 to be 2.6 (95% CI = 2.0 to 4.0), 2.7 (95% CI = 2.0 to 4.0), and 2.3 (95% CI = 2.1 to 4.0), respectively, under t_e = December 24, 2019, January 1, 2020, and January 8, 2020. These results indicate that the inferred R_0 values are relatively insensitive to the assumed emergence time of the basal genotype outside of France. At later assumed values of t_e, our estimates for η were higher, indicating that later emergence times were compensated for by a higher transmission rate between infected individuals outside of France and susceptible individuals within France.

We reconstructed the unobserved state variables for the multiple-introductions model using SMC simulations parameterized with R_0 and η values that were sampled from the parameter spaces shown in Fig. 8, using the same approach we used for reconstructing state variables on the mock segregating sites trajectory. These reconstructed variables are shown in Fig. 9. As expected for an epidemic with an $R_0 > 1$, the total number of infected individuals increased exponentially over the time period considered (Fig. 9d–f). In Fig. 9g–i, we plot the reconstructed cumulative number of recovered individuals over time. These cumulative trajectories indicate that by mid-March 2020, approximately 0.009% to 2.044% of individuals in France had recovered from infection from this SARS-CoV-2 lineage. These cumulative predictions can be roughly compared against findings from a serological study that was conducted over this time period in France[37]. Based on a survey of 3221 individuals, this study found that 0.41% of

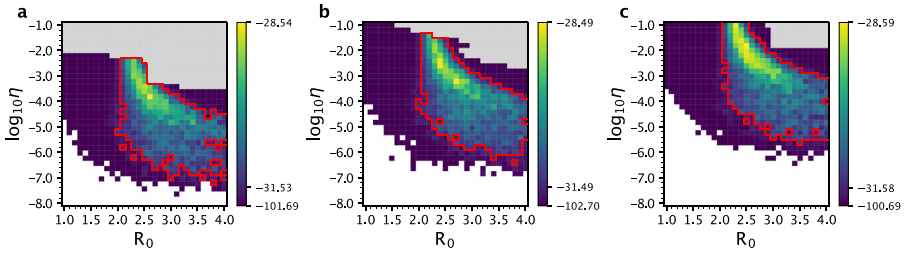

Fig. 8 | Joint estimation of the basic reproduction number R_0 and the transmission-reduction parameter η for the multiple-introductions model using the France data. The joint log-likelihood surface based on the estimated segregating site trajectory for the France data is shown under three different basal genotype emergence times: t_e = December 24, 2019 (**a**), January 1, 2020 (**b**), and January 8, 2020 (**c**). Each cell shows the mean log-likelihood value based on 10 SMC simulations. Blank cells indicate mean log-likelihood values of negative infinity. Gray cells indicate where log-likelihood values were not evaluated due to extended simulation time. The red lines in each panel denote the set of parameter combinations that fall within the 95% confidence interval. As in Fig. 7, a few 'islands' of parameter combinations are apparent due to the variation in the log-likelihood values obtained from the SMC simulations.

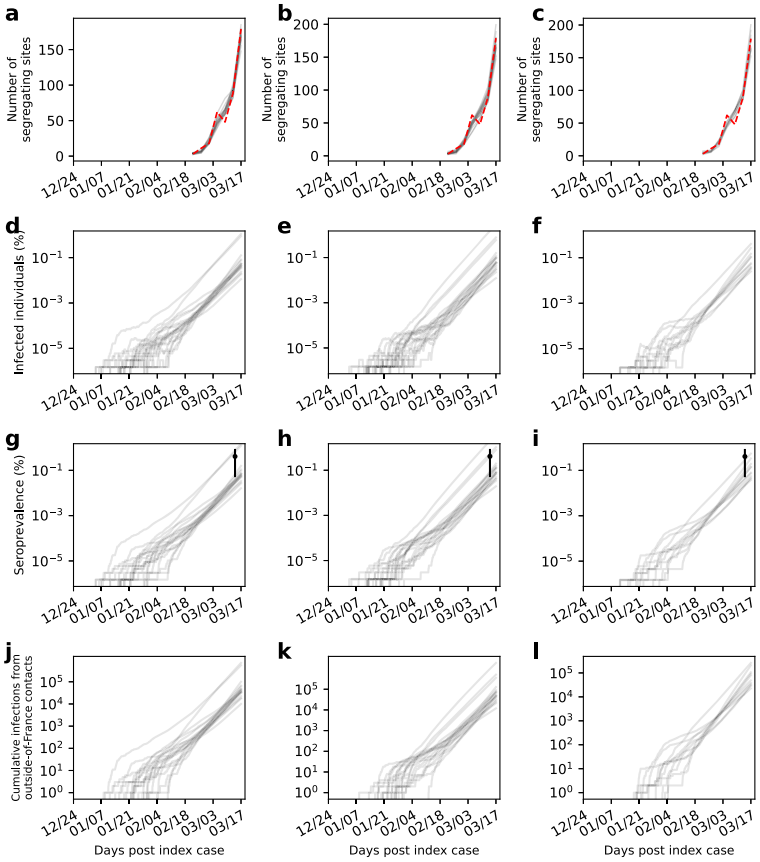

Fig. 9 | Trajectories of reconstructed state variables for the France data under the multiple-introductions model. State variables are reconstructed for the multiple-introductions model with three different values assumed for the emergence time of the basal genotype: t_e = December 24, 2019 (first column), January 1, 2020 (second column), and January 8, 2020 (third column). **a**–**c** Segregating site trajectory for the France SARS-CoV-2 data (red), alongside reconstructed segregating site trajectories (gray). **d**–**f** Reconstructed dynamics of the number of infected individuals ($E_1 + E_2 + I$) over time, shown in percent of France's population. **g**–**i** Reconstructed dynamics of the cumulative number of recovered individuals over time, shown in percent of France's population. Independent estimates of the fraction of the population that has been infected with SARS-CoV-2 by mid-March are shown in black. Estimates are from a serological study conducted during the time window March 9-15, 2020[37]. **j**–**l** Reconstructed dynamics of the cumulative number of infections in France that resulted from contact with infected individuals outside of France. Reconstructed state variables shown in panels (**a**–**l**) were obtained by running the particle filter using R_0 and t_0 parameter values randomly sampled from within the 95% CI region, with a further condition that the log-likelihood from the run exceeded the 95% CI region log-likelihood cutoff shown in Fig. 8a–c, respectively.

individuals (95% confidence interval = 0.05% to 0.88%) had gotten infected with SARS-CoV-2 by March 9 to 15, 2020 (Fig. 9g–i). Our estimates fall in line with these independent estimates. Of note, our estimates should fall on the low side of these independent estimates because other, smaller clades were also circulating in France during the time period studied and infections with viruses from these other clades would also contribute to seropositivity levels. We also emphasize that this is necessarily a rough comparison because seroconversion does not occur exactly at the point of recovery. It can occur over a broader range of times, ranging from prior to recovery to many days following symptom onset[38]. Finally, in Fig. 9j–l, we plotted the reconstructed cumulative number of infections that resulted directly from contact with individuals outside of France. By the first sampled time window (ending on February 22, 2022), our SMC results indicate that there were very likely repeated introductions of this lineage into France, with the majority of sampled particles pointing towards hundreds of introductions of this lineage into France by this time point.

Discussion

Here, we developed a statistical inference approach to estimate epidemiological parameters from virus sequence data. Our inference approach is a "tree-free" approach in that it does not rely on the reconstruction of viral phylogenies to estimate model parameters. One benefit of using such an approach for parameter estimation of emerging viral pathogens is that, early on in an epidemic, phylogenetic uncertainty present in time-resolved viral phylogenies is significant, and tree-based phylodynamic inference approaches would need to integrate over this uncertainty. This is oftentimes computationally intensive, especially when many sequences have been sampled. The computational complexity of our "tree-free" approach, in contrast, does not scale with the number of sampled sequences. Instead, the runtime required for parameter inference depends on the number of genotypes that evolve over the course of the model simulations. This number in turn is affected by the proposed basic reproduction number, the proposed time of the index case in the single introduction model, and the magnitude of the per genome, per transmission mutation rate μ. A second benefit to our tree-free approach is that it can estimate the time of the index case (in a single-introduction scenario), whereas tree-based inference methods estimate the time of the most recent common ancestor. This is a benefit when the question of interest focuses on when a viral lineage emerges and starts to spread. Instead of viral phylogenies being the data that statistically interface with the epidemiological models, our approach uses a population genetic summary statistic of the sequence data, namely the number of segregating sites present in time-binned sets of viral sequences. Our inference approach benefits from being plug-and-play in that it can easily accommodate different epidemiological model structures.

Based on fits to a simulated data set, we have shown that segregating site trajectories can be used to estimate the basic reproduction number R_0 and the timing of the index case t_0 in cases where a single introduction can be assumed. We further fit a multiple-introductions epidemiological model to a segregating site trajectory that was calculated from SARS-CoV-2 sequence data from France, estimating a basic reproduction number R_0 of approximately 2.3-2.7. These results are consistent with previous estimates from an epidemiological analysis and consistent with a serological study conducted in mid-March 2020.

Our inference approach relies on several assumptions that are shared by existing phylodynamic inference methods. Most notably, it relies on an assumption that all mutations are phenotypically neutral. However, a recent analysis of SARS-CoV-2 sequences has shown evidence for purifying selection, even early on during the pandemic[39]. Indeed, within the set of SARS-CoV-2 sequences from France, we observe 170 nonsynonymous mutations and 138 synonymous mutations (a ratio of 1.23:1). Given the number of nonsynonymous sites ($n = 68{,}540$) and the number of synonymous sites ($n = 19{,}255$) in the SARS-CoV-2 genome, we would expect, under neutrality, a ratio of 3.56:1. This underrepresentation of nonsynonymous genetic variation points towards purifying selection in our analyzed dataset. A more recent analysis also raises the possibility of adaptive evolution occurring during early 2020[40]. Incorporating non-neutral genetic variation into inference approaches such as ours and existing phylodynamic ones is complicated, although some statistical approaches have started to tackle this goal[9]. In the context of our segregating sites inference approach, directly incorporating non-neutral evolution will increase model complexity considerably, and assumptions would need to be made about the distribution of mutational fitness effects. Rather than incorporating non-neutral evolution within our approach, we can for now consider how the occurrence of non-neutral evolution would impact our parameter estimates. With purifying selection at play, we would expect to see less genetic variation than in its absence. As such, the number of segregating sites in any time window would be lower than it would be under neutrality. Our inference approach, assuming neutrality, would therefore bias R_0 estimates to be low and, in single-introduction models, the timing of the index case t_0 to be late. In multi-introduction models, our estimate of η would be biased high.

Our approach also assumes infinite sites and the absence of homoplasies. While these assumptions are limiting over longer periods of sequence evolution, our approach is intended to be used for emerging viral pathogens, sampled over shorter periods of time, when levels of genetic diversity are still low. As such, these assumptions will likely not be violated in cases where this approach will come in useful. We would also like to note that the infinite sites assumption could in principle be relaxed, but this would make the simulations in the inference approach substantially more costly. Furthermore, as time goes on, not only do chances of repeated mutations at sites increase, but genetic diversity increases. As such, phylogenetic uncertainty will decrease, such that existing tree-based phylodynamic inference approaches will become increasingly informative and segregating site trajectories less informative.

While our inference approach does adopt assumptions of phenotypic neutrality and infinite sites, it does not assume a constant sampling rate or a specific sampling process throughout the time period over which sequences are collected. As we have shown in Fig. 1b, sampling effort does impact the segregating sites trajectory: the greater the sampling effort, the larger the number of segregating sites. For our inference approach to perform effectively, sampling effort therefore needs to be matched between the simulations and the empirical data. This matching of sampling effort is implemented in the particle filter. However, the number of samples sequenced per time window is not particularly informative of model parameters (except in the case of extremely high sampling effort when certain low R_0 model parameterizations cannot appropriately evaluate the expected number of segregating sites in a time window because the number of sampled sequences exceed the number of simulated recoveries). The reason why the number of samples is not particularly informative of model parameters is because, under our approach, sampling of individuals does not impact the underlying epidemiological dynamics: individuals are sampled upon recovery, once they are no longer infectious. That the number of observed samples is not highly informative of model parameters we see as a benefit of our approach because sampling effort and testing rates can change dramatically over the course of an emerging pandemic or over the early period of an emerging viral lineage as surveillance efforts ramp up. In contrast, sampling times of sequences have been shown to be highly informative of model parameters in the case of birth-death models, with sampling process misspecification resulting in the possibility of arriving at biased parameter estimates[41].

While the number of sampled sequences is largely uninformative of model parameters, our approach does have to make an assumption

of when individuals are sampled. In our simulated dataset and in our application to SARS-CoV-2, we assumed that individuals were sampled as they recovered. This sampling scheme decision was based on our understanding that the time of symptom onset often follows peak viral load for many emerging viral pathogens[42] and an assumption that most testing early on in a pandemic involves individuals who develop symptoms. It is important to note that if the assumed sampling scheme is mismatched with the empirical sampling scheme, parameter estimates may be biased. For example, if individuals were instead sampled as they transitioned from the exposed class to the infectious class, rather than upon recovery, and we assumed in our model that individuals were sampled upon recovery, then our R_0 estimates would be biased high.

Finally, we would like to note that setting the per genome, per transmission mutation rate to a constant value does not correspond to an assumption of a constant molecular clock. A constant molecular clock requires that the number of substitutions per unit time remains the same. Our assumption is that the mean number of nucleotide changes that occur during a transmission event between a donor and a recipient (at the consensus level) stays constant over time. This would almost certainly be the case unless the fidelity of the viral polymerase was evolving over the period considered. Changes in the substitution rate could come about if the generation interval between transmission events changes due, for example, to the implementation of non-pharmaceutical interventions or increased symptom awareness. A shortening of the generation interval (defined as the time between infection and onward transmission) would increase the number of transmission events that occur per unit time and thereby result in an increase in the substitution rate. In contrast, a lengthening of the generation interval would result in fewer transmission events occurring per unit time, thereby decreasing the population-level substitution rate. Changes in the generation interval can emerge from an underlying epidemiological model, such that our assumption of a constant per genome, per transmission event mutation rate does not preclude or conflict with the observation of changes in the substitution rate over time.

The analysis we presented here focuses on statistical inference using sequence data alone. In recent years, there has also been a growing interest in combining multiple data sources – for example, sequence data and epidemiological data or serological data - to more effectively estimate model parameters. The few existing studies that have incorporated additional data while performing phylodynamic inference have shown the value in pursuing this goal[7,43,44]. As a next step, we aim to extend the segregating sites approach developed here to incorporate epidemiological data and/or serological data more explicitly. Straightforward extension is possible due to the state-space model structure that is at the core of the particle filtering routine we use.

Our analysis focused on phylodynamic inference based on sequence data belonging to a single viral lineage, with either a single index case or multiple introductions from an outside reservoir. Our approach, however, can be expanded in a straightforward manner to multiple viral lineages. This is especially useful in cases like SARS-CoV-2, where many regions have witnessed the introduction of multiple clades[10,45]. In this case, a single segregating sites trajectory could be calculated for each clade, such that multiple segregating site trajectories could be simultaneously fit to under specified constraints such as the basic reproduction number being the same across all clades. Different clades could also be allowed to differ in their reproductive numbers, such that questions relating to the selective advantage of some clades over others could be addressed. As such, this inference method, designed for emerging pathogens with low levels of genetic diversity, may continue to be useful for endemic pathogens to address questions related to the emergence of new viral lineages.

Methods

Brief overview of inference approach

Mutations occur during viral replication within infected individuals and these have the potential to be transmitted. During the epidemiological spread of an emerging virus or viral lineage, the virus population (distributed across infected individuals) thus accrues mutations and diversifies genetically. This joint process of viral spread and evolution can be simulated forward in time using compartmental models, with patterns of epidemiological spread leaving signatures in the evolutionary trajectory of the virus population. Parameters of these compartmental models that govern patterns of epidemiological spread can thus in principle be estimated using viral sequence data. Here, similar in spirit to existing inference approaches based on summary statistics[46–50], we develop a statistical inference approach that fits compartmental epidemiological models to time series of a low-dimensional summary statistic calculated from sequence data. Specifically, we use trajectories of the number of segregating sites from samples of the viral population taken over time for statistical inference. Because we propose the use of our method early on in an epidemic (or during the early expansion of a viral lineage), we focus primarily on estimating the basic reproduction number R_0 using this inference approach.

Epidemiological model simulations and calculation of segregating site trajectories

To simulate mock data of segregating site trajectories, we specify a compartmental epidemiological model and simulate the model under demographic stochasticity using Gillespie's τ-leap algorithm. Here, we use a susceptible-exposed-infected-recovered (SEIR) model whose stochastic dynamics are governed by the following equations:

$$S_{t+\Delta t} = S_t - n_{S\to E} \tag{1}$$

$$E_{t+\Delta t} = E_t + n_{S\to E} - n_{E\to I} \tag{2}$$

$$I_{t+\Delta t} = I_t + n_{E\to I} - n_{I\to R} \tag{3}$$

$$R_{t+\Delta t} = R_t + n_{I\to R} \tag{4}$$

where:

$$n_{S\to E} \sim Pois\left(\beta \frac{S_t}{N} I_t \Delta t\right) \tag{5}$$

$$n_{E\to I} \sim Pois(\gamma_E E_t \Delta t) \tag{6}$$

$$n_{I\to R} \sim Pois(\gamma_I I_t \Delta t) \tag{7}$$

where β is the transmission rate, N is the host population size, γ_E is the rate of transitioning from the exposed to the infected class, γ_I is the rate of recovering from infection, and Δt is the τ-leap time step used. R_0 is given by β/γ_I. The epidemiological dynamics of this model can be simulated from the above equations alone. Additional complexity is needed to incorporate virus evolution throughout the course of the simulation. To incorporate virus evolution, we partition exposed individuals and infected individuals into genotype classes, with genotype 0 being the reference genotype present at the start of the simulation. Mutations to the virus occur at the time of transmission, with the number of mutations that occur in a single transmission event given by a Poisson random variable with mean μ, the per-genome, per-transmission event mutation rate. We assume infinite sites such that new mutations necessarily result in new genotypes. New mutations

and new genotypes are both assigned integer indices in order of their appearance. When new mutations are generated at a transmission event, the new genotype harbors the same mutation(s) as its parent genotype plus any new mutations. We use a sparse matrix approach to store genotypes and their associated mutations to save on memory.

There are three types of events that occur in the SEIR model simulations: transitions from exposed to infected; transitions from infected to recovered; and transmission. To simulate transitions from exposed to infected, during a time step Δt, $n_{E\to I}$ individuals are drawn at random from the set of individuals who are currently reside in the exposed class. These individuals will transition to the infected class during this time step, while retaining their current genotype statuses. To simulate transitions from infected to recovered, during a time step Δt, $n_{I\to R}$ individuals are drawn at random from the set of individuals who are currently residing in the infected class. These individuals will transition to the recovered class during this time step. To simulate transmission, during a time step Δt, we add $n_{S\to E}$ new individuals to the set of exposed individuals. For each newly exposed individual, we randomly choose (with replacement) a currently infected individual as its 'parent'. If no mutations occur during transmission, then this newly exposed individual enters the same genotype class as its parent. If one or more mutations occur during transmission, this newly exposed individual enters a new genotype class, and the sparse matrix is extended to document the new genotype and its associated mutations (given as integers, without a bitstring or explicit genome structure).

We start the simulation with one infected individual carrying a viral genotype that we consider as the reference genotype (genotype 0). To calculate a time series of segregating sites, we define a time window length T ($T > \Delta t$) of a certain number of days and partition the simulation time course into discrete, non-overlapping time windows. During simulation, we keep track of the individuals that recover (transition from I to R) within a time window. For each time window i, we then sample n_i of these individuals at random, where n_i is the number of sequences sampled in a given time window based on the sampling scheme chosen. Because we have the genotypes of the sampled individuals from the sparse matrix, we can calculate the number of segregating sites s_i in any time window i. Since s_i is the number of polymorphic sites across the sampled individuals in time window i, it is simply calculated from the set of mutations harbored by the sequences of the sampled individuals. While in our simulations, we sample individuals as they recover, alternative sampling schemes can instead be assumed. For example, individuals could be sampled as they transition from the exposed to the infected class, or while they are in the infected class. We chose to sample upon recovery based on symptom development (and thereby testing) often occurring following peak viral load.

Implementation of the transmission heterogeneity model

We implement transmission heterogeneity in the epidemiological model by splitting the infected classes into a high-transmission and a low-transmission class, as has been done elsewhere[6,10]. For an SEIR model, the model extended to incorporate transmission heterogeneity becomes:

$$S_{t+\Delta t} = S_t - n_{S\to E} \tag{8}$$

$$E_{t+\Delta t} = E_t + n_{S\to E} - n_{E\to I_h} - n_{E\to I_l} \tag{9}$$

$$I_{h,t+\Delta t} = I_{h,t} + n_{E\to I_h} - n_{I_h\to R} \tag{10}$$

$$I_{l,t+\Delta t} = I_{l,t} + n_{E\to I_l} - n_{I_l\to R} \tag{11}$$

$$R_{t+\Delta t} = R_t + n_{I_h\to R} + n_{I_l\to R} \tag{12}$$

where:

$$n_{S\to E} \sim Pois\left(\beta_h \frac{S_t}{N} I_{h,t}\Delta t\right) + Pois\left(\beta_l \frac{S_t}{P} I_{l,t}\Delta t\right) \tag{13}$$

$$n_{E\to I} \sim Pois(\gamma_E E_t \Delta t) \tag{14}$$

$$n_{E\to I_h} \sim Bin(n_{E\to I}, p_H) \tag{15}$$

$$n_{E\to I_l} = n_{E\to I} - n_{E\to I_h} \tag{16}$$

$$n_{I_h\to R} \sim Pois(\gamma_I I_{h,t}\Delta t) \tag{17}$$

$$n_{I_l\to R} \sim Pois(\gamma_I I_{l,t}\Delta t) \tag{18}$$

The parameter p_H quantifies the proportion of exposed individuals who transition to the high-transmission I_h class. Parameters β_h and β_l quantify the transmission rates of the infectious classes that have high and low transmissibility, respectively. We set the values of β_h and β_l based on a given parameterization of overall R_0 and the parameter p_H. To do this, we first define, as in previous work[6,10], the relative transmissibility of infected individuals in the I_h and I_l classes as $c = \frac{\beta_h}{\beta_l}$. We further define a parameter P as the fraction of secondary infections that result from a fraction p_H of the most transmissible infected individuals. Based on given values of p_H and P, we set c, as in previous work[10], to $\frac{[\frac{1-p_H}{p_H}]}{[\frac{1}{P}-1]}$. With c defined in this way, p_H can be interpreted as the proportion of most infectious individuals that result in P of secondary infections. We set P to 0.80, to make p_H easily interpretable relative to the "20/80" rule in disease ecology[22]. Recognizing that $R_0 = \frac{p_H\beta_h + (1-p_H)\beta_l}{\gamma_I}$ in this model, we can then solve for β_l: $\frac{R_0\gamma_I}{p_H c + (1-p_H)}$, and set $\beta_h = c\beta_l$. Note that the interpretation of p_H in the context of the disease ecology rule is an approximation since this calculation does not take into consideration variation in individual R_0 that results from differences in the duration of infection or variation in individual R_0 that results from differences in the number of secondary infections that are due to stochastic effects.

Epidemiological inference using time series of segregating sites

Our inference approach relies on particle filtering, also known as Sequential Monte Carlo (SMC), to estimate model parameters and reconstruct unobserved (latent) state variables. Particle filtering calculates the likelihood of a parameterized model (more precisely, the probability of observing the time-series data marginalized over the unobserved state variables) by recurrently updating a set of particles (Figure S10). In our case, each of these particles holds a state-space model, which includes a process model component that simulates underlying epidemiological and evolutionary dynamics and an observation model that relates these latent state variables to the observed segregating sites data (Figure S11). The process model includes the unobserved epidemiological variables (e.g., S, E, I, and R) and the evolutionary components of the model (viral genotypes and mutations). From one observed segregating sites data point to the next one, the model is simulated using Gillespie's τ-leap algorithm, as described in the section above.

At the end of each time window, when the simulation reaches the next observed segregating sites data point, the observation model is used to calculate the probability of observing the observed data point given the underlying process model. This probability is calculated as follows. We calculate the expected number of segregating sites from

the model simulation by performing k 'grabs' of sampled individuals, with each grab consisting of the following steps:

- pick (without replacement) n_i individuals from the set of individuals who recovered during time window i, where n_i is the number of samples present in the empirical dataset in window i. This step mimics the process of sample collection at the same effort as in the observed data. We control for sampling effort because the extent of sampling impacts the number of segregating sites.
- calculate the simulated number of segregating sites s_i^{sim}, based on the genotypes of the sampled n_i individuals.

Between grabs, the replacement of previously sampled individuals occurs. We then calculate the mean number of segregating sites for window i by taking the average of all k s_i^{sim} values. Finally, we calculate the probability of observing s_i segregating sites in window i, given the model-simulated mean number of segregating sites, using a Poisson probability mass function parameterized with the mean s_i^{sim} value and evaluated at s_i. As a special case in the calculation of this probability, if the number of individuals who recovered during a given time window i is less than the number that needs to be sampled (n_i), then the particle's probability of observing the number of segregating sites s_i is set to 0. The calculated probabilities serve as the weights for the particles.

Particle weights obtained at the end of each window are used 1) to resample particles for the next time window according to their assigned weights and 2) to calculate the likelihood of a parameterized model. In the particle filtering algorithm, the likelihood is obtained by averaging particle weights within each window and then multiplying these average particle weights across all time windows with observations. For time windows without observations ($n_i = 0$), particle weights are assigned a value of 0 if the virus has died out stochastically and 1 if the virus continues to persist in the population. These weights are used for resampling, but do not contribute to the calculation of the likelihood. We adopt this approach to filter out particles during early time windows that have undergone stochastic extinction.

Latent state variables are reconstructed by randomly sampling a particle at the end of an SMC simulation and plotting the values of its simulated latent state variables over time. All of our SMC simulations were performed with 200 particles and $k = 50$ grabs. Note that the complexity of this inference method is largely independent of the number of input sequences. This stands in contrast to phylodynamic inference approaches that frequently down-sample sequences to reduce runtime.

Converting simulated sequences into nucleotide sequences for the performance comparison against PhyDyn

Simulated sparse matrices were converted to nucleotide alignments by first generating a reference sequence with the same length as the maximum number of mutations in the sparse matrix and choosing an A, C, G, or T nucleotide at each site with equal probability. A mutated sequence was generated for each genotype represented in the sparse matrix by replacing the reference allele at that position with another nucleotide chosen with equal probability. The final FASTA alignment was generated by identifying the simulated sequence associated with each sampled individual. Generation of the simulated FASTA file was done using Python v3.9.4 with Numpy v1.19.4.

The simulated FASTA alignment was used to generate a BEAST2 XML file from a template XML which was generated in part using Beauti v2.6.6. This template used a JC69 nucleotide substitution model with no invariant sites. We assumed an uncorrelated log-normally distributed relaxed clock with a uniform [0.0, 1E-2] prior on the mean and a uniform [0.0,2.0] prior on the standard deviation.

A single-deme structured coalescent prior as defined by the following equations was implemented using PhyDyn v1.3.8:

$$\frac{dE}{dt} = \frac{\beta IS}{N} - \gamma_E E \tag{19}$$

$$\frac{dI}{dt} = \gamma_E E - \gamma_I I \tag{20}$$

$$\frac{dR}{dt} = \gamma_I I \tag{21}$$

where $\beta = R_0\gamma_I$. A population size of 10^5 with a single initially infected individual was used. We assume infected individuals remain exposed for an average of 2 days ($1/\gamma_E$) and infectious ($1/\gamma_I$) for an average of 3 days. R_0 was estimated using a uniform [1.0, 10.0] prior. All sampled sequences were assigned to the infected ("I") class.

Sampled parameters and trees were logged every 1000 states and all MCMC chains were run for at least 209 M (Fig. 3b), 64 million (Fig. 5c), 150 million (Figure S8c) iterations. The first 10% of MCMC chains were discarded as burn-in and the ESS values of all parameters were >200, as identified by Tracer v1.7.1 (10.1093/sysbio/syy032).

Epidemiological model structure and parameterization used in the SARS-CoV-2 analysis

The process model we use in our application to SARS-CoV-2 sequence data from France is based on a previously published epidemiological model[31]. We base our process model on this published model to allow for a direct comparison of inferred R_0 values between our sequence-based analysis and their analysis that focuses on SARS-CoV-2 spread in France over a similar time period. Their analysis was based on fitting an epidemiological model to a combination of case, hospitalization, and death data. Their model structure, once implemented using Gillespie's τ-leap algorithm, is given by:

$$S_{t+\Delta t} = S_t - n_{S\to E1} \tag{22}$$

$$E_{1,t+\Delta t} = E_{1,t} + n_{S\to E1} - n_{E1\to E2} \tag{23}$$

$$E_{2,t+\Delta t} = E_{2,t} + n_{E1\to E2} - n_{E2\to I} \tag{24}$$

$$I_{t+\Delta t} = I_t + n_{E2\to I} - n_{I\to R} \tag{25}$$

$$R_{t+\Delta t} = R_t + n_{I\to R} \tag{26}$$

where:

$$n_{S\to E1} \sim Pois\left(\beta\frac{S_t}{N}I_t\Delta t\right) + Pois\left(\beta\frac{S_t}{N}E_{2,t}\Delta t\right) \tag{27}$$

$$n_{E1\to E2} \sim Pois(\gamma_{E1}E_{1,t}\Delta t) \tag{28}$$

$$n_{E2\to I} \sim Pois(\gamma_{E2}E_{2,t}\Delta t) \tag{29}$$

$$n_{I\to R} \sim Pois(\gamma_I I_t\Delta t) \tag{30}$$

The parameters are the transmission rate β, the rate of transitioning from the E_1 class to the E_2 class γ_{E1}, the rate of transitioning from the E_2 class to the I class γ_{E2}, and the rate of transition from the I class to the R class γ_I. The average duration of time spent in the E_1 class given by $1/\gamma_{E1} = 4$ days, the average duration of time spent in the E_2 class given by $1/\gamma_{E2} = 1$ day, and the average duration of time spent in

the infected class given by $1/\gamma_I = 3$ days. Their model assumes that the transmission efficiency β of exposed class 2 (E_2) and that of the infected class I are the same; their model considers E_2 and I as distinct classes to interface with the case data, where symptoms are assumed to not appear before an individual has transitioned to class I. We maintain the model structure with E_1, E_2, and I rather than reducing it to a model structure with just a single E and a single I class to keep the same overall distribution of infection times as in their model.

Because SARS-CoV-2 dynamics are characterized by substantial levels of transmission heterogeneity[10,23,51] and we have shown in Fig. 1 that transmission heterogeneity impacts segregating site trajectories, we expanded the compartmental epidemiological model for SARS-CoV-2 described above to include transmission heterogeneity in a manner similar to the one we used in Fig. 1. Based specifically on the analysis by Paireau and colleagues[52], we set p_H to 0.10, such that 10% of infections are responsible for 80% of secondary infections. Analogous to the approach we undertook for the simulated data, we jointly estimated R_0 and t_0 using the segregating site trajectory shown in Fig. 6b.

Based on phylogenetic analyses that have indicated that early introductions of SARS-CoV-2 into focal regions likely resulted from multiple introductions rather than a single one, we considered a modified version of the epidemiological model that would allow for multiple introductions. The modification relied on the incorporation of infections within France that resulted from direct contact with infected individuals outside of France, termed the viral "reservoir". Similar to the approach adopted by some existing phylodynamic analyses[12], the viral population dynamics in this reservoir are simplified to exponential growth. This infected population from outside of France acts as another source of infection for susceptible individuals within France, allowing for multiple introductions of SARS-CoV-2 into France.

As in the focal region, new genotypes are expected to emerge in the outside reservoir. As we assume an infinite sites model, the genotypes that emerge in the outside reservoir and in the focal region will not overlap except in the basal genotype that is first introduced to the focal region. For this reason, and because the basal genotype is expected to be considerably more common than any of the viral genotypes that stem from it, we consider only the repeated introduction of the basal genotype into France. Starting at the time of emergence of the basal genotype in the outside reservoir (t_e), we let the number of individuals infected with this basal genotype Y_t grow exponentially:

$$Y_t = e^{r(t-t_e)} \tag{31}$$

where r is the intrinsic growth rate of the basal genotype. Based on empirical estimates[53,54], we set the intrinsic growth rate to 0.22 day^{-1}. To set t_e, we first identified the genotype sampled in France that is genetically closest to the reference strain Wuhan/Hu-1 (MN908947.3). This basal genotype differs from Wuhan/Hu-1 by 4 nucleotides: C241T, C3037T, C14408T, and A23403G. Using GISAID data, we then identified sequences with collection locations outside of France that carried all four of these mutations that define the basal genotype. The earliest of these sequences including the four basal genotype-defining mutations was collected on January 25, 2020, in Australia, suggesting that the basal genotype had been circulating prior to January 25, 2020. Considering the potential delay between emergence and the time of first detection, we considered three distinct t_e values: December 24th, 2019, January 1st, 2020, and January 8th, 2020.

Individuals infected in this outside reservoir can transmit their infection to susceptible individuals within France. The rate at which they transmit the infection is reduced relative to the rate at which infected individuals within France transmit the infection to susceptible individuals within France. We let the factor by which transmission is reduced be given by the factor η. During a τ-leap timestep, the number of individuals within France who become infected from contact with an infected individual outside of France is therefore given by:

$$n^O_{S\to E1} \sim Pois\left(\beta\eta\frac{S_t}{N}Y_t\Delta t\right) \tag{32}$$

As we are considering only the transmission of the basal genotype from infected individuals in the outside reservoir to susceptible individuals in France, all of these newly infected individuals will carry the basal genotype unless mutation occurs during the transmission process. Our simplifying assumption that only the basal genotype can be introduced into France from the outside reservoir ignores the possibility that genotypes that are derived from the basal genotype enter France from the outside reservoir. Strictly speaking, we think this assumption is unlikely to be met. However, at very early time points in France's epidemic, most of the genotypes outside of France should still be the basal genotype, and only at later time points should the frequencies of derived genotypes increase outside of France. Introduction of these derived genotypes at these later time points could result in the establishment of viral sublineages in France. However, because autochthonous infections would be high at this point, these viral sublineages would very likely go unsampled. As such, we do not think that our assumption of only the basal genotype being introduced into France would have a dramatic effect on our results. We can consider, however, the effects that violation of this assumption would have on our parameter estimates: if derived genotypes were introduced into France and sampled (or their descendants sampled), then the number of segregating sites that would have evolved within France would be lower than we are currently taking it to be. As such, our current estimate of R_0 would be biased high.

Estimation of the per genome, per transmission event mutation rate

We set the per-genome, per-transmission mutation rate parameter μ to 0.33. This is based on the fit of a Poisson distribution to the number of de novo substitutions that were observed in 87 transmission pairs of SARS-CoV-2 from four studies[32–35]. Accession numbers for 78/87 of these transmission pairs are available in Table S1. Accession numbers for the remaining pairs were provided by the corresponding authors of the relevant publication[34]. Sequence data were aligned to Wuhan/Hu-1 (MN908947.3)[55] using MAFFT v.7.464[56]. Insertions relative to Wuhan/Hu-1 were removed and the first 55 and last 100 nucleotides of the genome were masked. De novo substitutions for each pair were identified in Python v.3.9.4 (http://www.python.org) using NumPy v.1.19.4[57]. Ambiguous nucleotides were permissively included in the identification of de novo substitutions (e.g., an R nucleotide was assumed to match both an A and a G). The mean number of substitutions between transmission pairs is the maximum likelihood estimate for the rate parameter of the Poisson distribution. The 95% confidence interval was calculated using the exact method using SciPy v.1.5.4[58].

The value for $\mu = 0.33$ is consistent with population-level substitution rate estimates for SARS-CoV-2, which range from 7.9×10^{-4} to 1.1×10^{-3} substitutions per site per year[28,59]. With a genome length of SARS-CoV-2 of approximately 30,000 nucleotides and a generation interval of approximately 4.5 days[60], these population-level substitution rates would correspond to per genome, per transmission mutation rates of between 0.29 and 0.41, respectively.

Estimation of segregating site trajectories for the France data

We downloaded all complete and high-coverage SARS-CoV-2 sequences with complete sampling dates sampled through March 17th, 2020 (https://doi.org/10.55876/gis8.230123mt) in France and uploaded through April 29th, 2021 from GISAID[61]. Sequences were aligned to Wuhan/Hu-1 using MAFFT v.7.464. Insertions relative to Wuhan/Hu-1 were removed. Any sequences with fewer than 28000 A, C, T, or G

characters were removed. Following this filtering protocol, our dataset included 479 sequences. We masked the first 55 and last 100 nucleotides in the genome as well as positions marked as "highly homoplasic" in early SARS-CoV-2 sequencing data (https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/archived_vcf/problematic_sites_sarsCov2.2020-05-27.vcf). Pairwise SNP distances were calculated in a manner that accounted for IUPAC ambiguous nucleotides in Python using NumPy. To subset these data to a single clade circulating within France, we identified the connected components of this pairwise distance matrix with a cutoff of 1 SNP in Python using SciPy and identified the shared SNPs relative to Wuhan/Hu-1 between all sequences in each connected component. The largest connected component contained 308 sequences which shared the substitutions C241T, C3037T, C14408T, and A23403G. Our final dataset included these 308 as well as 124 sequences from connected components that shared these four substitutions relative to Wuhan/Hu-1. We included connected components in which all sequences had an *N* at any of the four clade-defining sites of the largest connected component. Two sequences were excluded as they differed from all other sequences in the dataset by > 7 SNPs. This dataset includes 112 of the 186 sequences analyzed in Danesh et al.[11]. Sequences were binned into four-day windows, aligned such that the last window ended on the latest sampling date. The number of segregating sites in each window was calculated in Python using NumPy. Ambiguous nucleotides were permissively considered in the calculation of segregating sites, e.g., an *N* nucleotide was assumed to match all four nucleotides, whereas an *R* nucleotide was assumed to match only *A* and *G* nucleotides. This matching assumption results in a lower bound estimate for the number of segregating sites in any time window. If we instead count an *N* nucleotide at a site as a mutation, the number of segregating sites in each time window is much larger (Figure S12a). However, it is unlikely that an *N* nucleotide indicates a mutation; it is much more likely that an *N* indicates an inability to call a nucleotide based on low read depth or poor quality scores at a *site*. If we count *N* nucleotides as matching observed variation but count other ambiguous nucleotides (e.g., *R*) as mutations, the segregating site trajectory is barely affected (Figure S12b). This is because there are very few non-*N* ambiguous nucleotides in the dataset. As such, our parameter estimates on the France dataset are unlikely to be impacted by our assumption of ambiguous nucleotides matching observed genetic variation at their respective sites.

Phylogenetic analysis of SARS-CoV-2 sequences from France

To confirm that the subset of sequences from France obtained from finding connected components formed an evolutionary lineage/clade, we first combined the 479 sequences sampled from France with 100 randomly-selected complete, high-coverage sequences sampled from outside France through March 17th, 2020 and uploaded to GISAID through April 29th, 2021. These sequences were aligned to Wuhan/Hu-1 using MAFFT, insertions were removed, and the sites described above were masked. This alignment was concatenated with the aligned sequences from France. IQ-Tree v. 2.0.7[62] was used to construct a maximum likelihood phylogeny, and ModelFinder[63] was used to find the best fit nucleotide substitution model (GTR + F + I). Small branches were collapsed. TreeTime v. 0.8.0[64] was used to remove any sequences with more than four interquartile distances from the expected evolutionary rate, rooting at Wuhan/Hu-1. Treetime was also used to generate a time-aligned phylogeny assuming a clock rate of 1×10^{-3} substitutions per site per year with a standard deviation of 5×10^{-4} substitutions per site per year, a skyline coalescent model, marginal time reconstruction, accounting for covariation, and resolving polytomies. Maximum likelihood phylogenies were visualized in Python using Matplotlib v. 3.3.3[65] and Baltic (https://github.com/evogytis/baltic).

Reporting summary

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

Data availability

The simulated data generated in this study are available at https://github.com/koellelab/segregating-sites. The transmission pair data used to estimate the per-genome, per-transmission event mutation rate is provided in Table S1. The SARS-CoV-2 viral genome sequences used in the France analysis are available from GISAID (Supplementary information; https://doi.org/10.55876/gis8.230123mt). Due to the size of datasets, source data (excluding genome sequences downloaded from GISAID) are available at https://github.com/koellelab/segregating-sites.

Code availability

Python code used for generation of all figures is available on GitHub: https://github.com/koellelab/segregating-sites.

References

1. Stadler, T. & Bonhoeffer, S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. R. Soc. B* **368**, 20120198 (2013).
2. Popinga, A., Vaughan, T., Stadler, T. & Drummond, A. J. Inferring epidemiological dynamics with Bayesian coalescent inference: the merits of deterministic and stochastic models. *Genetics* https://doi.org/10.1534/genetics.114.172791 (2014).
3. Ratmann, O. et al. Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. *Mol. Biol. Evolution* **34**, 185–203 (2017).
4. Volz, E. M. et al. Phylodynamic analysis to inform prevention efforts in mixed HIV epidemics. *Virus Evolution* **3**, vex014 (2017).
5. Stadler, T., Kühnert, D., Rasmussen, D. A. & du Plessis, L. Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data. *PLoS Curr* https://doi.org/10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f (2014).
6. Volz, E. M. & Siveroni, I. Bayesian phylodynamic inference with complex models. *PLoS Comput Biol.* **14**, e1006546 (2018).
7. Vaughan, T. G. et al. Estimating Epidemic Incidence and Prevalence from Genomic Data. *Mol. Biol. Evol* **36**, 1804–1816 (2019).
8. Rasmussen, D. A., Boni, M. F. & Koelle, K. Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. *Mol. Biol. Evol* **31**, 258–271 (2014).
9. Rasmussen, D. A. & Stadler, T. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. *Elife* **8**, e45562 (2019).
10. Miller, D. et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat. Commun.* **11**, 5518 (2020).
11. Danesh, G. et al. Early phylodynamics analysis of the COVID-19 epidemic in France. *Peer Community J.* **1**, e45 (2021).
12. Geidelberg, L. et al. Genomic epidemiology of a densely sampled COVID-19 outbreak in China. *Virus Evol* **7**, veaa102 (2021).
13. Volz, E. M. Complex population dynamics and the coalescent under neutrality. *Genetics* **190**, 187–201 (2012).
14. Stadler, T. Sampling-through-time in birth–death trees. *J. Theor. Biol.* **267**, 396–404 (2010).
15. Stadler, T. et al. Estimating the basic reproductive number from viral sequence data. *Mol. Biol. Evol.* **29**, 347–357 (2012).
16. Boskova, V., Bonhoeffer, S. & Stadler, T. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLoS Comput. Biol.* **10**, e1003913 (2014).
17. Kühnert, D., Stadler, T., Vaughan, T. G. & Drummond, A. J. Phylodynamics with Migration: A Computational Framework to Quantify

Population Structure from Genomic Data. *Mol. Biol. Evol.* **33**, 2102–2116 (2016).
18. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4**, vey016 (2018).
19. Bouckaert, R. et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol.* **15**, e1006650 (2019).
20. Stadler, T., Kuhnert, D., Bonhoeffer, S. & Drummond, A. J. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl Acad. Sci.* **110**, 228–233 (2013).
21. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
22. Woolhouse, M. E. J. et al. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proc. Natl Acad. Sci.* **94**, 338–342 (1997).
23. Sun, K. et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science* **371**, eabe2424 (2021).
24. Althouse, B. M. et al. Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLoS Biol.* **18**, e3000897 (2020).
25. Lemieux, J. E. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, eabe3261 (2021).
26. Koelle, K. & Rasmussen, D. A. Rates of coalescence for common epidemiological models at equilibrium. *J. R. Soc. Interface* **9**, 997–1007 (2012).
27. Keeling, M. J. & Rohani, P. *Modeling Infectious Diseases in Humans and Animals*. (Princeton University Press, 2008).
28. Pekar, J., Worobey, M., Moshiri, N., Scheffler, K. & Wertheim, J. O. Timing the SARS-CoV-2 index case in Hubei province. *Science* **372**, 412–417 (2021).
29. Nee, S., Holmes, E. C., May, R. & Harvey, P. Extinction rates can be estimated from molecular phylogenies. *Philos. Trans. R. Soc. Lond. B* **344**, 77–82 (1994).
30. Gámbaro, F. et al. Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020. *Eurosurveillance* **25**, (2020).
31. Salje, H. et al. Estimating the burden of SARS-CoV-2 in France. *Science* eabc3517 https://doi.org/10.1126/science.abc3517.(2020)
32. Popa, A. et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci. Transl. Med* **12**, eabe2555 (2020).
33. Braun, K. M. et al. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLoS Pathog.* **17**, e1009849 (2021).
34. Lythgoe, K. A. et al. SARS-CoV-2 within-host diversity and transmission. *Science* **372**, eabg0821 (2021).
35. San, J. E. et al. Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evol* **7**, veab041 (2021).
36. Worobey, M. et al. The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
37. Le, Vu,S. et al. Prevalence of SARS-CoV-2 antibodies in France: results from nationwide serological surveillance. *Nat. Commun.* **12**, 3025 (2021).
38. Iyer, A. S. et al. Persistence and decay of human antibody responses to the receptor binding domain of SARS-CoV-2 spike protein in COVID-19 patients. *Sci. Immunol.* **5**, eabe0367 (2020).
39. Ghafari, M. et al. Purifying Selection Determines the Short-Term Time Dependency of Evolutionary Rates in SARS-CoV-2 and pH1N1 Influenza. *Mol. Biol. Evol* **39**, msac009 (2022).
40. Neher, R. A. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. *Virus Evol* **8**, veac113 (2022).
41. Volz, E. M. & Frost, S. D. W. Sampling through time and phylodynamic inference with coalescent and birth-death models. *J. R. Soc. Interface* **11**, 20140945–20140945 (2014).
42. Linton, N. M., Akhmetzhanov, A. R. & Nishiura, H. Correlation between times to SARS-CoV-2 symptom onset and secondary transmission undermines epidemic control efforts. *Epidemics* **41**, 100655 (2022).
43. Rasmussen, D. A., Ratmann, O. & Koelle, K. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol.* **7**, e1002136 (2011).
44. Li, L. M., Grassly, N. C. & Fraser, C. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Mol. Biol. Evolution* **34**, 2982–2995 (2017).
45. Gonzalez-Reiche, A. S. et al. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* eabc1917 https://doi.org/10.1126/science.abc1917.(2020)
46. Leventhal, G. E. et al. Inferring Epidemic Contact Structure from Phylogenetic Trees. *PLoS Comput Biol.* **8**, e1002413 (2012).
47. Ratmann, O., Donker, G., Meijer, A., Fraser, C. & Koelle, K. Phylodynamic Inference and Model Assessment with Approximate Bayesian Computation: Influenza as a Case Study. *PLoS Comput Biol.* **8**, e1002835 (2012).
48. Kim, K., Omori, R. & Ito, K. Inferring epidemiological dynamics of infectious diseases using Tajima's D statistic on nucleotide sequences of pathogens. *Epidemics* **21**, 21–29 (2017).
49. Saulnier, E., Gascuel, O. & Alizon, S. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLOS Computational Biol.* **13**, e1005416 (2017).
50. Plazzotta, G. & Colijn, C. *Phylodynamics without trees: estimating R0 directly from pathogen sequences*. http://biorxiv.org/lookup/doi/10.1101/102061 (2017).
51. Adam, D. C. et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med* **26**, 1714–1719 (2020).
52. Paireau, J. et al. Early chains of transmission of COVID-19 in France, January to March 2020. *Eurosurveillance* **27**, 2001953 (2022).
53. Dehning, J. et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* **369**, eabb9789 (2020).
54. Musa, S. S. et al. Estimation of exponential growth rate and basic reproduction number of the coronavirus disease 2019 (COVID-19) in Africa. *Infect. Dis. Poverty* **9**, 96 (2020).
55. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
56. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
57. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
58. SciPy 1.0 Contributors. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
59. Duchene, S. et al. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution* **6**, veaa061 (2020).
60. Griffin, J. et al. Rapid review of available evidence on the serial interval and generation time of COVID-19. *BMJ Open* **10**, e040263 (2020).
61. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
62. Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evolution* **37**, 1530–1534 (2020).
63. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

64. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
65. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Acknowledgements

The research reported in this paper was supported by the National Institute of General Medical Sciences grant NIH/NIGMS R01 GM124280 R01 GM 12480 and R01 GM124280-03S1 (K.K.), the National Institute of Allergy and Infectious Diseases Centers of Excellence for Influenza Research and Response (CEIRR) contract # 75N93021C00017 (K.K.), and NIH NIAID F31AI154738 (M.A.M.). We thank the Koelle lab, Aaron King, Sally Otto, and Ailene MacPherson for feedback, as well as the BIRS Mathematics and Statistics of Genomic Epidemiology workshop for the opportunity to discuss this work.

Author contributions

Y.P.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing, Visualization. M.A.M.: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Writing, Visualization. K.K.: Conceptualization, Methodology, Investigation, Writing, Supervision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-38809-7.

Correspondence and requests for materials should be addressed to Katia Koelle.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

2.3 Supplementary information

Epidemiological inference for emerging viruses using segregating sites

Supplementary Information

Authors: Yeongseon Park[1], Michael Martin[1,4], Katia Koelle[2,3,*]

[1] Graduate Program in Population Biology, Ecology, and Evolution, Emory University, Atlanta, GA 30322, USA

[2] Department of Biology, Emory University, Atlanta, GA 30322, USA

[3] Emory Center of Excellence for Influenza Research and Response (CEIRR), Atlanta GA, USA

Corresponding author *: katia.koelle@emory.edu

Present affiliation [4]: Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

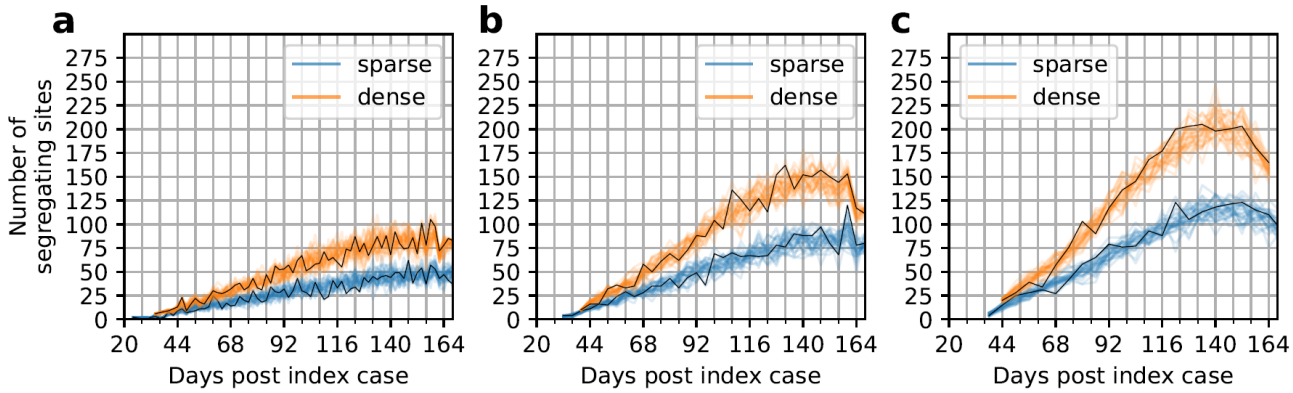

Figure S1. Segregating site trajectories under different time window lengths. Segregating site trajectories for the simulation shown in Figure 1a under dense (orange) and sparse (blue) sampling effort, when trajectories are calculated using time window lengths of (a) 2 days; (b) 4 days (as in Figure 1b); and (c) 6 days. Under the dense sampling scheme, sampling effort is 20 sequences per 2 day time window (a), 40 sequences per 4 day time window (b), and 60 sequences per 6 day time window (c). Under the sparse sampling scheme, sampling effort is 10 sequences per 2 day time window (a), 20 sequences per 4 day time window (b), and 30 sequences per 6 day time window (c). 30 randomly-sampled segregating site trajectories are shown for each sampling effort. Black lines each show a single representative segregating site trajectory. These lines are included to highlight the extent of sampling noise present under different window sizes.

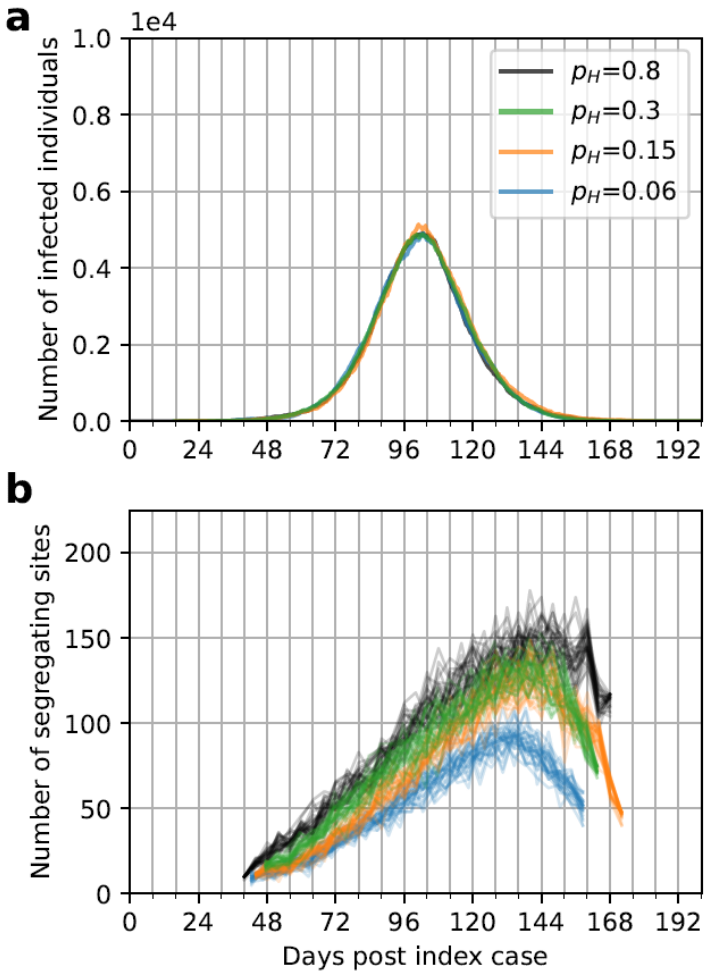

Figure S2. Segregating site trajectories under different levels of transmission heterogeneity. (a) Simulated dynamics of infected individuals (I) under an SEIR model with an R_0 of 1.6 and incorporating various levels of transmission heterogeneity compared to those of the original R_0 = 1.6 simulation without transmission heterogeneity. Transmission heterogeneity simulations shown are all shifted in time to align their epidemic peaks with the simulation without transmission heterogeneity (black line; p_H = 0.8). The transmission heterogeneity simulations considered span from low levels of transmission heterogeneity (p_H = 0.3), to intermediate levels of transmission heterogeneity (p_H = 0.15), to high levels of transmission heterogeneity (p_H = 0.06). (b) Segregating site trajectories for the simulations shown in (a). All simulations are densely sampled (40 sequences sampled per 4-day time window).

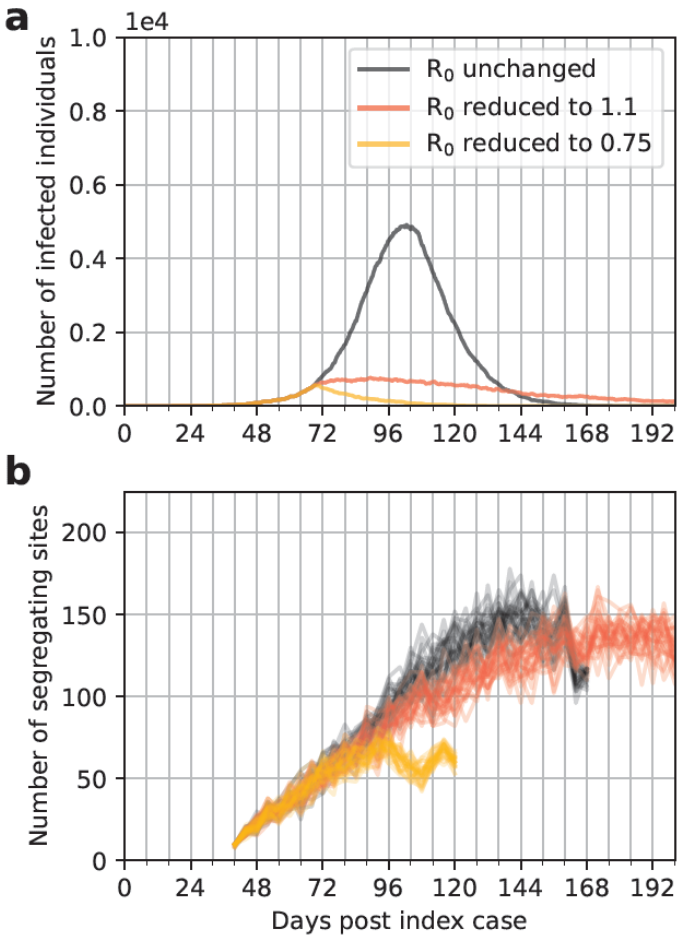

Figure S3. Segregating site trajectories under transmission reduction scenarios implemented at a later time point of the simulated epidemic. (a) Simulated dynamics of infected individuals (*I*) under an SEIR model. Changes in R_0 occurred when the number of infected individuals reached 1000. The simulation in red has R_0 decreasing to 1.1. The simulation in yellow has R_0 decreasing to 0.75. The simulation in black has R_0 remaining at 1.6. (b) Segregating site trajectories for the three simulations shown in (a). All simulations are densely sampled (40 sequences sampled per 4-day time window).

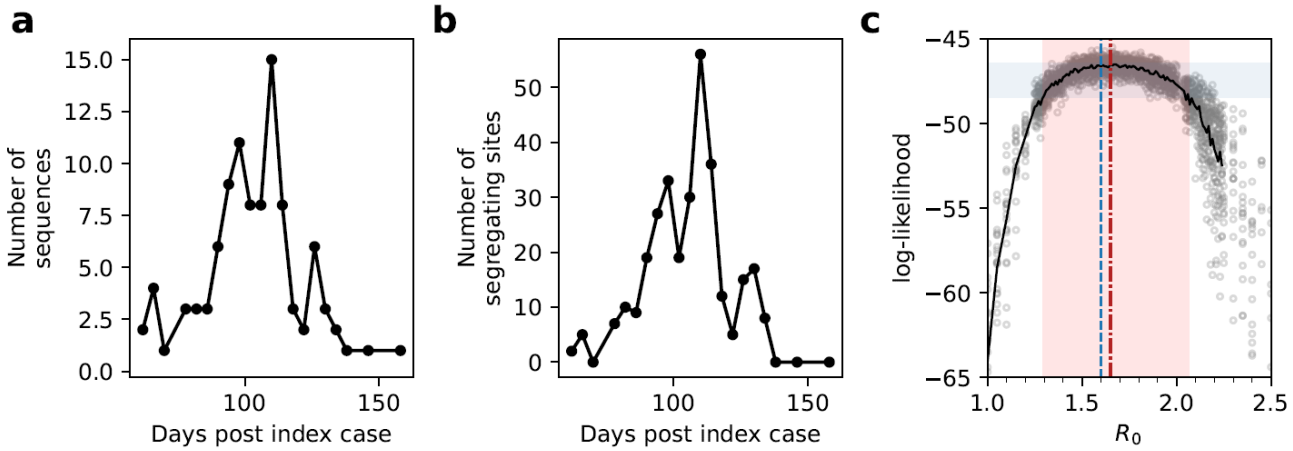

Figure S4. Epidemiological inference on a simulated trajectory of segregating sites, with lower sampling effort than in Figure 2. (a) The number of sampled sequences over time, by time window. Sampling was done in proportion to the number of individuals recovering in a time window. In all, 100 sequences were sampled over the course of the simulated epidemic. (b) Segregating site trajectory from the set of sampled sequences. (c) Estimation of R_0 using SMC. The maximum likelihood estimate for R_0 was 1.65 [95% CI = 1.30 to 2.06].

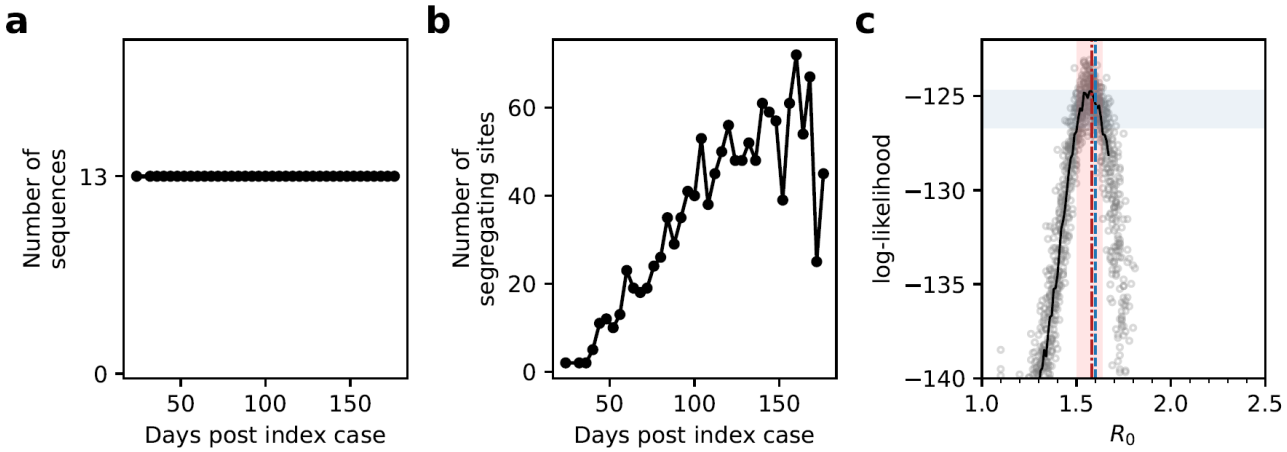

Figure S5. Epidemiological inference on a simulated trajectory of segregating sites, with uniform rather than proportional sampling. (a) The number of sampled sequences. Uniform sampling was performed by sampling 13 sequences per 4-day time window. Time windows with fewer than 13 sequences available were not included in the analysis. As such, here, only 494 sequences were used for inference. (b) Simulated segregating site trajectory from the sampled sequences. (c) Estimation of R_0 using SMC. The estimate for R_0 was 1.58 [95% CI = 1.51 to 1.62].

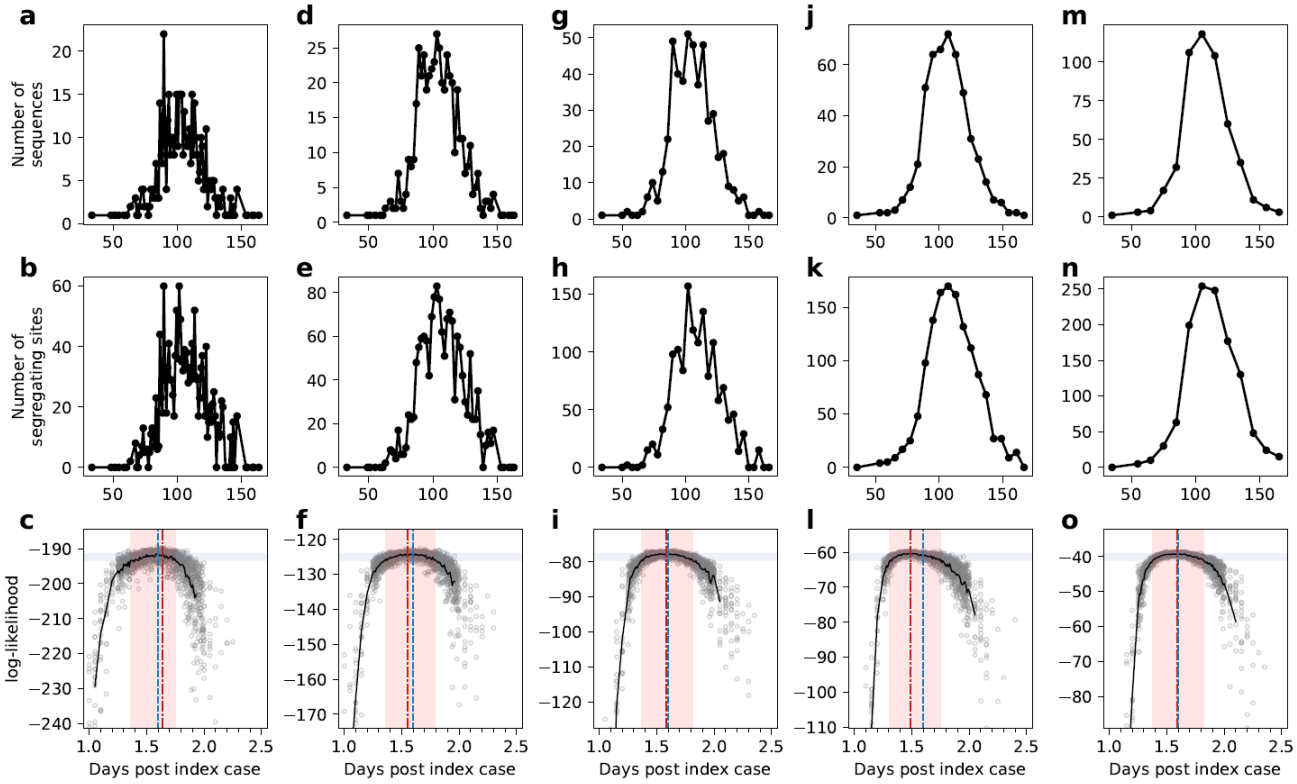

Figure S6. Epidemiological inference on the same set of sampled sequences as in Figure 2, binned at different time window lengths. For all time window lengths, the same set of 500 proportionally-sampled sequences were used. Columns show results by time window length: 1, 2, 4, 6, and 10 days, respectively. Top row (panels a,d,g,j,m): The number of sampled sequences over time, binned by time window. Middle row (panels b,e,h,k,n): Segregating site trajectories from the set of binned sequences. Bottom row (panels c,f,i,l,o): Estimation of R_0 using SMC. Panel c: the estimate for R_0 was 1.64 [95% CI = 1.36 to 1.75]. Panel f: the estimate for R_0 was 1.55 [95% CI = 1.36 to 1.79]. Panel i (reproducing the results shown in Figure 2): the estimate for R_0 was 1.58 [95% CI =1.37 to 1.81]. Panel l: the estimate for R_0 was 1.49 [95% CI = 1.31 to 1.75]. Panel o: the estimate for R_0 was 1.59 [95% CI = 1.38 to 1.82].

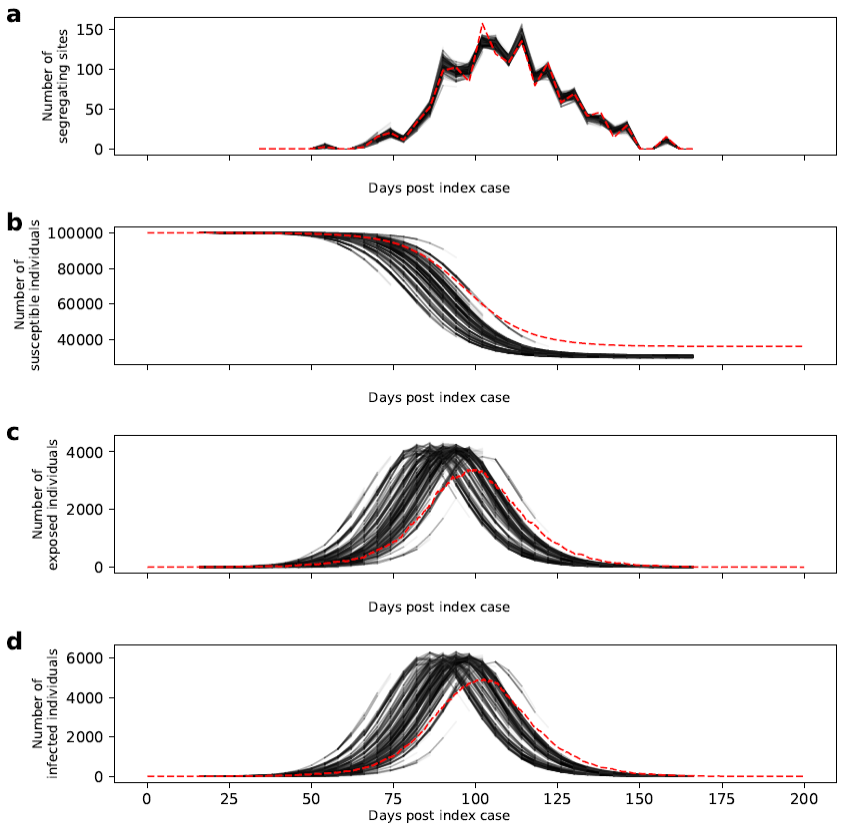

Figure S7. Resampling of particles during SMC allows for loss of low-weight particles with latent state variables that deviate from true ones. (a) Simulated trajectory of the number of segregating sites (dashed red), alongside reconstructed trajectories of the number of segregating sites (black lines). Gray lines show reconstructed segregating site trajectories from particles that were randomly sampled throughout the SMC procedure. (b) Simulated dynamics of susceptible individuals (dashed red), alongside reconstructed dynamics of susceptible individuals (black lines). Gray lines show reconstructed susceptible dynamics from particles that were randomly sampled throughout the SMC procedure. (c) Simulated dynamics of exposed individuals (dashed red), alongside reconstructed dynamics of exposed individuals (black lines). Gray lines show reconstructed dynamics of exposed individuals from particles that were randomly sampled throughout the SMC procedure. (d) Simulated dynamics of infected individuals (dashed red), alongside reconstructed dynamics of infected individuals (black lines). Gray lines show reconstructed dynamics of infected individuals from particles that were randomly sampled throughout the SMC procedure. Reconstructed dynamics from randomly sampled particles show state variables spanning from the sampled time point to the previous time point only. SMC simulations were run with R_0 = 1.7 and t_0 = 16, corresponding to the parameter combination with the highest mean log-likelihood value (see Figure 3a).

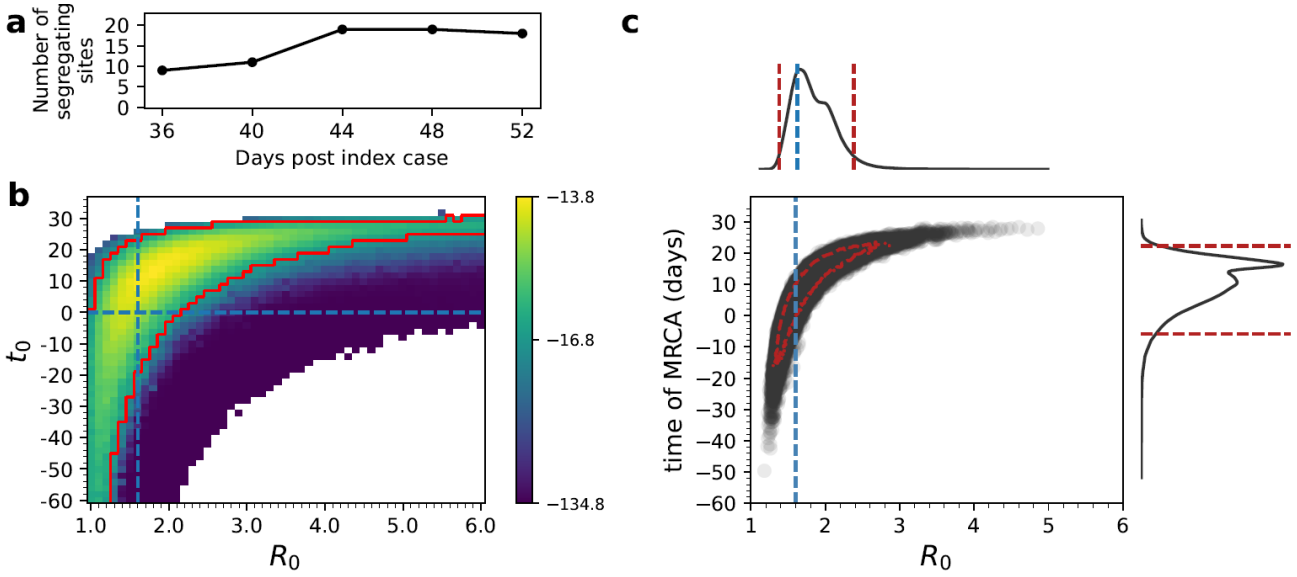

Figure S8. Joint estimation of the basic reproduction number (R_0) and the timing of the index case (t_0) using early samples from the short, μ = 0.4 simulation, with comparison against PhyDyn. (a) Simulated trajectory of the number of segregating sites using early sequences. Sequences were binned into 4-day windows, with 10 individuals sampled from each time window. (b) The log-likelihood surface based on a segregating site trajectory shown in panel (a). As in Figure 3a, the log-likelihood value shown in each cell is the mean log-likelihood value calculated from 20 SMC simulations and the 95% CI boundary shown in red contains sets of parameter combinations that fall within 2.966 log-likelihood units of the maximum log-likelihood. Blank cells had mean log-likelihood values of negative infinity. (c) Joint density plot for R_0 and the time of the most recent common ancestor (tMRCA), as estimated using PhyDyn [6] on the same set of 50 sampled sequences. Dashed red line in the joint density plot shows the 95% HPD interval of the joint density. The simulation was parameterized with a per genome, per transmission mutation rate of $\mu = 0.4$.

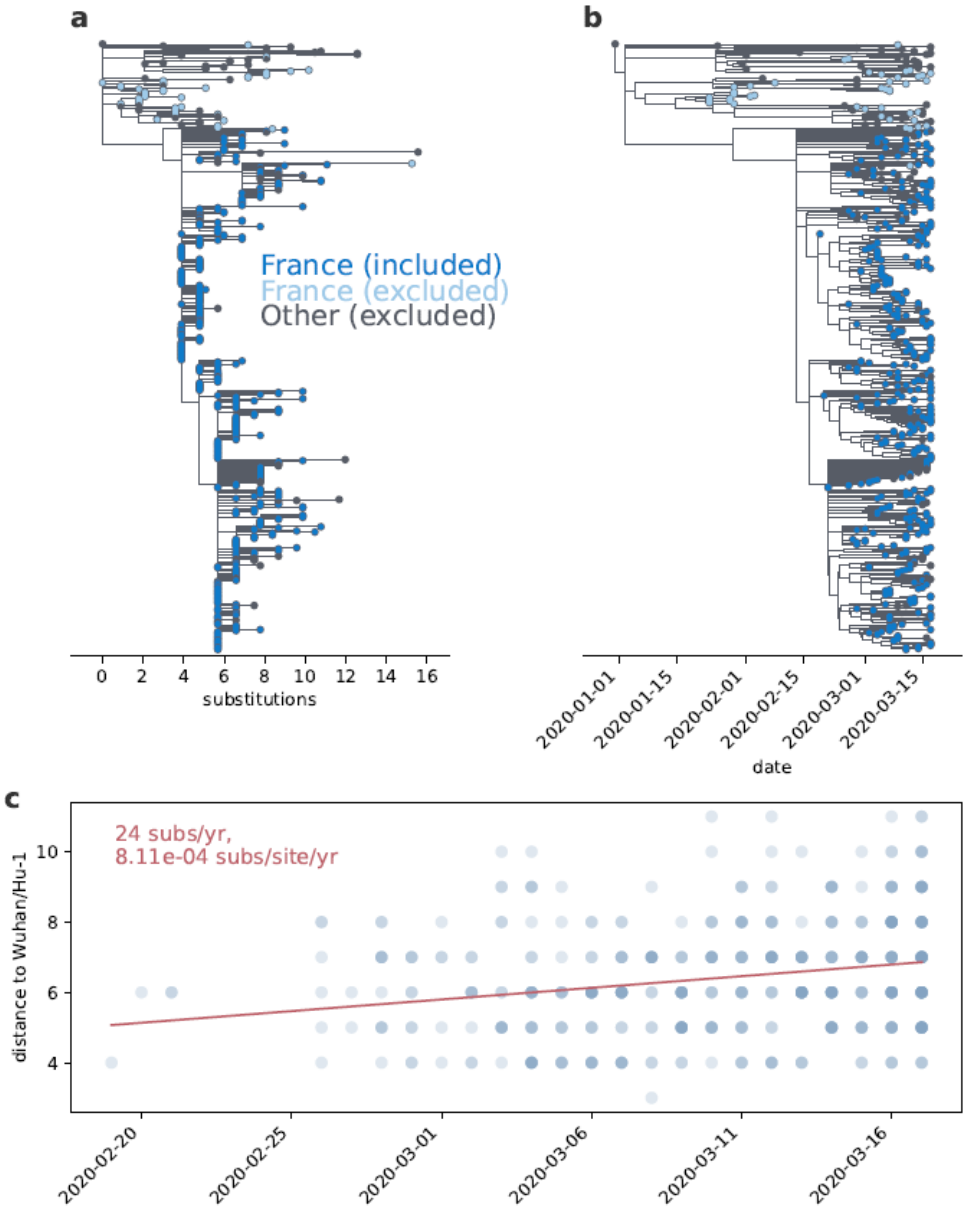

Figure S9. Inferred phylogenies for the sequences sampled from France, January 23-March 17, 2020. (a) Divergence tree, showing the number of nucleotide substitutions from Wuhan/Hu-1. Sequences from France are colored in blue, with dark blue coloring indicating sequences that were included in our single-lineage analysis and light blue coloring indicating sequences that were excluded from our analysis. Tips colored in gray denote genetically similar sequences sampled from outside of France during this time period. (b) Time-aligned maximum likelihood phylogeny, with coloring of sequences as in panel (a). (c) Plot showing genetic distances between sequences in the focal (dark blue) clade and the reference sequence Wuhan/Hu-1. A linear fit to these data yields a substitution rate of 8.11 x 10^{-4} substitutions per site per year, comparable to other reported substitution rates inferred for SARS-CoV-2 [1].

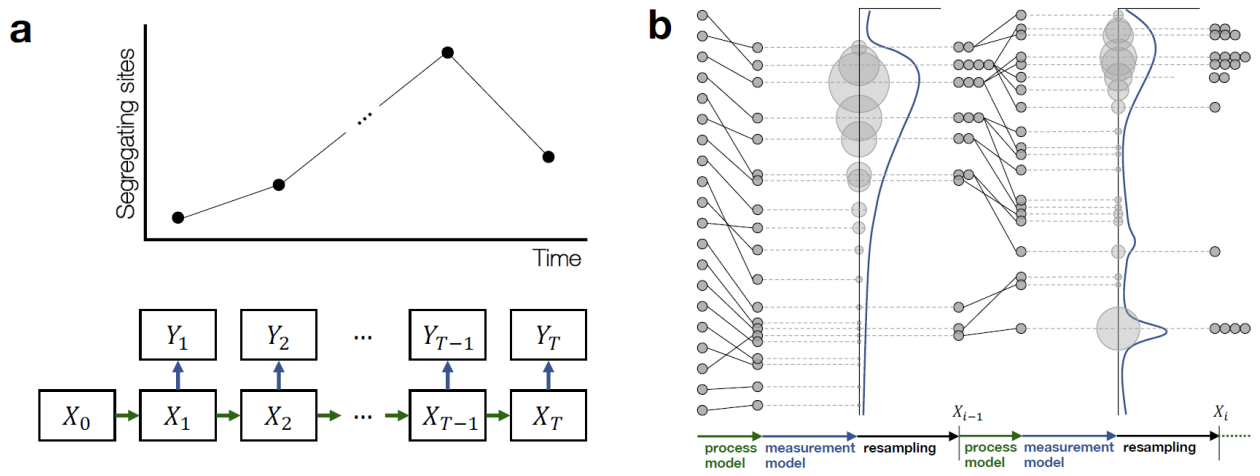

Figure S10. Depiction of the segregating sites inference approach using particle filtering. (a) Panel showing an observed segregating sites trajectory (top) and a schematic of a state-space model (bottom). A segregating site trajectory is obtained from available sequence data by first binning viral samples into consecutive, non-overlapping time windows according to their collection dates. The trajectory is then calculated by counting the number of polymorphic sites in the set of viral sequences in each time window. In a state-space model, the process model simulates underlying dynamics of latent variables over time. Here, the process model comprises the boxes labeled X_i (with i = 0, 1, …T-1, T) and the arrows between these boxes. The measurement model (depicted by the arrows between X_i and Y_i) relates the underlying state variables to the observed data Y_i. The observed data are the number of segregating sites over the time windows. (b) The particle filtering algorithm starts with a number of particles (shown as gray circles aligned in the first column), each initialized with initial state X_0. During a time window, the process model of each particle is simulated forward, arriving at a latent state X_1 at the end of the first time window. This is depicted by the black lines connecting the columns of particles. The measurement model is then used to calculate the weight of each particle (represented by the size of the light gray circles), which is defined as the probability of observing a given number of segregating sites for the time window i (s_i) based on each particle's simulated dynamics. We used Poisson distribution with rate parameter λ = mean s_i^{sim} (see Figure S11). Based on their weights, particles are sampled with replacement to generate a new set of particles for the next window. This resampling is shown using dotted lines that horizontally connect light gray circles to the gray-colored particles. This process continues until the last time window. Overall likelihoods of a given model parameterization are calculated by averaging the weights of the particles during each observation time window and then multiplying these average weights across time points.

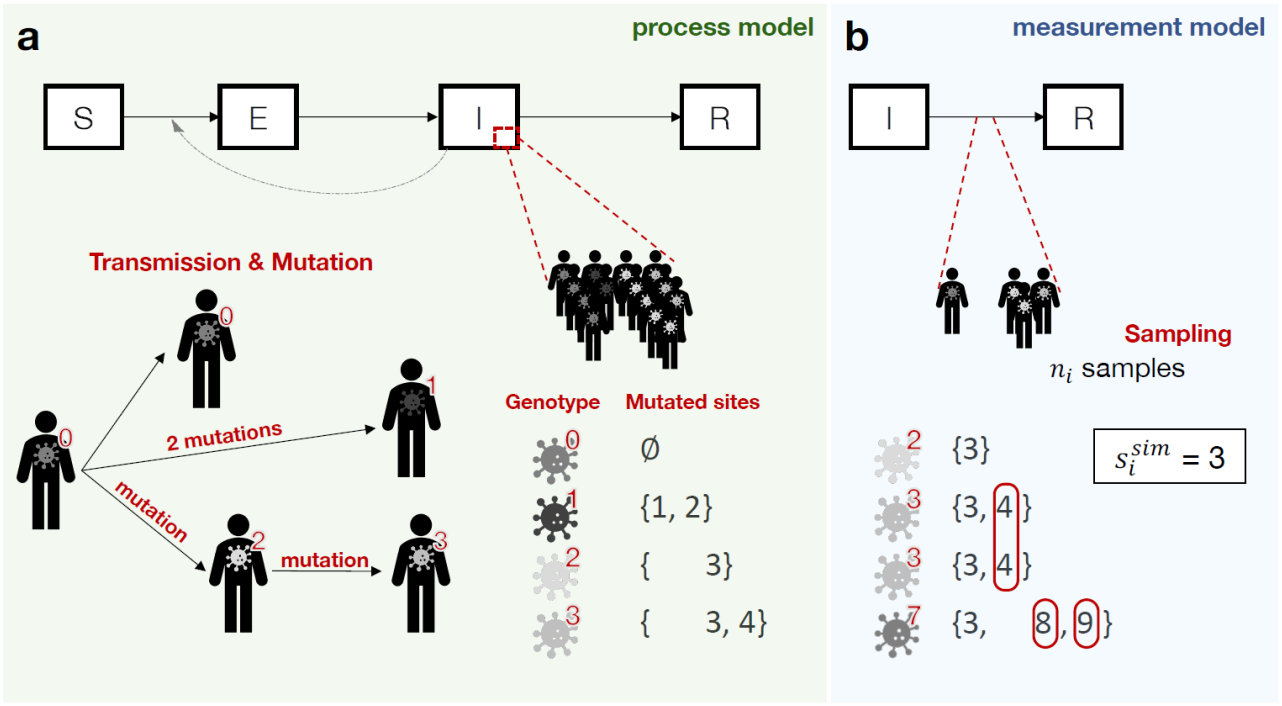

Figure S11. Depiction of the state space model. (a) Depiction of an epidemiological SEIR process model. The latent variables *X* in this process model are {*S*, *E*, *I*, and *R*}. Exposed and infected individuals are categorized by genotype, each of which has a unique set of mutations. Genotypes are shown as red-colored integers. A viral genotype of a donor is inherited by a recipient unless one or more mutations occur during the transmission event (bottom left). The occurrence of one or more mutations at transmission results in the recipient being infected with a new genotype, with this new genotype harboring the new mutation(s) as well as inheriting the existing set of mutations from the donor genotype. New mutations are numbered chronologically upwards from the current maximum mutation number. (b) Depiction of the measurement model. In addition to simulating the epidemiological dynamics specified by the process model, we keep track of the number of individuals of each genotype that have recovered during a given time window. In time window *i*, we then randomly sample n_i of these recovering individuals, where n_i denotes the number of viral samples that are binned in window *i* in the empirical data set. We calculate the number of segregating sites in this simulated sample of n_i sequences. Here, with $n_i = 4$, the number of segregating sites is $s_i^{sim} = 3$. These correspond to mutations 4, 8, and 9, because, for each of these sites, not all four sampled individuals carry the mutation. We repeat this process *k* times, with *k* 'grabs' of $n_i = 4$ recovering individuals and then calculate the mean number of segregating sites for that time window across these *k* grabs. Y_i in the state space model is given by this mean number of segregating sites in this time window *i*.

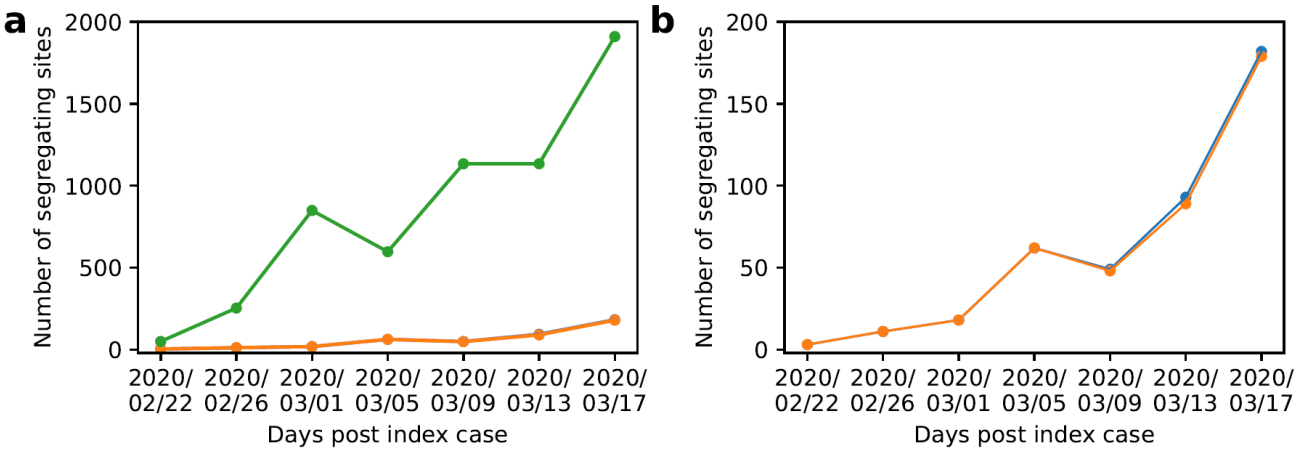

Figure S12. The effect of ambiguous nucleotides on the segregating site trajectory for France. (a) Segregating site trajectories calculated under different ambiguous nucleotide assumptions. The blue line shows the segregating site trajectory under the permissive assumption described in the main text (e.g., an *R* nucleotide matches both an *A* nucleotide and a *G* nucleotide). Under this assumption, an *N* nucleotide cannot increase the number of segregating sites observed in a time window, and an *R* nucleotide can only do so if the site at which it is found otherwise carries only *T*, *C*, and/or *Y* nucleotides. The green segregating site trajectory assumes that all ambiguous sites are mutations. Under this assumption, an *N* nucleotide increases the number of segregating sites observed in a time window unless that site already has all 4 nucleotides present, and an *R* nucleotide increases the number of segregating sites unless both an *A* and a *G* nucleotide are already present at that site. The orange segregating site trajectory assumes that *N* nucleotides match existing genetic variation but that other ambiguous nucleotides are considered as mutations. (b) Segregating site trajectories, as in panel (a), showing only the ambiguous nucleotide assumptions that correspond to the orange and blue lines in panel (a).

Table S1. Transmission pairs used to estimate the per-genome, per-transmission event mutation rate μ. Accession numbers of the consensus sequences from the donor and the recipient of the transmission pair are provided.

Study	Donor	Recipient	# SNPs
Popa et al. 2020	EPI_ISL_419656	EPI_ISL_437993	0
Popa et al. 2020	EPI_ISL_419656	EPI_ISL_437994	0
Popa et al. 2020	EPI_ISL_437994	EPI_ISL_437995	0
Popa et al. 2020	EPI_ISL_437994	EPI_ISL_438017	0
Popa et al. 2020	EPI_ISL_437994	EPI_ISL_438003	0
Popa et al. 2020	EPI_ISL_437994	EPI_ISL_438005	0
Popa et al. 2020	EPI_ISL_437994	EPI_ISL_437998	0
Popa et al. 2020	EPI_ISL_437994	EPI_ISL_583869	0
Popa et al. 2020	EPI_ISL_438005	EPI_ISL_438013	0
Popa et al. 2020	EPI_ISL_438005	EPI_ISL_438014	0
Popa et al. 2020	EPI_ISL_437998	EPI_ISL_438008	1
Popa et al. 2020	EPI_ISL_438008	EPI_ISL_438007	1
Popa et al. 2020	EPI_ISL_583869	EPI_ISL_438011	0
Popa et al. 2020	EPI_ISL_583869	EPI_ISL_583870	0
Popa et al. 2020	EPI_ISL_583869	EPI_ISL_438019	0
Popa et al. 2020	EPI_ISL_583869	EPI_ISL_438016	0
Popa et al. 2020	EPI_ISL_583869	EPI_ISL_583880	0
Popa et al. 2020	EPI_ISL_438016	EPI_ISL_438022	0
Popa et al. 2020	EPI_ISL_438022	EPI_ISL_438038	1
Popa et al. 2020	EPI_ISL_583870	EPI_ISL_438020	1
Popa et al. 2020	EPI_ISL_583870	EPI_ISL_438018	0
Popa et al. 2020	EPI_ISL_583870	EPI_ISL_583871	1
Popa et al. 2020	EPI_ISL_583870	EPI_ISL_475770	1
Popa et al. 2020	EPI_ISL_583871	EPI_ISL_583876	0
Popa et al. 2020	EPI_ISL_583871	EPI_ISL_583877	0
Popa et al. 2020	EPI_ISL_583871	EPI_ISL_583872	0
Popa et al. 2020	EPI_ISL_583872	EPI_ISL_583875	0
Popa et al. 2020	EPI_ISL_583872	EPI_ISL_583878	0
Popa et al. 2020	EPI_ISL_583875	EPI_ISL_438052	1
Popa et al. 2020	EPI_ISL_583875	EPI_ISL_438053	0
Popa et al. 2020	EPI_ISL_583875	EPI_ISL_438051	0
Braun et al. 2021	EPI_ISL_484813	EPI_ISL_484807	0
Braun et al. 2021	EPI_ISL_484919	EPI_ISL_484926	1
Braun et al. 2021	EPI_ISL_484919	EPI_ISL_484950	0
Braun et al. 2021	EPI_ISL_484926	EPI_ISL_484950	1
Braun et al. 2021	EPI_ISL_484921	EPI_ISL_484961	0
Braun et al. 2021	EPI_ISL_484921	EPI_ISL_484818	0
Braun et al. 2021	EPI_ISL_484961	EPI_ISL_484818	0
Braun et al. 2021	EPI_ISL_484952	EPI_ISL_484911	0
Braun et al. 2021	EPI_ISL_484973	EPI_ISL_484977	0
Braun et al. 2021	EPI_ISL_484976	EPI_ISL_495484	0
Braun et al. 2021	EPI_ISL_495461	EPI_ISL_509895	0
Braun et al. 2021	EPI_ISL_509876	EPI_ISL_509982	0
Braun et al. 2021	EPI_ISL_509876	EPI_ISL_509991	0
Braun et al. 2021	EPI_ISL_509876	EPI_ISL_509986	0
Braun et al. 2021	EPI_ISL_509982	EPI_ISL_509991	0
Braun et al. 2021	EPI_ISL_509982	EPI_ISL_509986	0
Braun et al. 2021	EPI_ISL_509991	EPI_ISL_509986	0
Braun et al. 2021	EPI_ISL_509897	EPI_ISL_509878	0
Braun et al. 2021	EPI_ISL_509897	EPI_ISL_509866	2
Braun et al. 2021	EPI_ISL_509878	EPI_ISL_509866	2
Braun et al. 2021	EPI_ISL_428254	EPI_ISL_428256	0
Braun et al. 2021	EPI_ISL_436627	EPI_ISL_436628	0
Braun et al. 2021	EPI_ISL_425176	EPI_ISL_427427	0
James et al. 2020	EPI_ISL_467433	EPI_ISL_467467	2
James et al. 2020	EPI_ISL_467433	EPI_ISL_467435	1
James et al. 2020	EPI_ISL_467433	EPI_ISL_467458	1
James et al. 2020	EPI_ISL_467446	EPI_ISL_467468	0
James et al. 2020	EPI_ISL_467446	EPI_ISL_467451	0
James et al. 2020	EPI_ISL_467444	EPI_ISL_467466	1
James et al. 2020	EPI_ISL_467444	EPI_ISL_467456	1
James et al. 2020	EPI_ISL_467444	EPI_ISL_467455	1

Popa et al. 2020	EPI_ISL_438051	EPI_ISL_438085	0	James et al. 2020	EPI_ISL_467444	EPI_ISL_467432	1
Popa et al. 2020	EPI_ISL_475770	EPI_ISL_583881	1	James et al. 2020	EPI_ISL_467444	EPI_ISL_467433	0
Popa et al. 2020	EPI_ISL_583880	EPI_ISL_438025	0	James et al. 2020	EPI_ISL_467444	EPI_ISL_467442	0
Popa et al. 2020	EPI_ISL_438039	EPI_ISL_438063	0	Lythgoe et al. 2021	NA	NA	0
Popa et al. 2020	EPI_ISL_438100	EPI_ISL_438098	1	Lythgoe et al. 2021	NA	NA	2
Popa et al. 2020	EPI_ISL_438035	EPI_ISL_438034	0	Lythgoe et al. 2021	NA	NA	1
Popa et al. 2020	EPI_ISL_438035	EPI_ISL_438036	0	Lythgoe et al. 2021	NA	NA	0
Popa et al. 2020	EPI_ISL_438035	EPI_ISL_438037	0	Lythgoe et al. 2021	NA	NA	0
Braun et al. 2021	EPI_ISL_421299	EPI_ISL_421306	0	Lythgoe et al. 2021	NA	NA	0
Braun et al. 2021	EPI_ISL_421323	EPI_ISL_421290	0	Lythgoe et al. 2021	NA	NA	1
Braun et al. 2021	EPI_ISL_421327	EPI_ISL_421319	0	Lythgoe et al. 2021	NA	NA	2
Braun et al. 2021	EPI_ISL_421328	EPI_ISL_421325	0	Lythgoe et al. 2021	NA	NA	0
Braun et al. 2021	EPI_ISL_421332	EPI_ISL_421287	0				

Sequence data downloaded from GISAID.

Data Availability

GISAID Identifier: EPI_SET_230123mt
doi: 10.55876/gis8.230123mt

All genome sequences and associated metadata in this dataset are published in GISAID's EpiCoV database. To view the contributors of each individual sequence with details such as accession number, Virus name, Collection date, Originating Lab and Submitting Lab and the list of Authors, visit 10.55876/gis8.230123mt

Data Snapshot

- EPI_SET_230123mt is composed of 13,963 individual genome sequences.
- The collection dates range from 2019-10-22 to 2020-11-02;
- Data were collected in 103 countries and territories;
- All sequences in this dataset are compared relative to hCoV-19/Wuhan/WIV04/2019 (WIV04), the official reference sequence employed by GISAID (EPI_ISL_402124). Learn more at https://gisaid.org/WIV04.

Supplementary References

1. Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution* **6**, veaa061 (2020).
2. Volz, E. M. & Siveroni, I. Bayesian phylodynamic inference with complex models. *PLoS Comput Biol* **14**, e1006546 (2018).
3. Le Vu, S. *et al.* Prevalence of SARS-CoV-2 antibodies in France: results from nationwide serological surveillance. *Nat Commun* **12**, 3025 (2021).

Chapter 3

Common misspecification of the generation interval leads to the underestimation of R in phylodynamic inference

3.1 Abstract

Generation intervals are distributions that describe the time between infection and onward transmission. They are a key epidemiological quantity because, together with the reproduction number R, they determine the population-level growth rate of a pathogen and its doubling time. Conversely, when fitting epidemiological models to data, assumed generation intervals impact R inference. This is well-known from studies that have used case data for R inference, with many studies emphasizing the importance of choosing an accurate distribution for the generation interval. In the phylodynamic inference of R, the generation interval distribution is often not explicitly mentioned, and the impact of generation interval misspecification has been studied less.

Here, we explore the impact of a commonly assumed (but rarely empirically accurate) exponential generation interval distribution on the estimation of R in phylodynamic inference. Using phylodynamic simulations and inference on these simulated datasets, we find that during the early exponential growth of an epidemic, if the generation interval is assumed to be exponentially distributed when it actually has a lower variance, then estimates of R will be biased low. Furthermore, uncertainty in the biased R estimates will be small. This underestimation under exponential distribution is largely restored by using the longer mean generation interval, suggesting that the underestimation could be explained by the $r - R$ relationship. Our work highlights the importance of acknowledging implicit generation interval assumptions in phylodynamic inference and points to the need for methodological developments in phylodynamic inference to provide greater flexibility in the specification of accurate generation intervals.

3.2 Introduction

The time between the infection of an individual and onward transmission from that individual is known as the generation interval (Svensson, 2007). Together with the reproduction number R, the generation interval is key in determining how fast an infectious disease will spread through a population: a disease with a shorter generation interval will spread more rapidly through a population than a disease with a longer generation interval, provided that they have the same supercritical R. Generation intervals also play an important role in R estimation. During the exponential growth of an early epidemic, reproduction numbers are generally calculated from intrinsic growth rates r, which could be inferred from case data to quantify the rate at which the number of incidence changes over time. The calculation of R from r depends on the relationship between r and R, which is a function of the generation interval (Wallinga and Lipsitch, 2007).

The generation interval between a donor-recipient pair is a specific, single length of time. However, there is inherent variation in the generation interval between transmission pairs as well as between a donor and their recipients in instances where the realized number of secondary infections exceeds one. To capture this variation, the generation interval of an infectious disease is generally described with a distribution rather than by a single value. Often, the empirical shape of the distribution is not known and is difficult to estimate as infection of an individual is rarely observed (Park et al., 2022).

In place of the empirical distribution, theoretical distributions are often used to model the generation interval. These include the gamma distribution, the Weibull distribution, and the log-normal distribution, all of which have only positive/non-negative support. This ensures that an individual does not transmit an infection prior to themselves becoming infected. The theoretical distributions that are most commonly used to model the generation interval often have high probability densities

around their mean values, such that the variance of the distribution is lower than its mean. This reflects observed biological characteristics of most infectious diseases, such as secondary infections rarely occurring immediately following a primary infection. Appropriately capturing the variation in the generation interval (rather than only the mean of this distribution) is crucial when estimating the reproduction number from the epidemic growth rate, as it contributes to the quantitative relationship between r and R (Wallinga and Lipsitch, 2007; Park et al., 2019).

Many models implicitly assume a generation interval distribution. In the classic susceptible-infectious-recovered (SIR) compartmental model, the generation interval is exponentially distributed, with the mean being the inverse of the rate of leaving the infectious compartment. More realistic generation interval distributions can be implemented by incorporating an additional compartment in the model that represents individuals who have been exposed (E) but have not yet become infectious. In these SEIR models, an infected individual stays in the exposed compartment before moving to the infectious compartment, delaying the time until a secondary infection can occur. This results in the generation interval distribution that is skewed toward later time-since-infection time points.

In practice, however, exponentially distributed generation intervals are often assumed, likely because of mathematical simplicity. These distributions have the highest probability density at small values and generally higher coefficients of variation than empirical generation interval distributions. The assumption of exponentially distributed generation interval (when the true generation interval distribution has a smaller coefficient of variation) is known to lead to lower estimates for R when using time series of case data for R inference (Wallinga and Lipsitch, 2007). This is because the intrinsic growth rate r is usually estimated first from case data, and the R corresponding to this growth rate is smaller under an assumed exponential distribution than the R corresponding to this growth rate under a generation interval

distribution with a smaller (< 1) coefficient of variation (Wallinga and Lipsitch, 2007).

While misspecification of the generation interval distribution is increasingly being recognized as introducing biases in epidemiological inference based on incidence data, little attention has focused on generation interval misspecification in phylodynamic inference. In phylodynamics inference approaches, the generation interval distribution is oftentimes implicitly assumed. For example, birth-death models implicitly assume an exponentially distributed generation interval because they assume a constant birth rate (corresponding to a constant transmission rate) and a constant death rate (corresponding to an exponentially distributed infectious period).

Here, we explore the impact of generation interval misspecification on phylodynamic inference of the reproduction number R. Because the most common assumption in phylodynamic inference is that the generation interval distribution is exponentially distributed, we focus specifically on the impact of this misspecification when the 'true' (that is, empirical) generation interval distribution has a smaller coefficient of variation of less than one. Our approach relies on the simulation of mock viral sequence datasets and the subsequent application of phylodynamic inference approaches to these mock datasets to quantify the biases in R estimation under the implicit assumption of an exponentially distributed generation interval.

3.3 Methods

3.3.1 Model structure for simulating mock datasets

We forward-simulated epidemiological dynamics using a susceptible-exposed-infectious-recovered (SEIR) model without susceptible depletion, which reduces to an exposed-infectious (EI) model (Figure 3.1A). In this model, a newly infected individual enters the exposed (E) compartment. An exposed individual either becomes infectious at rate γ or is sampled, which occurs at rate ψ. Here, we assume that the sampled

individual does not transmit the infection after being sampled for the pathogen genome and thus is removed from the exposed compartment. Individuals who transition to the infectious class (I) recover from this class at rate δ or are removed from this compartment through sampling at rate ψ. In this model, an individual stays in the exposed (E) and infectious (I) compartments for an average of L_E and L_I days, respectively. L_E is given by $\frac{1}{\gamma+\psi}$ and L_I is given by $\frac{1}{\delta+\psi}$. The generation interval of this model can be approximated as the convolution of two exponential distributions with means of L_E and L_I (Figure 3.1B). With this compartmental structure, the reproduction number R in the model is given by:

$$R = (\frac{\beta}{\delta + \psi})(\frac{\gamma}{\gamma + \psi}) \tag{3.1}$$

In our model, we assumed that the sampling occurs from both exposed and infectious compartments so that the assumption regarding the sampling process aligns with the assumption in the constant birth-death-sampling model. The sampling probability p_s is calculated as:

$$p_s = (\frac{\psi}{\gamma + \psi}) + (\frac{\gamma}{\gamma + \psi})(\frac{\psi}{\delta + \psi}) \tag{3.2}$$

The first term captures the probability that an individual who becomes infected is sampled while in the E compartment. The second term captures the probability that an individual who becomes infected is sampled while in the I compartment. The second term includes the factor $(\frac{\gamma}{\gamma+\psi})$ because not all individuals who become infected transition to the infectious class (some are sampled in the E class and therefore removed from being infected).

We simulated the epidemiological dynamics from this compartment model using Gillespie's tau-leap algorithm (Gillespie, 2001). For the rate of becoming infectious

(γ) and the rate of recovery (δ), we used 1/4 days^{-1} and 1/3 days^{-1}, respectively. For the sampling rate, we used $\psi = 0.0015$ days^{-1}, which, using equation (2), corresponds to the sampling probability p_s of approximately 1%. The per-capita transmission rate (β) was set to 1.01 days^{-1}, which, using equation (1), corresponds to a reproduction number of 3.0. Each simulation started with an index case in the infectious (I) compartment. Simulations were run for 40 days using a τ of one minute.

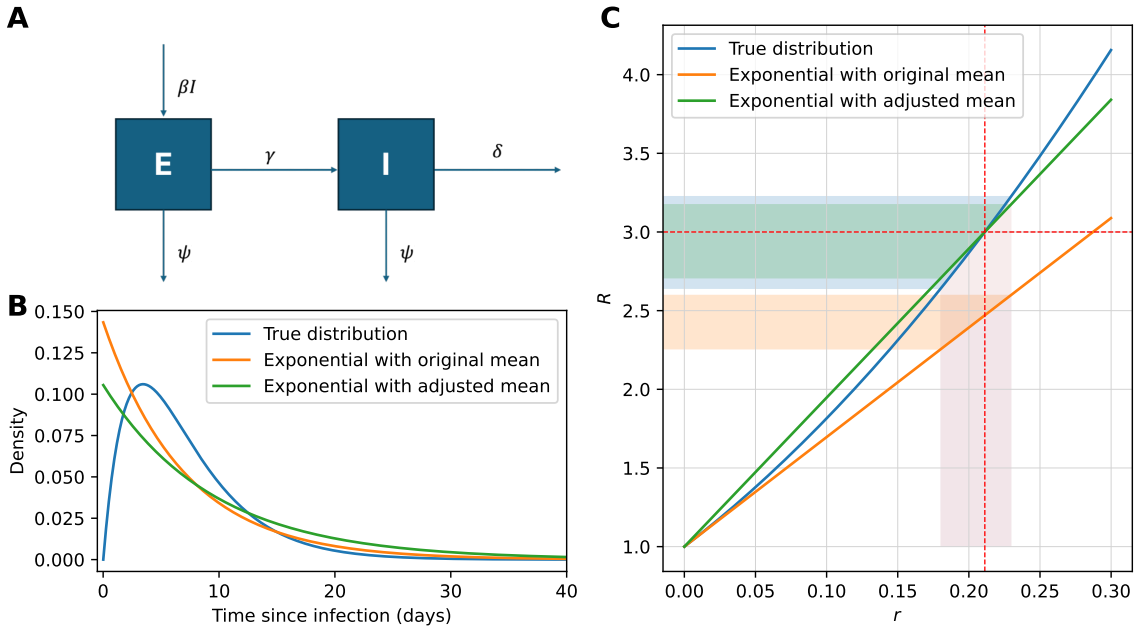

Figure 3.1: (A) Epidemiological model, consisting of exposed and infectious individuals. Individuals become infected by contact with infectious individuals and enter the exposed compartment. Individuals leave the exposed compartment by either transitioning to becoming infectious (at rate γ) or by being sampled (at rate ψ). Individuals become infectious by transitioning from the exposed class and become no longer infectious by recovering (at rate δ) or by being sampled (at rate ψ). (B) True and misspecified generation interval distributions. The true distribution is the generation interval distribution specified by the epidemiological model shown in panel A. Also, exponential distributions with the same mean (orange line) as the true distribution and adjusted mean (green line) that have the same r as the true distribution are shown. (C) The relationship between the reproduction number R and the intrinsic growth rate r under various generation interval distribution assumptions. The red dashed lines indicate true R and r values.

In our stochastic epidemiological simulations, we kept track of who-infected-whom in each simulation. To simulate the evolutionary and epidemiological dynamics, we

assigned a viral genotype to each infected individual. At the beginning of a simulation, the index case has a "reference" genotype with ancestral alleles only, and all other genotypes have a set of derived alleles relative to this reference genotype. Denoting genotypes by the set of derived alleles they carry, the reference genotype has an empty set ($G_{ref} = \emptyset$) of mutated sites. In comparison, genotype 1 (the first new genotype produced) may carry two mutated sites. These would be chronologically indexed, such that $G_1 = \{1, 2\}$.

Once an individual is newly infected, the individual inherits the mutated sites of the infector's genotype plus any additional mutations that occur during the transmission event. Additional mutations would result in a new genotype. We model mutation during the transmission event as a Poisson process where the number of mutations that occur during a transmission event is drawn from a Poisson distribution with mean p_m. We set the per-genome, per-transmission mutation rate p_m to 0.33 based on estimates from SARS-CoV-2 transmission pairs (Park et al., 2023). Since we consider viral evolutionary dynamics early on during an epidemic, we assume that mutations always occur at a new site rather than hitting the same site multiple times. We, therefore, adopt an infinite sites assumption. Finally, we assume that all mutations are fitness-neutral, consistent with assumptions made in the majority of phylodynamic inference approaches.

Once a simulation has finished, viral genotypes of sampled individuals are converted to nucleotide sequences for downstream BEAST analyses. Ancestral alleles for each site are first randomly chosen from among the four nucleotide bases ('A', 'T', 'G', and 'C'). The reference genotype is set as the genome carrying all ancestral alleles. Derived alleles are then randomly chosen from among the remaining three nucleotide bases at each site. If the viral genotype of an individual G_i has site j as an element, virus i then has the derived allele at site j. (If site j is not included in the viral genotype, then virus i has the ancestral allele at the site.)

3.3.2 Bayesian phylodynamic analyses

All phylodynamic analyses were performed using BEAST v. 2.7.5 (Bouckaert et al., 2019) using three different models: (1) the exponential growth coalescent model (Drummond et al., 2002), (2) the multi-type birth-death model implemented in the BDMM Prime package (Scire et al., 2022), (Vaughan and Stadler, 2024), and (3) the PhyDyn coalescent model (Volz and Siveroni, 2018). The BDMM Prime is the extended version of the original BDMM package (Kühnert et al., 2016), which is for multi-type birth-death models with migration (Kühnert et al., 2016). This model can incorporate the structured population, which could also be parameterized as a compartment model (see 3.3.2 for details) by considering each compartment as a "subpopulation" or a "deme" and transition between each compartment as "migration" (Kühnert et al., 2016). The PhyDyn model is a coalescent-based model that allows for complex population dynamics as well as population structure using generalized coalescent rates (Volz, 2012).

For each of these three models, we specified the same model of sequence evolution: the Jukes-Cantor substitution model (Jukes and Cantor, 1969) and a strict molecular clock. For each model, we calculated the reproduction number from the estimated parameters of the model. For example, for the exponential growth coalescent model, we calculated R from the exponential growth rate, and for the birth-death model and for PhyDyn, we calculated R from the estimated transmission rate. We obtained the median and 95% HPD of the calculated reproduction numbers using a custom script (available on `https://github.com/yspark576/misspecified_generation_interval`). All BEAST MCMC chains were run sufficiently long to result in effective sample size (ESS) values larger than 200 for the parameters of interest. ESS values were calculated using LogAnalyser within the BEAST2 package. Convergence of the MCMC chains was further assessed using Tracer v. 1.7.2 (Rambaut et al., 2018).

To investigate the impact of generation interval misspecification on phylodynamic

inference, we first performed each of these three phylodynamic analyses under different assumptions of the generation interval. First, to assess the bias introduced by a misspecified exponential distribution, we compared the estimation under (1) the true generation interval distribution and (2) the generation interval distribution that has the same mean but is exponentially distributed. We further explored the results under (3) an exponential distribution parameterized with a mean generation time that would reproduce the true intrinsic growth rate r of the epidemic (Figure 3.1C). We performed this latter analysis to determine whether a misspecified exponential distribution would yield unbiased estimates of R when the mean of this distribution was set to a value that would reproduce the correct epidemic growth rate.

Phylodynamic analysis under the coalescent exponential model

The coalescent model with exponential growth does not directly estimate the reproduction number R. Instead, the model estimates the intrinsic growth rate r. The estimate of r is then used to calculate the estimated R using the r-R relationship provided in (Wallinga and Lipsitch, 2007). This calculation requires the specification of the generation interval distribution. Under the 'true' generation interval distribution of our exposed-infectious (EI) compartment model (Figure 3.1), the r-R relationship is given by (Wallinga and Lipsitch, 2007):

$$R = (1 + rL_E)(1 + rL_I). \tag{3.3}$$

Given our δ, γ, and ψ, the L_E and L_I were 3.976 and 2.987 days, respectively, close to the values of 4.0 and 3.0 days, respectively, which would have been the case in the absence of sampling. We, therefore, use this relationship to calculate R from estimated r values under generation interval assumption (1) (the true distribution). Under an exponential generation time distribution, the relationship between r and R

is given by (Wallinga and Lipsitch, 2007):

$$R = (1 + r' L_G) \tag{3.4}$$

where r' is the growth rate under an exponentially distributed generation interval and L_G is the mean generation interval, which is given by $L_E + L_I = 6.963$. We use this relationship to calculate R from estimated r' values under generation interval assumption (2) (the exponential distribution parameterized with the true mean generation interval L_G). For generation interval assumption (3), we used the same equation (4) but with the mean generation interval adjusted to match the true intrinsic growth rate. This adjusted mean (L'_G) is calculated by setting $r = r'$, and solving for L_G. Under $R = 3$, the adjusted mean generation interval is calculated to be $L'_G = 9.470$ days.

Phylodynamic analysis under the birth-death model

The true generation interval from the EI model structure (assumption 1) was modeled using the multi-type birth-death model implemented in the BDMM-Prime package for BEAST 2. In this multi-type birth-death model, the exposed (E) and infectious (I) compartments were considered as separate "demes". Transmission is considered as a "birth" from deme I to deme E. Individuals who become infectious "migrate" from deme E to deme I, and recovery of an individual corresponds to a "death" from deme I. Consistent with the structure of our simulation model, individuals in demes E and I can be sampled. Individuals, once sampled, are removed from being considered infected.

We set the migration rate from deme E to I to γ, the death rate from deme I to δ, and the sampling rate from demes E and I to ψ. We estimated the birth rate from deme I to deme E, which corresponds to β, along with the time to the index case (T_0), which is the time between the last sample date and the infection of the

index case. The birth rate from deme I to deme E corresponds to the parameter β in Figure 3.1A. We used a uniform prior distribution for β with lower and upper bounds of 0.337 and 3.368 days^{-1}, respectively, corresponding to lower and upper bounds on R of 1 and 10, respectively. For the time to the index case (T_0), the lower bound for the uniform prior from 0 to infinity, where t_0 could be anytime before the last sample date. For the analyses, we further provided accurate information about the deme (E or I) from which each individual was sampled.

The exponentially distributed generation interval distributions (assumptions 2 and 3) were implemented using a single-type birth-death model by setting the number of demes to 1 in the BDMM-Prime model. In the single-type birth-death model, each individual stays infectious for $1/(\delta + \psi)$, and thus, the mean generation interval from the model is $1/(\delta + \psi)$. Also, the sampling proportion p_s is fixed to the same value used in the true model. Based on this, the death rate δ and the sampling rate ψ for the model were calculated from $p_s = \frac{\psi}{\psi+\delta}$ and $\psi + \delta = 1/L_G$, as $\psi = p_s/L_G$ and $\delta = (1 - p_s)/L_G$. These parameters were then fixed at these calculated values in the analyses.

For assumption (3), L'_G was used instead of L_G to calculate the death rate δ and the sampling rate ψ. As for assumption (1), we estimated the transmission rate β and the time to the index case (T_0). We used a uniform prior for β with lower and upper bounds, respectively, corresponding to R values of 1 and 10, using the equation $\beta = R(\delta + \psi)$. For the time to the index case, we used the same uniform prior as for assumption (1). Table 3.1 summarizes the priors used in the BDMM-Prime model under all three generation interval distribution assumptions.

Phylodynamic analysis under a coalescent model with complex dynamics

To incorporate the EI model under a structured coalescent framework, we used the PhyDyn package implemented in BEAST2 (Volz and Siveroni, 2018). As in the multi-

type birth-death model, under the assumption of true generation interval distribution, transmission is modeled as a birth event from deme I to deme E, the transition from exposed to infectious compartments is modeled as a migration event from deme E to deme I, and recovery from the infectious class is modeled as a death from deme I. Sampling events from compartments E and I are modeled as death events.

As in the birth-death analyses, the values of parameters ψ, γ, and δ were fixed at their true values, and we estimated β. For β, uniform priors with lower and upper bounds that correspond to the R of 1 and 10, respectively, were used. As before, the compartment from which each sampled individual was sampled was specified. The exponential generation interval distributions were implemented with a single infected compartment where individuals become no longer infectious, either through recovery or sampling, with rates $1/L_G$ and $1/L'_G$, respectively, for generation interval distribution assumptions (2) and (3). With the recovery rate fixed, the per-capita transmission rate β was estimated. Table 3.2 summarizes the priors used in the PhyDyn model under all three generation interval distribution assumptions. Note that sampling is not explicitly incorporated in PhyDyn models. It is considered as a removal from the compartment.

3.4 Results

Below, we first compare the estimated R under the true and misspecified exponential distribution to determine whether the misspecification leads to bias in phylodynamic inference (indicated with blue and orange in Figures 3.3, 3.4, and 3.5). In the following section, we further investigate whether the bias could be explained by the $r - R$ relationship through the exponential distribution with adjusted mean (indicated with green in Figures 3.3, 3.4, and 3.5).

3.4.1 R is systematically underestimated under a misspecified exponential distribution with true mean

Coalescent exponential growth model

We simulated 30 mock datasets and attempted to first estimate the intrinsic growth rate r using the coalescent exponential growth model implemented in BEAST2 (Figure 3.2) for each of these datasets. Coalescent inference with 6 of the 30 mock datasets did not reach convergence even after 24 million chains. These six datasets each had a small sample size and contained little genetic variation. Below, we therefore focus on the 24 simulations where effective sample sizes exceeded 200. As we generated our simulated datasets using the EI model with an R of 3.0, the exponential growth rate is expected to be 0.211 per day from Equation 3.3.

Of the 24 datasets, 22 had the true value of r contained within their 95% HPD (Figure 3.2A). The median estimated r values are slightly higher than the true value of $r = 0.211$ per day, with a bias toward higher median growth rates. This bias was statistically significant (one-sample Wilcoxon rank test, $p < 0.05$, Figure 3.2B) and is consistent with the push-of-the-past effect (Nee et al., 1994). (Stochastic simulations of an emerging epidemic can result in various outcomes, including extinction. However, we only considered the replicates with surviving epidemics, which tend to have a "flying start". As a result, the effective growth rate from those simulations is higher, which can lead to the overestimation under the deterministic assumption of the coalescent model.)

We converted these r estimates into R estimates using equation 3.3 for the true generation interval distribution and using equation 3.4 for the exponential generation interval distribution with the true mean L_G. Under the true generation interval distribution (assumption 1), the true R was included in the 95% HPD interval in 22 out of 24 datasets (Figure 3.3A). As expected, the dataset that failed to recover the

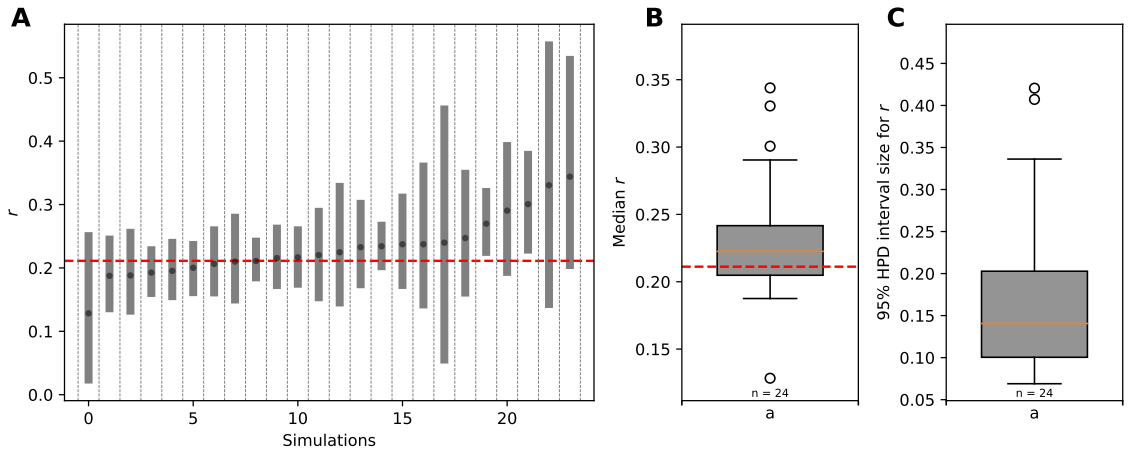

Figure 3.2: **Estimated growth rates (r) from the exponential-growth coalescent model.** (A) The median of the posterior distribution is indicated with dots, and the 95% HPD interval is shown as boxes surrounding the median. The dashed red line indicates the true r. Simulations are sorted by the estimated median of r. (B) Box plot for the median of the posterior distribution (C) Box plot for the size of 95% HPD interval.

true R was the dataset that failed to recover the true r. The median of the estimated R was slightly higher than the true value (one-sample Wilcoxon rank test, $p = 0.041$, Figure 3.3 B), reflecting the slight bias in estimates of growth rates.

Under the exponential distribution with the same mean as the true distribution, however, the median of the R_0 estimate was significantly lower than the true value (one-sample Wilcoxon rank test, $p < 0.001$), and the true R was included in the 95% HPD of only 12 of the 24 datasets. In the remaining datasets, the R was underestimated, and the 95% HPD failed to capture the true value. This suggests that even when the true r was recovered, R tends to be underestimated under a misspecified exponential distribution parameterized with the true mean. These results are consistent with findings based on case data (Wallinga and Lipsitch, 2007; Park et al., 2019). This underestimation could be explained by the higher growth rate under the exponentially distributed generation interval (Figure 3.1C). To match the estimated growth rate from the data, the model uses an exponential distribution assumption, which leads to a lower R estimate, thereby compensating for the distribution's naturally higher

growth dynamics.

In addition to the underestimation of R itself, the sizes of the 95% HPD interval estimates were also smaller under the misspecified exponential distribution compared to the true distribution (Mann-Whitney U rank test, $p < 0.001$, Figure 3.3C). This could also be explained by the $r - R$ relationship under different generation intervals (Figure 3.1C). Given the lower and upper bounds of the 95% HPD interval for r as (r_l, r_h), we can approximate the width of the 95% HPD interval for R by calculating $f(r_h) - f(r_l)$, since $R = f(r)$ is a monotonically increasing function. Since the derivative of $R = f(r)$ is greater for the true distribution compared to the exponential distribution, $f(r_h) - f(r_l)$ is greater for the true distribution, and thus the uncertainty is underestimated under the exponential distribution. The underestimate of uncertainty in estimates with a misspecified exponential distribution emphasizes that careful interpretation is crucial under the misspecified generation interval.

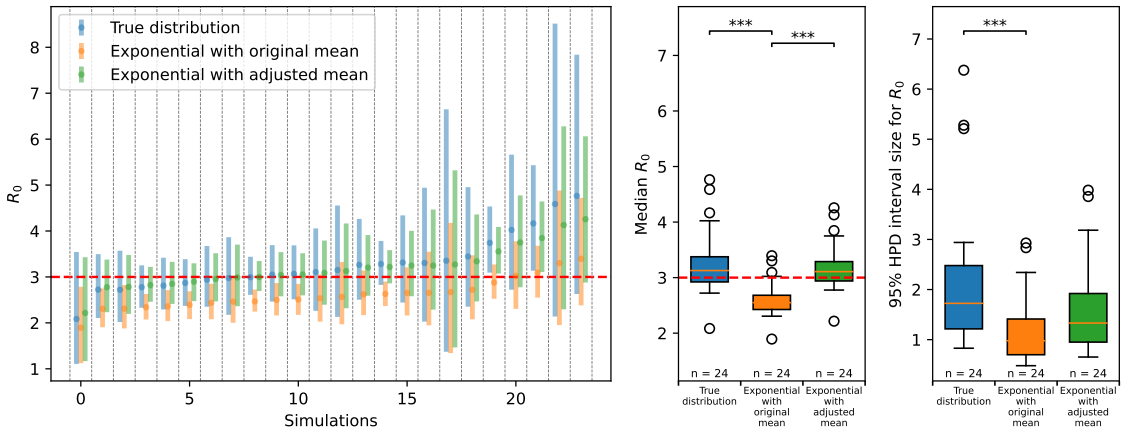

Figure 3.3: **Estimates of R from the exponential-growth coalescent model with different generation interval distributions.** (A) The median of the posterior distribution is indicated with dots, and the 95% HPD interval is shown as boxes surrounding the median. The dashed red line indicates the true R. Simulates are shown as ordered in Figure 3.2A. (B) Box plot for the median of the posterior distribution (C) Box plot for the size of 95% HPD interval. For (B) and (C), the asterisk indicates the p-value for the Mann-Whitney U rank test ($p < 0.001$ for '***,' $p < 0.01$ for '**,' and $p < 0.05$ for '*') comparing two distributions.

Birth-death model

Unlike in the exponential-growth coalescent model, where r is estimated and the generation interval distribution is only used thereafter to convert r to R, the birth-death model incorporates the generation interval distribution into the inference itself. Therefore, separate analyses were performed for each of the three generation interval distribution assumptions using the same simulated datasets. In Figure 3.4, we show the same data set as shown in Figure 3.3 for the exponential-growth coalescent model.

For the birth-death model, to obtain the estimation for R, we converted the sampled β into R using equation 3.3.1 and obtained the median and 95% HPD interval. Among the 24 datasets, 20 datasets had 95% HPD interval that successfully captured the true value (Figure 3.4A). The four datasets that failed to capture the true value underestimated the reproduction number. The median estimated R values were slightly lower than the true value (one-sample Wilcoxon rank test, $p = 0.047$) with a mean of 2.896 (Figure 3.4B).

Under the exponentially distributed generation interval, implemented as a single deme, the true values were captured in only 4 of the 24 datasets (Figure 3.4A). The median of the estimated R was significantly lower than the true value one-sample Wilcoxon rank test, $p < 0.0001$) with a mean of 2.412 (Figure 3.4B). This suggests that the R is underestimated even when the $r - R$ relationship was not used directly to estimate the reproduction number as in the exponential growth coalescent model. Even when the generation interval is indirectly assumed in the model, the generation interval distribution affects the estimation of the reproduction number.

As in the exponential growth coalescent model, the size of the 95% HPD interval was also smaller under the exponential distribution (Mann-Whitney U rank test, $p < 0.001$; Figure 3.4C), with a mean of 0.858 and 0.513 for the true and exponential distribution, respectively. This again indicates that misspecification of the generation interval distribution using an exponential distribution will bias estimates of R to be

low and that the uncertainty in these estimates will also be too low.

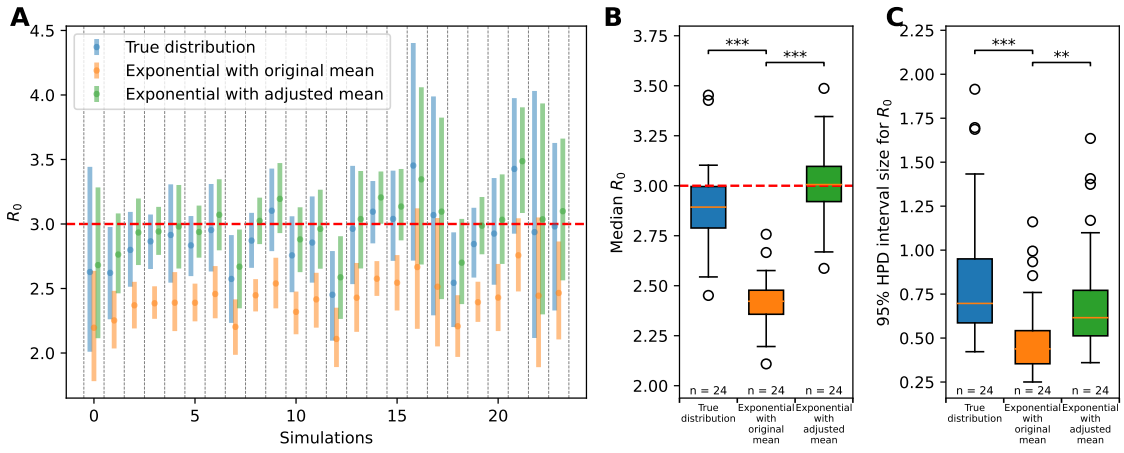

Figure 3.4: **Estimates of R from the BDMM model with different generation interval distributions.** (A) The median of the posterior distribution is indicated with dots, and the 95% HPD interval is shown as boxes surrounding the median. The dashed red line indicates the true R. Simulates are shown as ordered in Figure 3.2A. (B) Box plot for the median of the posterior distribution (C) Box plot for the size of 95% HPD interval. For (B) and (C), the asterisk indicates the p-value for the Mann-Whitney U rank test ($p < 0.001$ for '***,' $p < 0.01$ for '**,' and $p < 0.05$ for '*') comparing two distributions.

Coalescent model with complex dynamics (PhyDyn model)

Similarly to the birth-death model, the PhyDyn model incorporates the generation interval distribution into the inference itself. Separate analyses were, therefore, performed again for each of the three generation interval distribution assumptions using the same simulated datasets. Figure 3.5 shows the same 24 simulated datasets from Figures 3.3 and 3.4.

As in the BDMM model, we obtained samples of the transmission rate β from the MCMC chains and converted them to R to obtain the posterior distribution for R. Under the true distribution, the reproduction number was well estimated, in general, as expected. The PhyDyn analyses presented here are still preliminary, as ESS hasn't been reached in many replicates. Across the 24 datasets, the obtained 95% HPD interval captured the true value of R in 22 datasets (Figure 3.5A). The two datasets

overestimated the reproduction number. The medians of the posterior distribution were slightly higher, with a mean of 3.149 (Figure 3.5B), but was not significantly different from the true value (one-sample Wilcoxon rank test, $p = 0.08$).

Consistent with the results from the coalescent and birth-death models, the misspecified exponentially distributed generation interval parameterized with the true mean leads to an underestimation of the reproduction number. Only 11 datasets had their 95% HPD interval capturing the true value, and the underestimation of R was observed in 13 datasets (Figure 3.5A). The median was significantly lower than the true value (one-sample Wilcoxon rank test, $p < 0.001$), and the mean value was 2.580 across 24 datasets (Figure 3.5B). The size of the 95% HPD interval under the exponential distribution was significantly smaller (Mann-Whitney U rank test, $p < 0.001$; Figure 3.5C). This shows that the underestimation of R and the uncertainty in the estimates under exponential distribution were observed across tree models, suggesting the importance of the generation interval distribution in phylodynamic inferences.

3.4.2 Underestimation of R can be explained by the $R - r$ relationship

The analyses in Section 3.4.1 indicated that R was systematically underestimated when the distribution of the generation interval was misspecified with an exponential distribution rather than with the true distribution that has a smaller variance. Similar underestimation is also observed in case-based inferences relying on the growth rate to estimate R, as the misspecified variation affects the $r - R$ relationships (Park et al., 2019; Gostic et al., 2020). The exponential distribution, which has higher variation than the true distribution, has a higher density at a shorter generation interval, which implies that infections tend to occur earlier than in the true distribution (orange and blue lines in Figure 3.1B, but also see (Park et al., 2019)). These infections

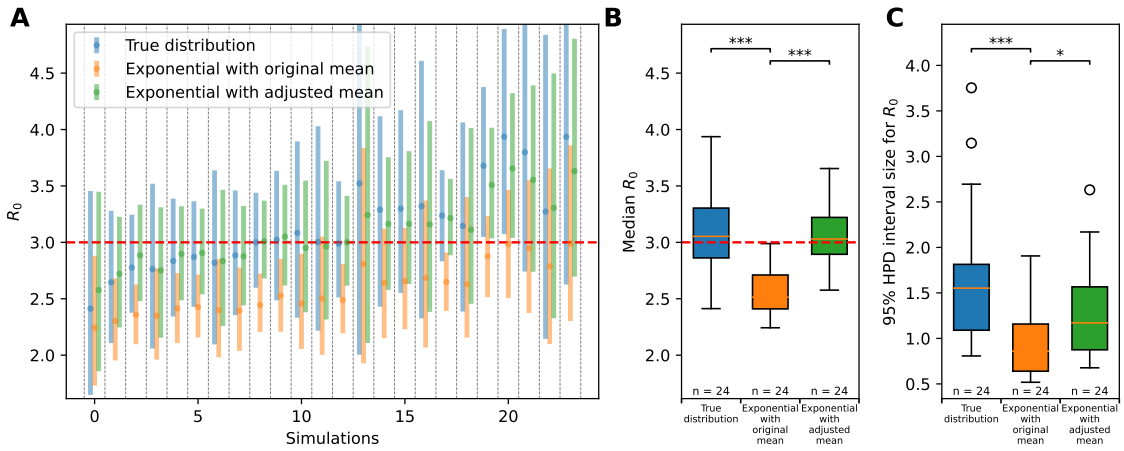

Figure 3.5: **Estimates of R from the PhyDyn model with different generation interval distributions.** (A) The median of the posterior distribution is indicated with dots, and the 95% HPD interval is shown as boxes surrounding the median. The dashed red line indicates the true R. Simulates are shown as ordered in Figure 3.2A. (B) Box plot for the median of the posterior distribution (C) Box plot for the size of 95% HPD interval. For (B) and (C), the asterisk indicates the p-value for the Mann-Whitney U rank test ($p < 0.001$ for '***,' $p < 0.01$ for '**,' and $p < 0.05$ for '*') comparing two distributions.

drive the faster growth of the epidemic and, therefore, lead to a higher growth rate given a reproduction number (Park et al., 2019). Namely, under the misspecified exponential distribution, the estimated r corresponds to a lower R compared to the true distribution with smaller variance (orange and blue lines in Figure 3.1C, but also see Park et al. (2019); Gostic et al. (2020); Diekmann and Heesterbeek (2000)).

In the following section, motivated by case-based inferences, we investigate whether the growth rate can explain the underestimation of R in phylodynamic inferences by performing phylodynamic analyses under a new exponential distribution that is expected to obtain the true growth rate given a true reproduction number (green line in Figure 3.1C). The new exponential distribution has a higher mean ($L_G^{'}$) of 9.470, equivalent to the true mean (L_G) of 6.963. The calculation of the adjusted mean is discussed in Section 3.3.2. If the growth rate is the main driver of the underestimation, estimates under the exponential distribution with the adjusted mean will perform better than those with the true mean.

Coalescent exponential model

We re-calculated the reproduction number from the estimated growth rate (3.2) assuming an exponential distribution with the adjusted mean ($L_G^{'}$). As expected, the underestimation observed with an exponential distribution parameterized with the same mean as the true distribution is no longer observed. Among the 24 analyzed datasets, the true R was recovered in 22 datasets, as under the true generation interval. The datasets that failed to recover the true R were the same datasets that failed to recover the true r and true R under the true generation interval distribution. The median of the posterior distribution of R was slightly higher than the true value, with a mean of 3.194 on average (one-sample Wilcoxon rank test, $p < 0.001$, Figure 3.3B).

The median of the estimated R was significantly different from that under the exponential distribution parameterized with the true mean (Mann-Whitney U rank test, $p < 0.001$) but was not significantly different from that under the true generation interval distribution (Mann-Whitney U rank test, $p > 0.05$). This further supports that the underestimation observed under exponential distribution with true mean is due to the $r - R$ relationship (Figure 3.1C). Unlike the exponential distribution with a true mean that has a higher growth rate, the new exponential distribution with an adjusted mean has a comparable growth rate, as the adjusted mean is longer than the true mean. Therefore, the new distribution does not lead to a lower R estimate to compensate for the higher growth, as observed under the exponential distribution using the true mean.

The size of the 95% HPD interval was not significantly different under the exponential distribution with true mean and under the exponential distribution with adjusted mean (Mann-Whitney U rank test, $p > 0.05$). However, the mean of the 95% HPD interval size was larger under the adjusted mean, with 1.603 and 1.178, respectively. Compared to the true distribution, the 95% HPD interval size was again comparable to those under the true distribution (Mann-Whitney U rank test, $p > 0.05$). However,

the mean of the interval size was slightly lower under the exponential distribution with adjusted mean, with the mean of 1.602, compared to the mean of 2.191 under the true distribution (Figure 3.3C).

Again, this can be explained with the slope for $R = f(r)$ under the true and exponential distribution (Figure 3.1C). The $R = f(r)$ function for the true distribution and the exponential distribution with the adjusted mean intersects at the true R. However, as $f(r)$ is a concave function, for r that is less than the true value, the R for the true distribution is lower than that for the exponential distribution with adjusted mean. Likewise, for r that is greater than the true value, the converted R is higher under the true distribution. Together, these lower R for $r < r_{true}$ and greater R for $r > r_{true}$ leads to the wider 95% HPD interval for the true distribution.

Birth-death model

Similarly, under the birth-death model, the exponential distribution with the longer mean resulted in higher R estimates that were closer to the true value. Across 24 datasets, the true R was recovered in 20 datasets (Figure 3.4A). Among the four datasets that failed to recover R, two underestimated the true value, and two overestimated the true value. The median of the posterior distribution of R was 2.991 on average, which was very close to the true value of 3.0 (one-sample Wilcoxon rank test, $p = 0.843$). The median estimates of R were significantly different from those under the exponential distribution with the original mean (Mann-Whitney U rank test, $p < 0.001$) but not significantly different from those under the true distribution (Mann-Whitney U rank test, $p > 0.05$; Figure 3.4B). As in the exponential growth coalescent model, using the adjusted mean could recover the true reproduction number, suggesting the importance of the growth rate in the estimation of the reproduction number. The size of the 95% HPD interval was also larger under the exponential distribution with the adjusted mean than those with the original mean (Mann-Whitney

U rank test, $p < 0.01$; Figure 3.4C). Compared to the true distribution, the 95% HPD interval size under the exponential distribution with adjusted mean is not significantly different. There was no significant difference between the 95% HPD interval size under the true distribution and exponential distribution with adjusted mean.

Unlike the coalescent exponential growth model, the birth-death model estimates the reproduction number directly without converting from the growth rate. To better understand whether the generation interval distribution affected the estimation of the growth rate, we further investigated the latent variables, in particular, the number of infected individuals over time. We found that the number of infected individuals increased at comparable rates, resulting in similar growth rate estimates despite differences in other parameters (Figure 3.6 upper panels). The underestimation of the reproduction number while the growth rates remain consistent across models suggests that the generation interval distribution affects the estimation of R primarily through the $r - R$ relationship, as in the coalescent exponential growth model.

Coalescent model with complex dynamics

Consistent with other models, the estimated R values were higher and closer to the true value if the adjusted mean ($L_G^{'}$) was used with the exponential distribution. The true R was recovered in 22 datasets, except for two datasets that overestimated R (Figure 3.5 A). The median of the R estimates was not significantly different from the true value (one-sample Wilcoxon rank test, $p = 0.16$) with a mean value of 3.08. The distribution of the median estimates from 24 datasets was higher than those from the exponential distribution with true mean (Mann-Whitney U rank test, $p < 0.001$; Figure 3.5B), but was not significantly different from those under true distribution (Mann-Whitney U rank test, $p > 0.05$). The size of the 95% HPD interval was larger than those from the exponential distribution with true mean (Mann-Whitney U rank test, $p < 0.05$; Figure 3.5C), but also was significantly lower than those from the

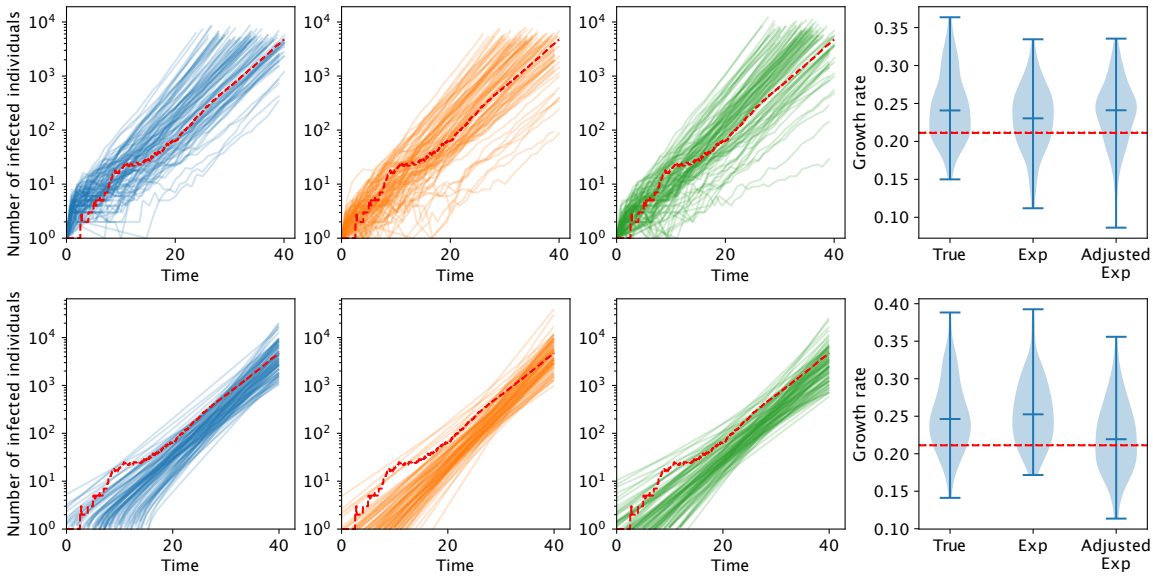

Figure 3.6: **Example trajectory of the number of infected individuals and growth rate.**
The number of infected individuals over time is sampled during MCMC chains for the birth-death model (upper panels) and coalescent with the complex dynamics model (lower panels) for a simulation replicate. Each line represents a trajectory from one chain, and 100 chains are shown. Red dashed lines show the true simulated trajectory. For the true distribution (shown in blue), the number of infected individuals is obtained from the number of individuals in the E and I compartments. For the exponential distribution with true and adjusted mean (shown in orange and green), the number of infected individuals is obtained from the number of individuals from the I compartment. The rightmost panels show the distribution growth rate calculated from the sampled trajectory. The vertical line inside the violin plot indicates the median of the distribution, and the red vertical line indicates the true growth rate calculated from the true reproduction number.

true distribution (Mann-Whitney U rank test, $p < 0.05$). Again, the underestimation of the R under the exponential distribution was not observed with a longer mean generation interval, which resulted in a comparable growth rate, suggesting that the underestimation is mainly governed by the $r - R$ relationship.

As in the birth-death model, the latent variables for epidemiological dynamics were investigated. The number of infected individuals over time showed similar patterns across generation interval distribution, and the growth rate estimated from each trajectory also showed similar distributions (Figure 3.6 lower panels). This further emphasizes the role of the $r - R$ relationship in the phylodynamic estimation of R.

3.5 Discussion

In the analysis of epidemiological case data, it is well known that the generation interval distribution shapes the relationship between the epidemic growth rate and the reproduction number (Wallinga and Lipsitch, 2007; Park et al., 2019). However, the impact of this relationship on phylodynamic analyses has been less well-studied. In this study, we focus on the impact that a commonly used but oftentimes misspecified, exponentially distributed generation interval distribution has on the estimation of R. We demonstrate that assuming an exponentially distributed generation interval when the true generation interval distribution has a smaller variance leads to a systematic underestimation of the reproduction number and erroneously high confidence in this underestimate.

Our results were consistent with those from case-based inferences. In case-based inferences, the growth rate can be estimated first from the incidence data and then used to calculate the reproduction number from the estimated growth rate (Wallinga and Lipsitch, 2007). In the exponential growth coalescent model, just as in case-based inferences, the growth rate is estimated first, and then the reproduction number is

estimated. Therefore, a similar bias is expected under the exponential growth coalescent model. However, even when the generation interval is incorporated through other model components in the BDMM and PhyDyn model, we observed the underestimation of the R under the exponential distribution. Furthermore, the true R was recovered when the growth rate was matched through the adjusted mean generation interval, even with the exponentially distributed generation interval.

In our findings, the exponential distribution with adjusted mean successfully recovered the true growth rate. Although this might suggest that the underestimation could be explained by matching the growth rate, we emphasize that this approach is not applicable to real-world data analysis. In our analyses, we could calculate the adjusted mean generation interval based on the true reproduction number. However, in the real world, we do not know the true reproduction number; it is the parameter we aim to estimate. Therefore, our results demonstrate the role of growth rate in the underestimation of the reproduction number rather than suggest a way to correct the bias in existing inference approaches.

Furthermore, we acknowledge that we do not know whether the tree shape under the exponential distribution with adjusted mean is close to that under the true distribution. However, tree shapes appear to be mainly driven by the number of tips in the tree (Plazzotta et al., 2016). In their study of tree shape under non-exponential infectious periods, the difference in tree shape features under different types of non-exponential infectious periods could be explained by the number of tips in the tree. This suggests that the exponential distribution with adjusted mean, which is expected to have a similar number of infected individuals since we matched the growth rate, may have similar tree shapes and features.

Despite its importance in inference, the generation interval is a rarely observed quantity, as it is very hard to know exactly when one is infected. Therefore, the exact distribution for the generation interval is also rarely known. However, (Park et al.,

2019) suggested that approximating the generation interval distribution as a gamma distribution could be parameterized based on the mean and standard deviation or through maximum-likelihood estimation. This provides a way to incorporate the generation interval into inference approaches and use the known $r - R$ relationship for gamma distribution (Wallinga and Lipsitch, 2007). In phylodynamic analyses, however, we are not aware of any package that allows for gamma-distributed generation intervals in inference. Although the exposed compartment can generate the gamma distribution when the duration in exposed and infectious compartments are similar, this represents very limited cases. Moreover, having multiple compartments can significantly slow down phylodynamic analyses, limiting its usability. Therefore, developing a flexible sequence-based inference approach that can incorporate more flexible generation interval distributions would be valuable for future research.

3.6 Supplementary information

Parameter	Prior	Unit
clockRate	Unif(0, Infinity)	
originBDMMPrime (T_0)	Unif(0, Infinity)	days
Assumption 1: True distn. from EI model		
birthRateAmongDemesCanonical.E_to_I (β)	Unif (0.3368, 3.3684)	days^{-1}
birthRateAmongDemesCanonical.I_to_E	Fixed at 0	days^{-1}
birthRateCanonical.E	Fixed at 0	days^{-1}
birthRateCanonical.I	Fixed at 0	days^{-1}
deathRateSPCanonical.E	Fixed at 0	days^{-1}
deathRateSPCanonical.I (δ)	Fixed at 1/3	days^{-1}
samplingRateSPCanonical (ψ)	Fixed at 0.0015	days^{-1}
removalProbCanonical	Fixed at 1	NA
migrationRateSPCanonical.E_to_I (γ)	Fixed at 1/4	days^{-1}
migrationRateSPCanonical.I_to_E	Fixed at 0	days^{-1}
startTypePriorProbs.E	Fixed at 0	NA
startTypePriorProbs.I	Fixed at 1	NA
Assumption 2: Exp. distn. with original mean		
birthRateCanonical (β)	Unif(0.1436, 1.4362)	days^{-1}
deathRateSPCanonical (δ)	Fixed at 0.1421	days^{-1}
samplingRateSPCanonical (ψ)	Fixed at 1.496×10^{-3}	days^{-1}
removalProbCanonical	Fixed at 1	NA
Assumption 3: Exp. distn. with adjusted mean		
birthRateCanonical (β)	Unif(1.056, 1.0559)	days^{-1}
deathRateSPCanonical (δ)	Fixed at 0.1045	days^{-1}
samplingRateSPCanonical (ψ)	Fixed at 1.100×10^{-3}	days^{-1}
removalProbCanonical	Fixed at 1	NA

Table 3.1: Parameters and priors for BDMM-Prime analyses

Parameter	**Prior**	**Unit**
clockRate	Unif(0, Infinity)	
Assumption 1: True distn. from EI model		
β	Unif (0.3368, 3.3684)	days^{-1}
γ	Fixed at 1/4	days^{-1}
δ	Fixed at 1/3	days^{-1}
ψ	Fixed at 0.0015	days^{-1}
E_0	Fixed at 0	
I_0	Fixed at 1	
Assumption 2: Exp. distn. with original mean		
β	Unif(0.1436, 1.4362)	days^{-1}
δ	Fixed at 0.1436	days^{-1}
I_0	Fixed at 1	
Assumption 3: Exp. distn. with adjusted mean		
β	Unif(0.1056, 1.0559)	days^{-1}
δ	Fixed at 0.1056	days^{-1}
I_0	Fixed at 1	

Table 3.2: Parameters and priors for PhyDyn analyses

Chapter 3 References

R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, and others. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

O. Diekmann and J. A. P. Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons, 2000.

A. J. Drummond, G. K. Nicholls, A. G. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320, 2002.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, 115(4):1716–1733, July 2001.

K. M. Gostic, L. McGough, E. B. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, R. Kahn, R. Niehus, J. A. Hay, P. M. De Salazar, J. Hellewell, S. Meakin, J. D. Munday, N. I. Bosse, K. Sherrat, R. N. Thompson, L. F. White, J. S. Huisman, J. Scire, S. Bonhoeffer, T. Stadler, J. Wallinga, S. Funk, M. Lipsitch, and S. Cobey. Practical considerations for measuring the effective reproductive number, Rt. *PLOS Computational Biology*, 16(12):e1008409, Dec. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008409. URL `http://dx.doi.org/10.1371/journal.pcbi.1008409`.

T. H. Jukes and C. R. Cantor. Evolution of Protein Molecules. *Mammalian Protein Metabolism*, pages 21–132, 1969. doi: 10.1016/b978-1-4832-3211-9.50009-7.

D. Kühnert, T. Stadler, T. G. Vaughan, and A. J. Drummond. Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Mol. Biol. Evol.*, 33(8):2102–2116, Aug. 2016.

S. Nee, E. C. Holmes, R. M. May, and P. H. Harvey. Extinction rates can be estimated from molecular phylogenies. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 344(1307): 77–82, Apr. 1994.

S. W. Park, D. Champredon, J. S. Weitz, and J. Dushoff. A practical generation-interval-based approach to inferring the strength of epidemics from their speed. *Epidemics*, 27:12–18, 2019.

S. W. Park, B. M. Bolker, S. Funk, C. J. E. Metcalf, J. S. Weitz, B. T. Grenfell, and J. Dushoff. The importance of the generation interval in investigating dynamics and control of new SARS-CoV-2 variants. *Journal of The Royal Society Interface*, 19(191):20220173, 2022.

Y. Park, M. A. Martin, and K. Koelle. Epidemiological inference for emerging viruses using segregating sites. *Nature Communications*, 14(1):3105, 2023.

G. Plazzotta, C. Kwan, M. Boyd, and C. Colijn. Effects of memory on the shapes of simple outbreak trees. *Scientific Reports*, 6(1), Feb. 2016. ISSN 2045-2322. doi: 10.1038/srep21159. URL `http://dx.doi.org/10.1038/srep21159`.

A. Rambaut, A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic biology*, 67 (5):901–904, 2018.

J. Scire, J. Barido-Sottani, D. Kühnert, T. G. Vaughan, and T. Stadler. Robust phylodynamic analysis of genetic sequencing data from structured populations. *Viruses*, 14(8):1648, July 2022.

A. Svensson. A note on generation times in epidemic models. *Math. Biosci.*, 208(1): 300–311, July 2007.

T. G. Vaughan and T. Stadler. Bayesian phylodynamic inference of multi-type population trajectories using genomic dat. Dec. 2024.

E. M. Volz. Complex population dynamics and the coalescent under neutrality. *Genetics*, 190(1):187–201, Jan. 2012.

E. M. Volz and I. Siveroni. Bayesian phylodynamic inference with complex models. *PLoS computational biology*, 14(11):e1006546, 2018.

J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2007.

Chapter 4

Epidemiologically clustered sequence in phylodynamic inferences

4.1 Abstract

The sample that represents the population is a key assumption in phylodynamic analyses. However, sequence datasets that are used in phylodynamic analyses oftentimes include epidemiologically clustered sequences. This is particularly likely during early epidemic growth of a virus or viral lineage when surveillance is targeted to an outbreak and when publicly available databases include sequences from household studies. These epidemiologically clustered sequences tend to be genetically highly similar to one another and thus may potentially bias sequence-based inferences of population-level growth rates. In this work, we investigate the bias introduced under the presence of epidemiologically clustered sequences in the phylodynamic estimation of the epidemic growth rate. We further evaluate various summary statistics that characterize genetic variation in randomly and non-randomly sampled populations to

determine their utility in detecting sampling bias, with the hope that these approaches could be applied to flag and correct for sampling bias prior to conducting downstream phylodynamic analyses. Single-dimensional statistics, such as the average pairwise nucleotide difference (π) and Watterson's θ, showed limited utility in identifying non-randomness. However, the distributions of pairwise tMRCA and nucleotide differences revealed distinctive patterns in non-randomly sampled datasets and appear to be more promising methods for detecting non-randomness.

4.2 Introduction

Viral genome sequences are increasingly used to quantitatively characterize the population dynamics of viral infectious diseases. With extensive genome sequencing efforts and well-established databases such as GISAID (Elbe and Buckland-Merrett, 2017) and NCBI GenBank that facilitated rapid sharing of pathogen genome sequences, there are more publicly available sequences than ever before. These data allow us to understand diverse aspects of disease dynamics (Martin et al., 2021), including the identification of newly emerging variants (Davies et al., 2021; Viana et al., 2022), the investigation of the introduction of a pathogen (Grubaugh et al., 2017), and transmission dynamics (Alpert et al., 2021). However, it is important to recognize the variation in the sampling efforts across and even within the sampled datasets, which can introduce biases unless well addressed. Genomic surveillance efforts for SARS-CoV-2 sequences varied across countries (Chen et al., 2022; Furuse, 2021; Brito et al., 2022), depending on the socioeconomic factors and the availability of sequencing laboratories (Brito et al., 2022). Sampling efforts also varied across time (Spott et al., 2024). Additionally, certain types of studies, for instance, studies focusing on household transmission or local outbreak clusters, may result in some epidemiological clusters being more closely related than others within a sequence dataset.

The importance of addressing sampling effort has been pointed out early on by (Frost et al., 2015), and numerous studies have investigated the effects of non-random sampling (Hall et al., 2016; Karcher et al., 2016). For example, Wohl et al. (2021) explored sampling schemes to reduce sampling bias and Parag et al. (2020) and Karcher et al. (2016) proposed models incorporating sampling processes in phylodynamics and phylogeography. However, existing studies have primarily addressed spatial and temporal non-random sampling.

Over-representation of epidemiologically clustered sequences can introduce significant biases in phylodynamic inference. While a certain degree of epidemiological clustering naturally emerges under random sampling conditions, some transmission clusters could be over-represented in a dataset from the aggregation of genome sequences from different types of studies, including household transmission or contact-trace studies. Even prior to the development of phylodynamic inference approaches, it was known that over-representation of a part of a population could bias summary statistics that quantify the extent of genetic (sequence) variation in a population (Tajima, 1995). Furthermore, studies have demonstrated that non-randomly sampled datasets with epidemiologically linked individuals can lead to an underestimation of effective population size and a failure to detect temporal population size changes in Bayesian skyline models (de Silva et al., 2012). Additionally, phylodynamic inference of subtrees or genetically closer individuals has been shown to underestimate effective population sizes under constant-size coalescent models as well as to underestimate exponential growth rates under exponential growth coalescent models (Dearlove et al., 2017).

Various approaches have been proposed to address biases introduced by epidemiologically clustered sequences. When possible, sequence datasets can be curated to exclude epidemiologically clustered sequences using alternative information sources such as news articles (Fraser et al., 2009). However, relevant metadata are often miss-

ing and unavailable, and privacy concerns restrict access to epidemiological information (Song et al., 2022). Furthermore, the increasing volume of sequences being deposited makes manual identification of clusters increasingly challenging. Some studies address these issues through the down-sampling of sequences.

Here, we first evaluate the bias introduced by epidemiologically clustered sequences on phylodynamic analysis using simulated datasets. For this, we simulated random and non-random datasets for comparison. We then analyze one-dimensional summary statistics quantifying genetic variation to identify potential signals of non-randomness in datasets. Furthermore, we explore higher-dimensional statistics to assess whether they provide additional informative signals for detecting non-random sampling.

4.3 Methods

4.3.1 Epidemiological and evolutionary simulations

To generate mock sequence datasets. We used a discrete-time branching process to simulate underlying epidemiological dynamics. Each epidemic starts with an index case at generation 0, and the population at generation 1 comprises all the individuals infected by the index case. Similarly, the population in Generation 2 comprises all the individuals infected by the individuals in Generation 1, and so on for each subsequent generation. The number of secondary cases from an infected individual in any generation is drawn from a negative binomial offspring distribution. The negative binomial distribution was parameterized using a mean (corresponding to the basic reproduction number R_0, set to 2.0) and an overdispersion parameter k. In most of our analyses, we set k to infinity, such that the offspring distribution became a Poisson distribution with mean R_0. In our analyses that focused on transmission heterogeneity, we considered a scenario with $k = 0.2$. In all of our simulations, we forward-simulated the branching process model for 12 generations and kept track of infector-infectee

relationships. As such, our simulations generated a full transmission history (Figure 4.1).

Each infected individual is characterized by the viral genome sequence they are infected with. The viral genome sequence of the index case was considered the reference genotype, and all other viral genome sequences were defined based on relative mutations compared to this reference genome. Viral mutations were assumed to occur at transmission due to the tight transmission bottleneck that characterizes acutely infecting viral pathogens such as SARS-CoV-2 and influenza A viruses (Martin and Koelle, 2021; Hannon et al., 2022; Li et al., 2022; McCrone et al., 2018). The number of new mutations that occur during transmission from a donor to an offspring is drawn from a Poisson distribution with mean p_m. We used a per transmission, per genome mutation rate of $p_m = 0.33$ for our simulation, based on estimates from SARS-CoV-2 transmission pairs (Park et al., 2023). Since this mutation probability is low and results in a limited amount of temporal signal in the simulated viral sequences, we also simulated a scenario with a higher $p_m = 2.0$. In all of our simulations, we assume infinite sites.

After the sampling process was simulated (see Section 4.3.1), simulated sequences were converted to a FASTA file where the ancestral alleles in the reference genome were randomly assigned to one of the four nucleotides (A, T, G, C) and the derived alleles were randomly assigned to one of the remaining three nucleotides. We first generated the reference sequence of 20kb, then randomly chose unique sites in this genome for each of the mutations that occurred, and then converted the ancestral allele to the derived allele for each simulated individual based on the mutations they carried.

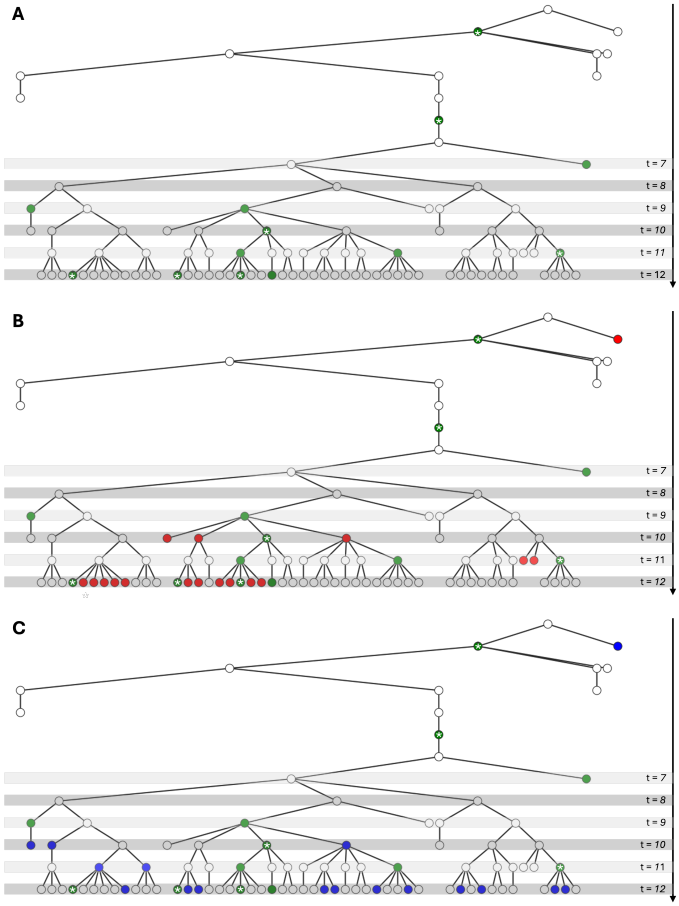

Figure 4.1: **A representative transmission tree from a forward simulation.** Each node represents an infected individual, and all individuals from a generation are horizontally aligned. Edges connecting two individuals represent the infector-infectee relationship, with the parent being the infector and the child being the infectee. Colored nodes represent sampled individuals. Panels (A-C) each show the same simulated transmission tree. Colored nodes comprise the viral sequence dataset. (A) Baseline dataset derived from the simulated transmission tree. Each individual was sampled with probability p_s. (B) Non-random sample (NS) dataset derived from the simulated transmission tree. A proportion of the individuals in the baseline dataset (green nodes) are contact-traced (marked with an asterisk inside the circles). All siblings of contact-traced individuals are sampled and added to the baseline dataset. (C) Random sample (RS) dataset derived from the simulated transmission tree. This dataset adds additional randomly sampled individuals to the baseline dataset to match the size of the non-random dataset in each generation.

4.3.2 Sampling of viral sequences from simulations

We generated two types of datasets for each simulation: a dataset that includes epidemiologically clustered individuals (non-random sample dataset; NS) and a dataset with only randomly sampled individuals (random sample dataset; RS). To generate these datasets, we first generated a baseline dataset, where we randomly sampled infected individuals with a sampling probability of $p_s = 0.02$ (Figure 4.1A). Under the transmission heterogeneity scenario (Figure 4.6), we sampled 0.5% of the total infected individuals, as "survived" epidemics tend to be larger than those without transmission heterogeneity

From this baseline dataset, the non-random sample dataset (NS) was first generated by adding epidemiologically clustered sequences (Figure 4.1B). To emulate a dataset with epidemiologically clustered sequences, we first chose a proportion p_c (here, 0.3) of sampled individuals to be contact-traced. If an individual is contact-traced, all of their "siblings" (that is, all of the individuals who have the same infector as the contact-traced individual) are included in the dataset.

We then also generated a random sample dataset (RS) that matched the non-random sample dataset in sample size by augmenting the baseline dataset with additional, randomly sampled individuals (Figure 4.1C). Specifically, for each generation, we obtained the number of added sequences in the NS dataset and sampled the same number of individuals randomly from that generation to the baseline dataset. The additional randomly sampled individuals were chosen from the subset of individuals not already sampled in the baseline dataset (so that no individual was sampled twice).

4.3.3 Summary statistics for characterizing the viral sequence datasets

We calculated summary statistics for each generation from the NS and RS datasets and compared these calculated summary statistics to determine whether they would be useful in identifying the presence of epidemiologically clustered sequences in viral datasets. We first considered classic population genetic summary statistics that quantify levels of genetic variation in a population: the average pairwise nucleotide difference and Watterson's θ. The average pairwise nucleotide difference was calculated from the number of different "alleles" between each possible pair of sequences. From n sequences, the average pairwise nucleotide difference π is calculated as follows:

$$\pi = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}$$

Watterson's θ is calculated from the number of segregating sites (i.e., the number of polymorphic sites) present in the sampled viral population. Since each site is more likely to be polymorphic when there are more sequences, the number of segregating sites depends on the number of samples. Therefore, to account for the sample size, Watterson's θ normalizes the number of segregating sites (S) using a correction factor (a_n):

$$\theta = \frac{S}{a_n}$$

where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$.

We also considered two higher-dimensional summary statistics. The first was the distribution of pairwise nucleotide differences, which is the fuller version of the one-dimensional average pairwise nucleotide difference described above. The second was the distribution of the pairwise time-to-most-recent-common-ancestor ($tMRCA$). This summary statistic was obtained by determining the time to the most recent

common ancestors for a pair of individuals (represented by the sequences). Since we know the true transmission history from the simulation, we calculated the exact value rather than inferring it from a reconstructed phylogenetic tree.

4.3.4 Assessment of bias in phylodynamic inference

We first evaluated bias in phylodynamic inference that may be introduced by epidemiologically clustered sequences. To this end, we estimated the exponential growth rate (r) under the coalescent model with exponential growth implemented in BEAST 2.7.5 (Bouckaert et al., 2019) for our NS and RS datasets. To perform BEAST analyses for multiple datasets, a template XML file was generated by modifying an example file generated by BEAUti program within the BEAST package. This template was used to generate an input XML file based on the simulated FASTA files. We assumed a JC69 nucleotide substitution model with no invariant sites and a strict clock model, consistent with the model of sequence evolution we used to generate the simulated sequence data. We used a uniform prior for the tree parameters, including the growth rate and the current population size.

For each dataset, we ran three independent MCMC chains, each chain having a length of more than 30 million states. We sampled parameters and trees in every 1,000 states. After discarding the first 10% of each chain as burn-in, we combined the chains using the "logcombiner" tool in the BEAST 2 package. We confirmed that the combined chain had an effective sample size (ESS) greater than 200 for the growth rate. Finally, summary statistics of the posterior distributions (e.g., mean, median, ESS, etc.) were calculated using "loganalyser."

4.4 Results

4.4.1 Characteristics of the simulated datasets

We simulated the above-described branching process until we obtained 200 independent replicates that did not go extinct within the first 12 generations. For each of these replicates, we generated non-random (NS) and random (RS) sequence datasets. The baseline dataset for each replicate included roughly 2% of the total infected individuals (Figure 4.5), as expected, given a sampling proportion of $p_s = 0.02$. However, because the replicates differed in the number of individuals that had become infected over the course of 12 generations, the absolute number of samples varied considerably across the 200 replicates, ranging from a minimum of 4 to a maximum of 769 samples Figure 4.5. The NS datasets have additional samples from contact-tracing each individual in the baseline dataset with a probability of $p_c = 0.3$ (Figure 4.5A). This additional sampling comprised roughly 1% of the total infected individuals (Figure 4.5B). By design, the RS datasets had the same number of sampled individuals as their corresponding NS datasets. The size of these "add-on" groups (in both the NS and RS datasets) ranged from a minimum of 0 to a maximum of 430 samples.

4.4.2 Phylodynamic inference

We estimated the intrinsic growth rate using the coalescent exponential growth model implemented in BEAST using 10 randomly chosen branching process simulations of the 200 we simulated. We managed to estimate growth rates for 8 out of these 10 simulations. Growth rate estimates from both the NS and RS viral datasets from these eight simulation replicates are shown in Figure 4.2. Because the true R_0 in our simulation was 2.0, the corresponding growth rate (r) is $ln(2.0) \approx 0.693$ per generation. The median estimate of r from the eight random sample (RS) datasets was 0.761 on average, and the 95% highest posterior density (HPD) included the true growth

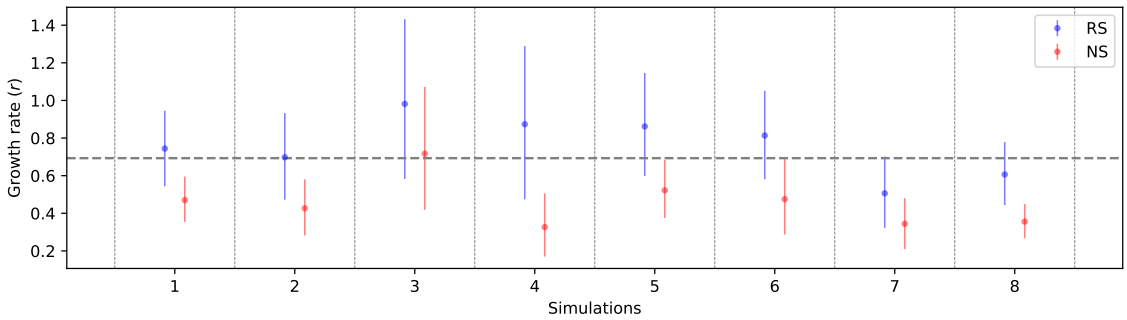

Figure 4.2: **Growth rates estimated under a coalescent exponential model using simulated datasets that do or do not contain epidemiologically clustered sequences**. Results on random sample datasets are shown in blue, and Results on non-random sample datasets are shown in red. Dots and error bars show the median and the 95% HPD of the posterior distribution for each dataset, respectively. The dotted line indicates the true growth rate, calculated from the basic reproduction number R_0 used in the forward simulations of the branching process model.

rate in each of these datasets. In contrast, the median estimate of r was 0.448 on average (corresponding to an R_0 of 1.57). Only two of the eight datasets had their 95% HPD interval capturing the true value. These findings indicate that the inclusion of epidemiologically clustered sequences can lead to considerable underestimation of the growth rate. This underestimation bias can be explained by contact-traced sequences being genetically more similar to one another than randomly sampled individuals within a generation, leading to shorter external branch lengths and, therefore, lower estimated growth rates.

4.4.3 Differences in one-dimensional summary statistics between random sample and non-random sample sequence datasets

Average pairwise nucleotide differences

Figure 4.3 compares the average pairwise nucleotide differences between the 200 random samples (RS) datasets and their matched 200 non-random samples (NS) datasets. In

this figure, we present comparisons on a per-generation basis, starting at generation 6, because earlier generations have fewer samples with which to calculate average pairwise nucleotide differences. When conditioned on the same underlying dynamics, average pairwise nucleotide differences in each generation were significantly lower in the NS datasets that epidemiologically clustered sequences than those of the RS datasets ($p < 0.05$ from both two-sided and one-sided paired sample t-test, Supplementary table 5.1). This is consistent with the expectation that epidemiologically clustered sequences tend to be genetically more similar, and thus, our expectation is that the average pairwise nucleotide difference between samples will be smaller.

Over generations, average pairwise nucleotide differences also increased in both the NS and the RS datasets as more mutations accumulated in the viral population. In later generations, the differences in the average pairwise nucleotide differences given the underlying dynamics became less apparent despite these differences still being statistically significant. The mean difference in the average pairwise nucleotide differences between the random sample and nonrandom sample datasets monotonically decreased from 1.035 average nucleotide differences in generation 6 to 0.061 average nucleotide differences in generation 12. This decrease can be explained by how the non-random sampling scheme is implemented. Because this sampling scheme contact-traces all siblings from 30% of the infected individuals that are in the baseline dataset, and because the expected number of siblings an infected individual has (which is $R_0 - 1 = 1$) does not change over the generations, the fraction of the sequences sampled in a given generation that derive from contact tracing is lower at higher generations. As such, non-random sample datasets at higher generations start to resemble those of the random sample datasets more closely, reducing the difference in this summary statistic at higher generations.

Although average pairwise nucleotide differences were significantly lower in the NS datasets than in the RS datasets when considered in a paired fashion (and

thereby conditioning on the same underlying epidemiological dynamics), in the real world, the underlying dynamics are rarely known. Therefore, to assess whether average pairwise nucleotide differences could inform the existence of epidemiologically clustered sequences, we need to determine whether the observed differences in the NS and the RS datasets differ from one another when they are considered in an unpaired fashion (such that we are not conditioning on the same underlying epidemiological conditions). We, therefore, compared the distribution of average pairwise nucleotide differences from the 200 simulation replicates between the NS and RS datasets. In the first three generations examined (generations 6 through 8), the distribution of the average pairwise nucleotide differences were significantly different from each other ($p < 0.05$, two-sample Kolmogorov-Smirnov test for goodness of fit). However, in the later generations, the distributions were no longer significantly different. Simulations at higher mutation rates p_m did not qualitatively change these results (Supplementary Figure 4.8). This suggests that average pairwise nucleotide differences are poor summary statistics to identify whether a dataset contains epidemiologically clustered samples that could bias phylodynamic estimates.

Watterson's theta

We next compared Watterson's θ across paired NS and RS datasets. Similar to our findings with average pairwise nucleotide differences, Watterson's θ was significantly lower in the NS datasets compared to the RS datasets ($p < 0.05$, both two-sided and one-sided paired sample t-test, Supplementary table 5.1; Figure 4.3). However, for Watterson's θ, the difference between random and nonrandom datasets increased over time, from 1.463 at generation 6 to 12.798 at generation 12.

To understand these results, recall that Watterson's θ quantifies genetic variation using the number of segregating sites across genome sequences. With a larger number of sampled sequences, each site in the genome has a higher probability of being observed

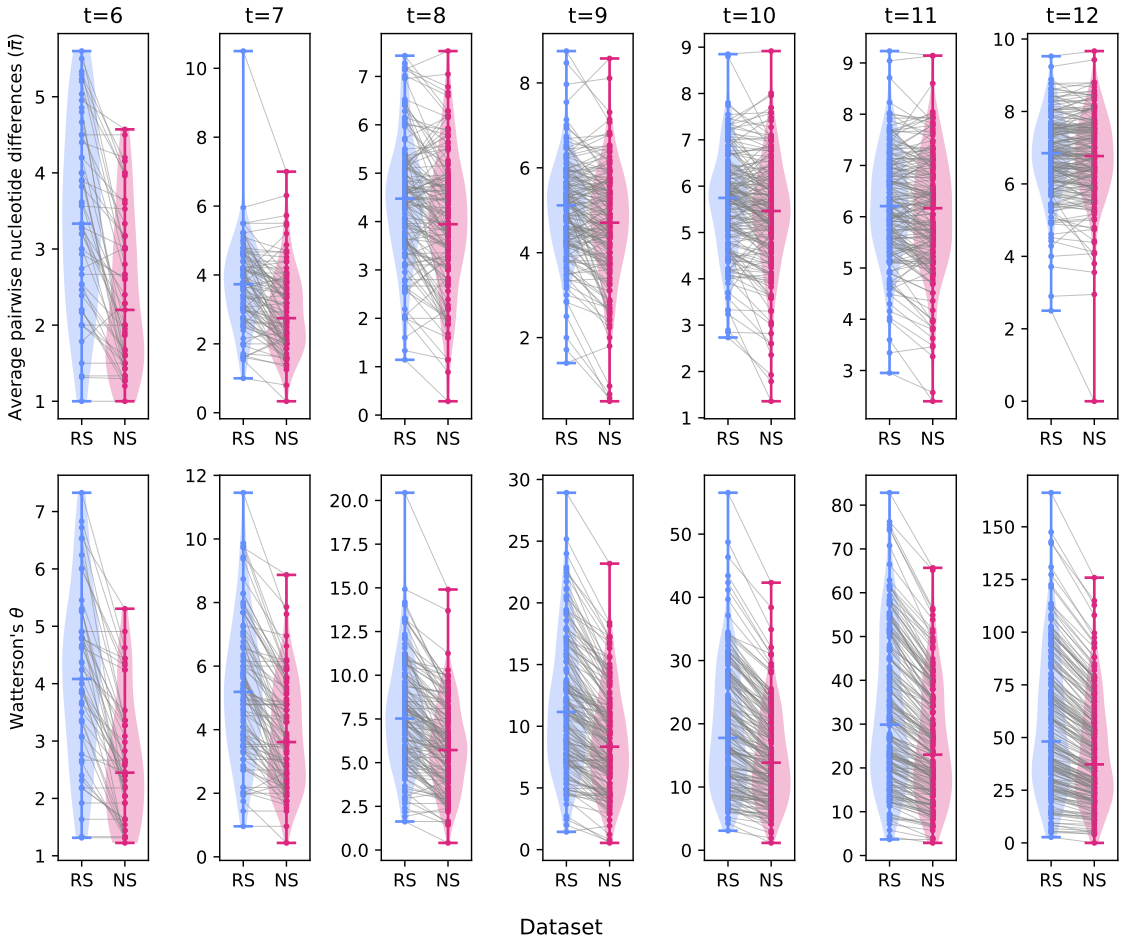

Figure 4.3: **Pairwise comparisons of summary statistics from random sample (RS) and non-random sample (NS) datasets by generation.** Average pairwise nucleotide differences (upper panels) and Watterson's theta (lower panels) were calculated from random and nonrandom datasets. Each dot denotes a summary statistic calculated from sequence samples in the generation indicated. Gray lines connect summary statistics calculated from datasets derived from the same simulation. Distribution of the summary statistics is shown as a violin plot, and the horizontal marker indicates the median of the distribution. Each column corresponds to a generation (t = 6 to t = 12).

as polymorphic. Therefore, in order to account for the sample size, Watterson's θ is obtained by normalizing the number of segregating sites with the correction factor (see equation 4.3.3). When the contact-traced sequences were added to the baseline dataset, these sequences added fewer segregating sites to the dataset than the samples added to the RS dataset. This is because the contact-traced sequences are genetically highly similar to the ones that are already included in the dataset (Supplementary Figure 4.9). This can explain the lower Watterson's θ in nonrandom samples. The reduced decrease of Watterson's θ in later generations could be explained by the increased sample size at higher generations.

We next compared the distribution of Watterson's θ from 200 simulation replicates in an unpaired fashion. Across all generations considered (generations 6 through 12), the two distributions were significantly different from each other ($p < 0.05$, two-sample Kolmogorov-Smirnov test for goodness of fit). This finding suggests that Watterson's θ is likely a more promising statistic than the average pairwise nucleotide difference for identifying when datasets may contain epidemiologically clustered sequences. In real-world situations, however, we do not have the distribution of Watterson's θ from a dataset and similar patterns are observed under a higher mutation rate (Supplementary 4.8). Instead, we only obtain a single Watterson's θ for sequences in a given generation (or block of time). This might hinder the use of one-dimensional summary statistics to determine if a dataset has epidemiologically clustered sequences that could bias phylodynamic analysis.

4.4.4 Differences in multi-dimensional summary statistics between random sample and non-random sample sequence datasets

Distribution of pairwise nucleotide differences

Distributions of pairwise nucleotide differences are shown in Figure 4.4 for both the RS and NS datasets. The distribution of pairwise nucleotide differences from the RS datasets had higher densities around the distribution's median value rather than at the upper bound, especially during later generations. In our model, the mutation occurs during a transmission, and the number of mutations occurring during a transmission follows the Poisson distribution. The number of transmissions between two individuals is the tMRCA. Therefore, the distribution of pairwise nucleotide differences could be obtained by summing the number of mutations by tMRCA times. Also, for both random and nonrandom datasets, the distribution expands to higher values over generations, as expected from the accumulated genetic variation in the viral population.

In contrast to the distributions observed for the RS datasets, the distributions of pairwise nucleotide differences in the NS datasets have higher densities at low pairwise nucleotide difference values. This is expected, especially during the earlier generations, where epidemiologically clustered sequences comprise a higher proportion of the samples. Under a higher mutation rate p_m, higher densities at low pairwise nucleotide differences were observed more clearly (Supplementary Figure 4.12). This is partly due to the distribution expanding toward larger values, which separates the main peak from the peak of epidemiologically clustered sequence pairs. Under the transmission heterogeneity, the increase in the density was higher, although the peak was not separated from the main peak.

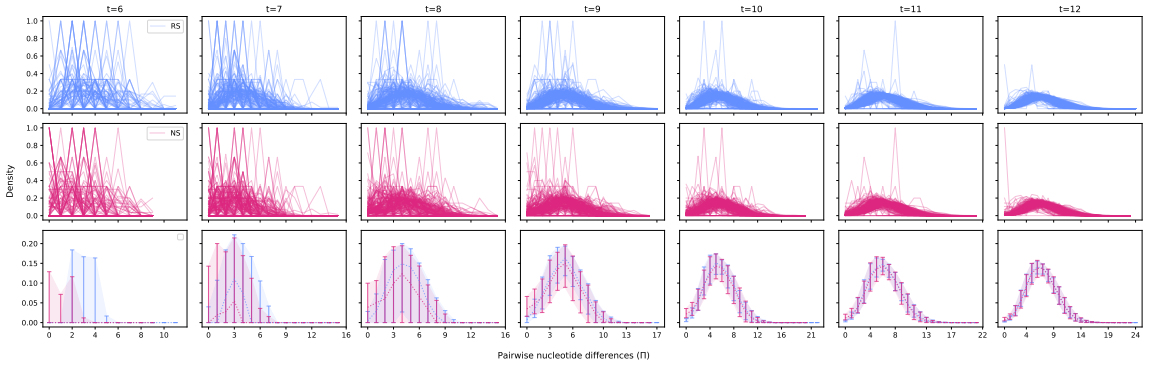

Figure 4.4: **Comparison of pairwise nucleotide difference distributions between random and nonrandom datasets, by generation.** Pairwise nucleotide diversity values were obtained between every pair of samples in the RS dataset and in the NS dataset by generation. Distributions of these pairwise tMRCAs are shown in the top row for the RS dataset and in the middle row for the NS dataset, with columns corresponding to generations. Each line shows the distribution from a single dataset in a single generation. The bottom row summarizes the individual distributions shown in the top and middle rows. The dotted lines show the median. The error bars show the 25% and 75% percentiles.

4.4.5 Transmission heterogeneity

We further explored the ability to detect datasets with epidemiologically clustered sequences under a scenario of high transmission heterogeneity (overdispersion parameter $k = 0.2$). Except for two replicates with no samples, the sample size ranged between 1 and 1820, showing considerable variation in the sample size. The add-on sample size also varied widely, from 0 to 7104 individuals, likely because some individuals with many siblings were chosen for the contact tracing.

As in the baseline scenario, the average pairwise nucleotide difference was significantly lower in the nonrandom dataset across all generations ($p < 0.05$ from both two-sided and one-sided paired sample t-test, Supplementary table 5.1; Supplementary Figure 4.7). The difference in the $\bar{\pi}$ between random and nonrandom datasets decreased at higher generations, as in our findings without transmission heterogeneity. However, the decrease in the transmission heterogeneity simulations was more pronounced than that in the previous simulations. However, unlike in our analyses of

the no-transmission heterogeneity simulations, the overall distributions of $\bar{\pi}$ from 200 replicates were significantly different between the NS and RS datasets, even in the later generations ($p < 0.05$, two-sample Kolmogorov-Smirnov test for all generations).

Watterson's theta calculated from the transmission heterogeneity simulations was significantly lower for the NS datasets than the RS datasets (Supplementary figure 4.7), consistent with our findings with the no-transmission heterogeneity simulations. The overall distribution of Watterson's θ from 200 replicates were significantly different from each other ($p < 0.05$, two-sample Kolmogorov-Smirnov test for all generations).

For the distribution of the pairwise tMRCA (Supplementary figure 4.11), the local peak at $t = 1$ had a very high density. This local peak remained visible even in later generations and lowered the densities at higher tMRCA values. This pattern was also observed in the distribution of the pairwise nucleotide difference (Supplementary figure 4.13).

The larger reduction in the one-dimensional summary statistics and higher pairwise tMRCA density at $t = 1$ in the transmission heterogeneity simulations compared to the no-transmission heterogeneity simulations can be explained by the number of contact-traced samples in the NS dataset under transmission heterogeneity. Under transmission heterogeneity, most of the individuals have a small number of "siblings", while a few have a very large number of siblings. If an individual from a larger "family" is contact-traced, this family cluster is larger than those from scenarios without transmission heterogeneity, resulting in more pairs that have lower genetic diversity and smaller tMRCA.

4.5 Discussion

In this study, we generated viral datasets that contained epidemiologically clustered sequences along with matched datasets that contained only randomly sampled sequences.

Using these datasets, we explored how epidemiologically clustered sequences may bias phylodynamic estimates of the epidemic growth rate. Furthermore, we explored the utility of one-dimensional and multi-dimensional summary statistics to evaluate their ability to identify datasets that may contain epidemiologically clustered sequences prior to downstream phylodynamic analyses.

According to Waples and Anderson (2017), a truly random sample is achieved when each individual is sampled with equal probability and independently of other individuals. However, non-random sampling can have various forms. In de Silva et al. (2012), non-random sampling was implemented by sampling the first individual along a lineage. Dearlove et al. (2017) focused on random subtrees and false clusters based on genetic distance, where a subtree is obtained by getting descendants of a randomly chosen individual and the false cluster was obtained by selecting sequences if branch lengths between each other are below a threshold. Here, we simulated non-random sample datasets as a combination of randomly sampled individuals and epidemiologically clustered individuals who are siblings of the randomly sampled individuals. This sampling methodology was designed to emulate scenarios where some of the sampled individuals are contact-traced and epidemiologically linked sequences are further added to the dataset or scenarios where publicly downloaded sequences contain clustered individuals.

Our findings demonstrate that the presence of epidemiologically clustered sequences can result in an underestimation of the epidemiological growth rate. Similar biases have been observed in nonparametric models of Bayesian skyline (de Silva et al., 2012). With more non-random samples, the estimation failed to capture the growth of the epidemic and underestimated the current effective population size. This study further found that higher proportions of non-randomly sampled individuals led to greater underestimation of the effective population size. (Dearlove et al., 2017) further demonstrated that when only clustered sequences were considered, underestimation

of growth rate and effective population size occurred under both exponential growth coalescent and constant size coalescent models.

To mitigate bias from clustered sequences, researchers have employed various strategies, including down-sampling sequences (Hedge et al., 2013; Rambaut and Holmes, 2009), or excluding all but one sequence from known epidemiological clusters (Fraser et al., 2009). However, considering the increasing numbers of sequences deposited from diverse studies, including contact-traced studies and household studies (Hare et al., 2021), accessing comprehensive metadata may not always be feasible. Consequently, we explored whether summary statistics could be useful to identify when a dataset may contain epidemiologically clustered sequences.

Our initial analyses focused on one-dimensional summary statistics derived from our simulated datasets. These statistics could be used for hypothesis testing for random sampling using empirical p-values to detect non-randomness from real-world datasets. Such p-values can be derived through the simulation of multiple replicates and counting those that yield summary statistics greater than or equal to the observed value (North et al., 2002). Our analyses reveal that the distribution of the π is significantly different between random and non-random datasets only in earlier generations, suggesting a limited utility as a tool to detect the non-randomness. On the other hand, Watterson's θ exhibits significant differences between random and non-random datasets, suggesting its potential utility as a metric for hypothesis testing. However, since the distribution from non-random datasets is nested within that of random datasets, observed statistics are unlikely to yield p-values below 0.05, even in the presence of non-random sampling.

The multi-dimensional summary statistics we calculated and used to compare the RS and NS datasets revealed more apparent differences between these datasets. Specifically, distributions of pairwise nucleotide differences showed increased density at lower values. While this pattern becomes less clear in later generations, the existence of pairs with lower nucleotide differences may serve as an indicator of non-

randomness in the dataset, aligning with genetic distance-based cluster identification approaches (Wertheim et al., 2013). Although the distribution of pairwise tMRCA also showed differences between random and non-random datasets, it is important to note that our analysis calculated pairwise tMRCAs using true transmission trees rather than trees inferred from sequence data. In real-world applications, transmission tree reconstruction from sequence data would introduce additional noise, likely diminishing any signals that could indicate that a dataset contains epidemiologically clustered sequences.

While we calculated the summary statistics from simulated datasets, we calculated the summary statistics for each generation rather than aggregating them. However, the generation of sequences is not known, only their collection date. We used generations rather than collection data to examine the usability of the summary statistics under optimal conditions, even if unrealistic. Additionally, we assumed that the true epidemiological dynamic parameters were known, which is another unrealistic assumption. Our results indicate that early epidemic stochasticity might limit the ability to reliably detect when a dataset contains non-random samples, suggesting the need for more parametric approaches, as proposed for clustering approaches (Poon, 2016).

4.6 Supplementary information

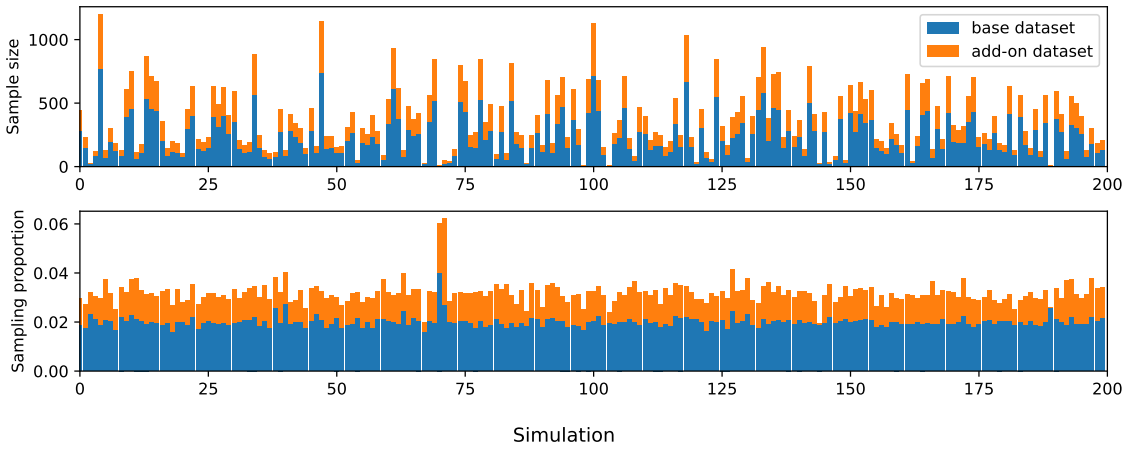

Figure 4.5: **Sample size (A) and realized fraction of sampled individuals (B) included the simulated datasets.**

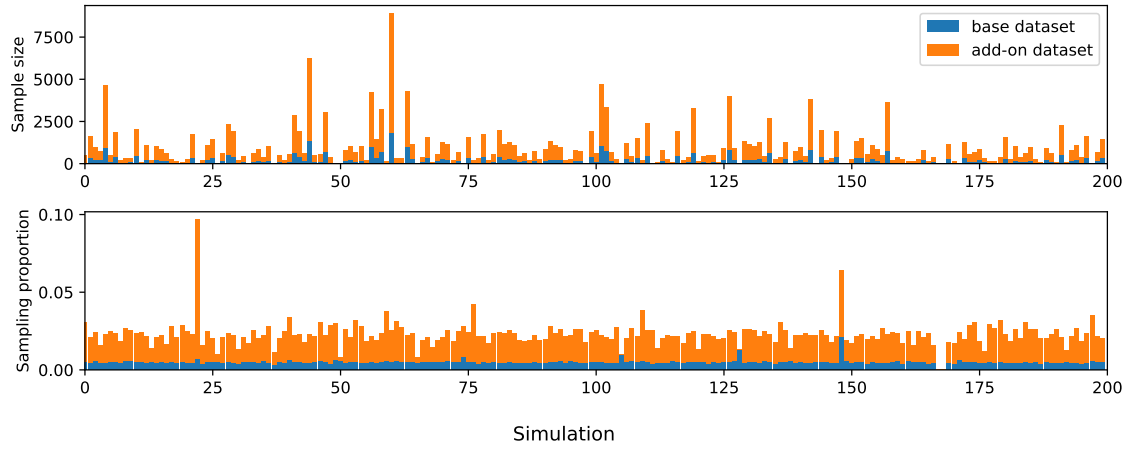

Figure 4.6: **Sample size (A) and realized fraction of sampled individuals (B) included the simulated datasets with transmission heterogeneity.**

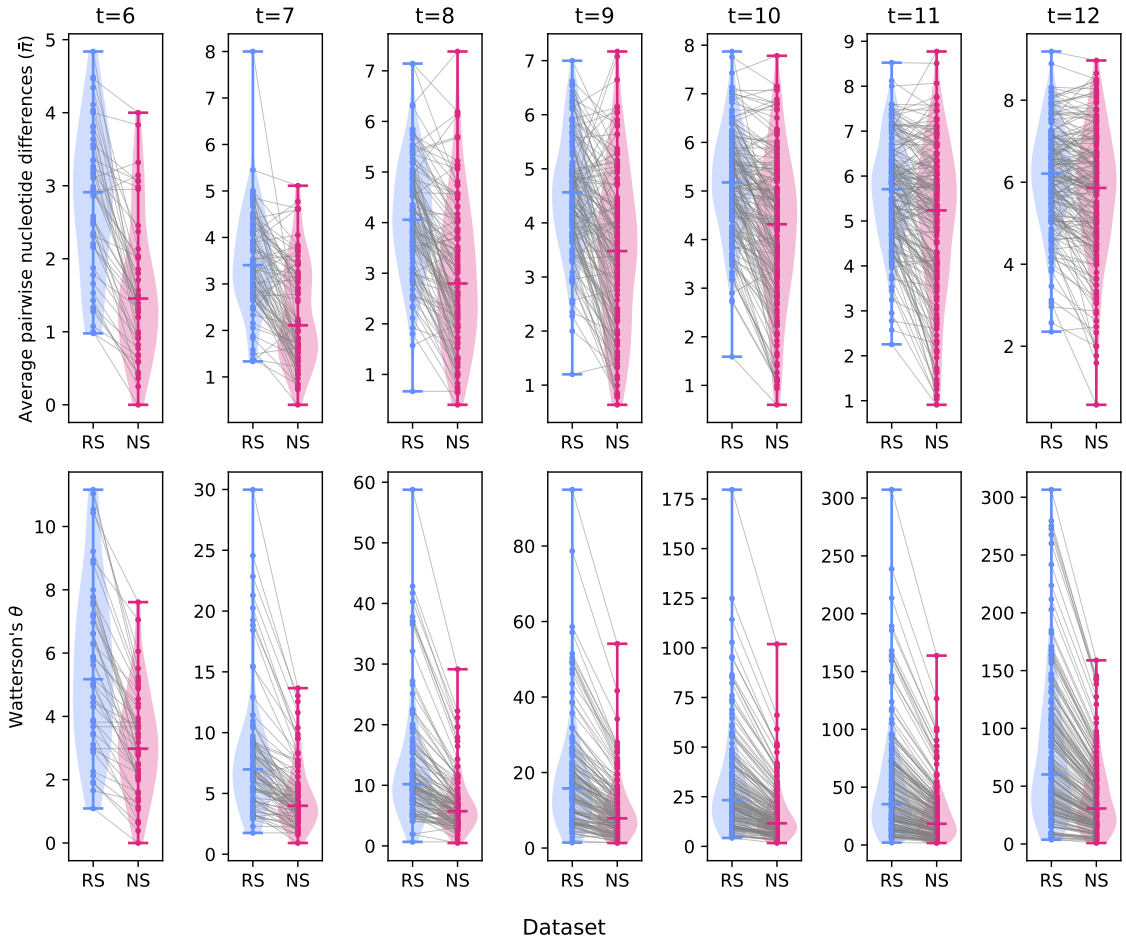

Figure 4.7: **Comparison of the summary statistics from random and non-random datasets over generations under transmission heterogeneity with** $k = 0.2$**.** The average pairwise nucleotide differences (upper panels) and Watterson's θ (lower panels) were calculated from random and nonrandom datasets. Each dots represent the calculated summary statistics from a dataset at each generation. The lines connect the datasets from the same simulation. The distribution of the summary statistics is shown as a violin plot, and the horizontal marker indicates the median of the distribution. Each column represents a generation from t = 6 to t = 12

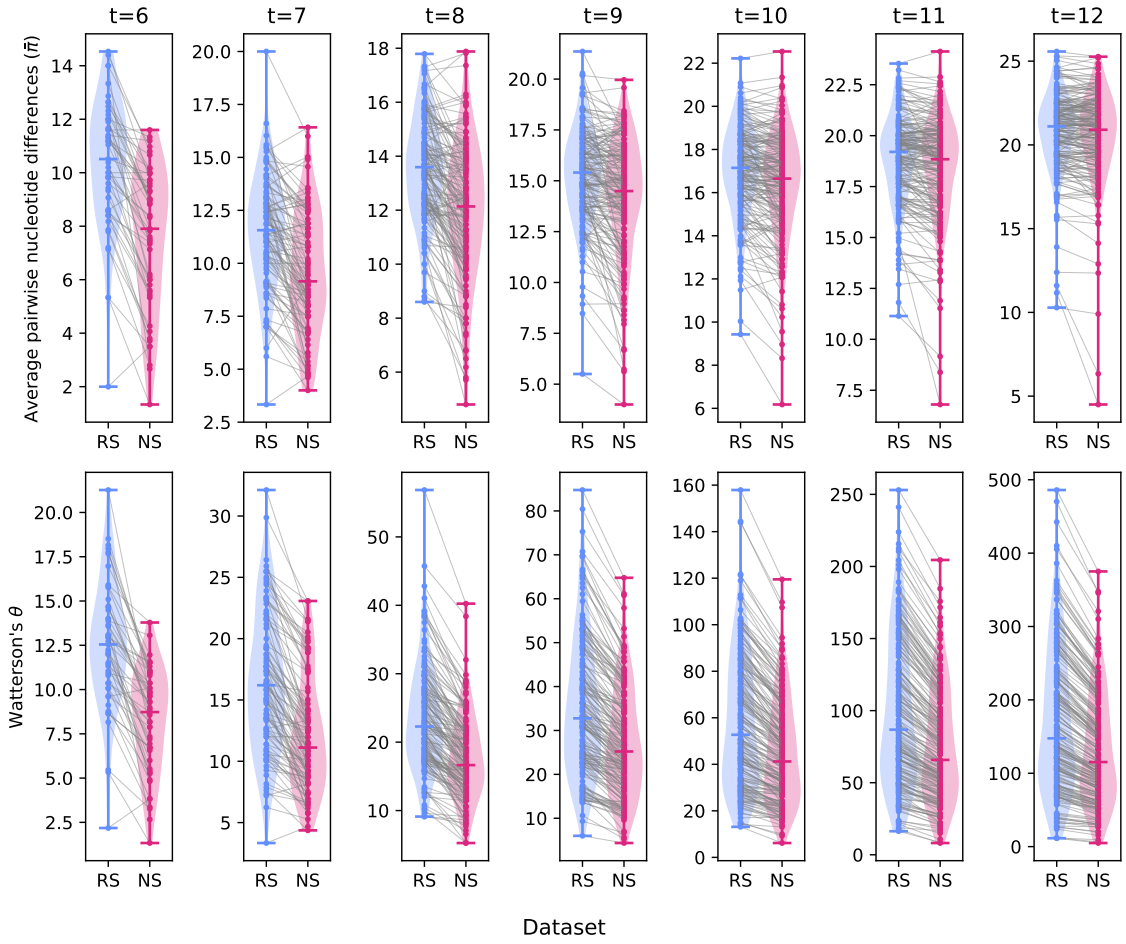

Figure 4.8: **Comparison of the summary statistics from random and non-random datasets over generations under higher p_m.** The average pairwise nucleotide differences (upper panels) and Watterson's θ (lower panels) were calculated from random and nonrandom datasets. Each dots represent the calculated summary statistics from a dataset at each generation. The lines connect the datasets from the same simulation. The distribution of the summary statistics is shown as a violin plot, and the horizontal marker indicates the median of the distribution. Each column represents a generation from t = 6 to t = 12

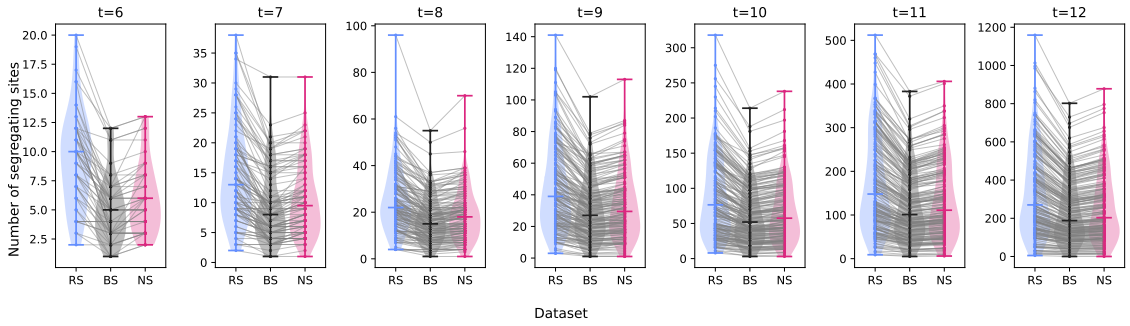

Figure 4.9: **Comparison of the number of segregating sites from the base, random, and nonrandom datasets over generations under higher p_m.** Each dots represent the calculated summary statistics from a dataset at each generation. The lines connect the datasets from the same simulation. The distribution of the summary statistics is shown as a violin plot, and the horizontal marker indicates the median of the distribution. Each column represents a generation from t = 6 to t = 12

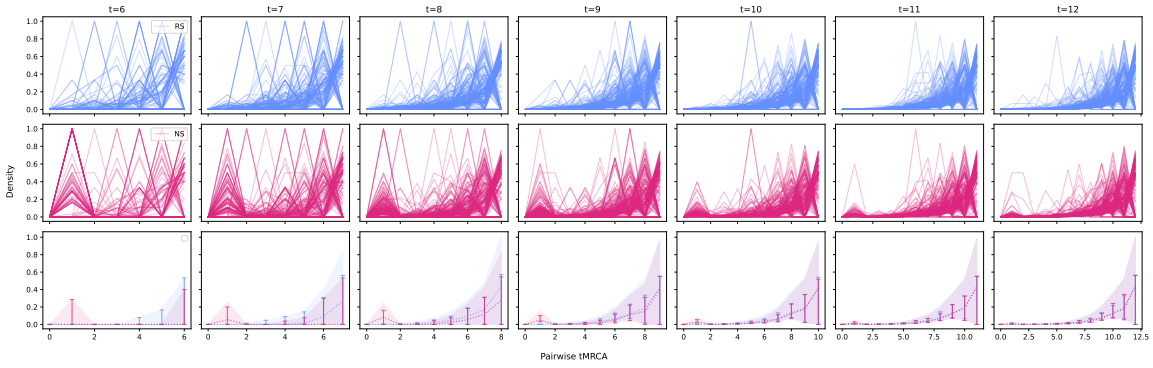

Figure 4.10: **Comparison of the distribution of tMRCA under three sampling schemes over generations under higher p_m.** The distribution of the pairwise tMRCA was obtained from every pair of samples in random (A) and nonrandom (B) datasets across generations. Each line represents the distribution from a dataset. (C) summarizes the individual distributions in (A) and (B). The dotted line represents the median. The error bars and colored area represent the 25% and 75% percentiles. Each column represent a generation from t = 6 to t = 12.

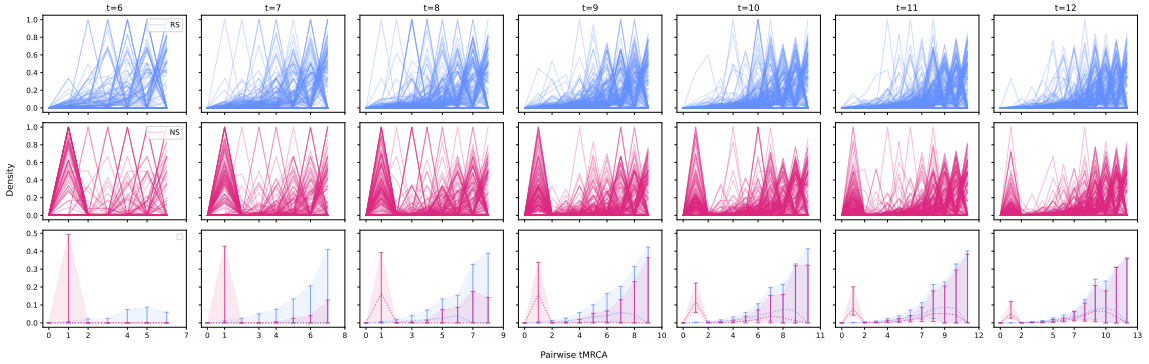

Figure 4.11: **Comparison of the distribution of tMRCA from random and nonrandom datasets over generations transmission heterogeneity with $k = 0.2$.** The distribution of the pairwise tMRCA was obtained from every pair of samples in random (A) and nonrandom (B) datasets across generations. Each line represents the distribution from a dataset. (C) summarizes the individual distributions in (A) and (B). The dotted line represents the median. The error bars and colored area represent the 25% and 75% percentiles. Each column represent a generation from t = 6 to t = 12.

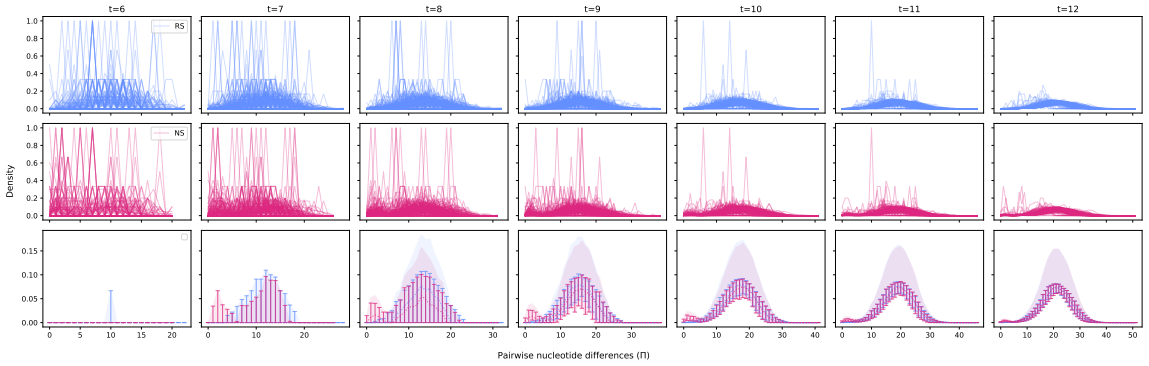

Figure 4.12: **Comparison of the distribution of pairwise nucleotide difference under three sampling schemes over generations under higher p_m.** The distribution of the pairwise tMRCA was obtained from every pair of samples in three datasets across generations: (A) BS, (B) RS, and (C) NS dataset. Each line represents the distribution from a dataset. (D) summarizes the individual distributions in (A), (B), and (C). The dotted line represents the median. The error bars and colored area represent the 25% and 75% quantiles. Each column represent a generation from t = 6 to t = 12.

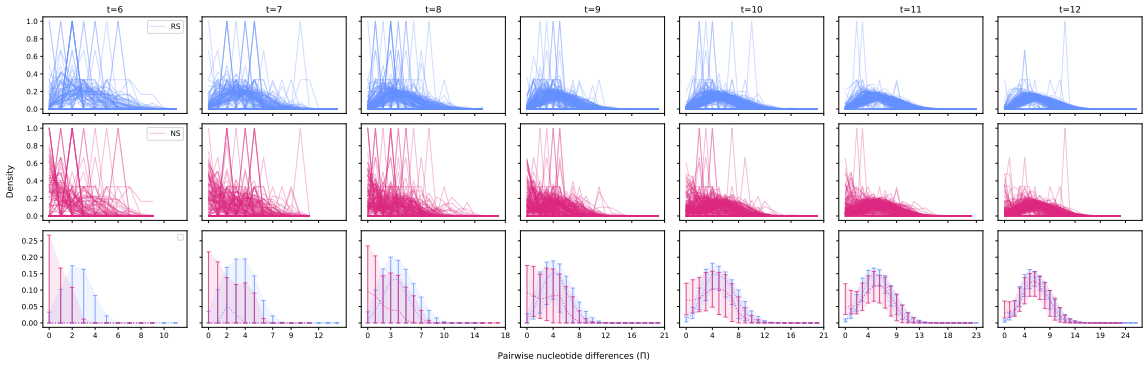

Figure 4.13: **Comparison of the distribution of pairwise nucleotide difference from random and nonrandom dataset over generations transmission heterogeneity with $k = 0.2$.** The distribution of the pairwise tMRCA was obtained from every pair of samples in three datasets across generations: (A) BS, (B) RS, and (C) NS dataset. Each line represents the distribution from a dataset. (D) summarizes the individual distributions in (A), (B), and (C). The dotted line represents the median. The error bars and colored area represent the 25% and 75% quantiles. Each column represent a generation from t = 6 to t = 12.

Baseline scenario		**Average pairwise nucleotide difference**			**Watterson's θ**		
Generation	Replications	Two-tailed t-test	One-tailed t-test	Komologorov-Smirnov test	Two-tailed t-test	One-tailed t-test	Komologorov-Smirnov test
4	**15**	**0.087**	**0.044**	**0.386**	**0.14**	**0.07**	**0.678**
5	**42**	**0.007**	**0.004**	**0.791**	**0.004**	**0.002**	**0.791**
6	**86**	**0**	**0**	**0.019**	**0**	**0**	**0.002**
7	**146**	**0**	**0**	**0.003**	**0**	**0**	**0.006**
8	**168**	**0**	**0**	**0.026**	**0**	**0**	**0**
9	**190**	**0**	**0**	**0.096**	**0**	**0**	**0**
10	**194**	**0**	**0**	**0.061**	**0**	**0**	**0**
11	**197**	**0**	**0**	**0.384**	**0**	**0**	**0.002**
12	**200**	**0.028**	**0.014**	**0.793**	**0**	**0**	**0.001**
High p_s		**Average pairwise nucleotide difference**			**Watterson's θ**		
Generation	Replications	Two-tailed t-test	One-tailed t-test	Komologorov-Smirnov test	Two-tailed t-test	One-tailed t-test	Komologorov-Smirnov test
4	**15**	**0.083**	**0.041**	**0.678**	**0.078**	**0.039**	**0.938**
5	**42**	**0**	**0**	**0.292**	**0**	**0**	**0.186**
6	**86**	**0**	**0**	**0.007**	**0**	**0**	**0**
7	**146**	**0**	**0**	**0.001**	**0**	**0**	**0**
8	**168**	**0**	**0**	**0**	**0**	**0**	**0**
9	**190**	**0**	**0**	**0.032**	**0**	**0**	**0**
10	**194**	**0**	**0**	**0.131**	**0**	**0**	**0.001**
11	**197**	**0**	**0**	**0.319**	**0**	**0**	**0.002**
12	**200**	**0**	**0**	**0.793**	**0**	**0**	**0.004**
Transmission heterogeneity		**Average pairwise nucleotide difference**			**Watterson's θ**		
Generation	Replications	Two-tailed t-test	One-tailed t-test	Komologorov-Smirnov test	Two-tailed t-test	One-tailed t-test	Komologorov-Smirnov test
4	**14**	**0.116**	**0.058**	**0.635**	**0.256**	**0.128**	**1**
5	**41**	**0**	**0**	**0.002**	**0**	**0**	**0.032**
6	**77**	**0**	**0**	**0**	**0**	**0**	**0.011**
7	**113**	**0**	**0**	**0**	**0**	**0**	**0**
8	**145**	**0**	**0**	**0**	**0**	**0**	**0**
9	**161**	**0**	**0**	**0**	**0**	**0**	**0**
10	**178**	**0**	**0**	**0**	**0**	**0**	**0**
11	**187**	**0**	**0**	**0**	**0**	**0**	**0**
12	**187**	**0**	**0**	**0.022**	**0**	**0**	**0**

Table 4.1: Statistical tests for one-dimensional summary statistics

Chapter 4 References

T. Alpert, A. F. Brito, E. Lasek-Nesselquist, J. Rothman, A. L. Valesano, M. J. MacKay, M. E. Petrone, M. I. Breban, A. E. Watkins, C. B. Vogels, C. C. Kalinich, S. Dellicour, A. Russell, J. P. Kelly, M. Shudt, J. Plitnick, E. Schneider, W. J. Fitzsimmons, G. Khullar, J. Metti, J. T. Dudley, M. Nash, N. Beaubier, J. Wang, C. Liu, P. Hui, A. Muyombwe, R. Downing, J. Razeq, S. M. Bart, A. Grills, S. M. Morrison, S. Murphy, C. Neal, E. Laszlo, H. Rennert, M. Cushing, L. Westblade, P. Velu, A. Craney, L. Cong, D. R. Peaper, M. L. Landry, P. W. Cook, J. R. Fauver, C. E. Mason, A. S. Lauring, K. St. George, D. R. MacCannell, and N. D. Grubaugh. Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. *Cell*, 184(10):2595–2604.e13, May 2021. ISSN 00928674. doi: 10.1016/j.cell.2021.03.061. URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867421004347`.

R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, and others. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

A. F. Brito, E. Semenova, G. Dudas, G. W. Hassler, C. C. Kalinich, M. U. G. Kraemer, J. Ho, H. Tegally, G. Githinji, C. N. Agoti, L. E. Matkin, C. Whittaker, Bulgarian SARS-CoV-2 sequencing group, T. Kantardjiev, N. Korsun, S. Stoitsova, R. Dimitrova, I. Trifonova, V. Dobrinov, L. Grigorova, I. Stoykov, I. Grigorova, A. Gancheva,

Communicable Diseases Genomics Network (Australia and New Zealand), A. Jennison, L. Leong, D. Speers, R. Baird, L. Cooley, K. Kennedy, J. De Ligt, W. Rawlinson, S. Van Hal, D. Williamson, COVID-19 Impact Project, R. Singh, S. Nathaniel-Girdharrie, L. Edghill, L. Indar, J. St. John, G. Gonzalez-Escobar, V. Ramkisoon, A. Brown-Jordan, A. Ramjag, N. Mohammed, J. E. Foster, I. Potter, S. Greenaway-Duberry, K. George, S. Belmar-George, J. Lee, J. Bisasor-McKenzie, N. Astwood, R. Sealey-Thomas, H. Laws, N. Singh, A. Oyinloye, P. McMillan, A. Hinds, N. Nandram, R. Parasram, Z. Khan-Mohammed, S. Charles, A. Andrewin, D. Johnson, S. Keizer-Beache, C. Oura, O. G. Pybus, N. R. Faria, Danish Covid-19 Genome Consortium, M. Stegger, M. Albertsen, A. Fomsgaard, M. Rasmussen, Fiocruz COVID-19 Genomic Surveillance Network, R. Khouri, F. Naveca, T. Graf, F. Miyajima, G. Wallau, F. Motta, GISAID core curation team, S. Khare, L. Freitas, C. Schiavina, G. Bach, M. B. Schultz, Y. H. Chew, M. Makheja, P. Born, G. Calegario, S. Romano, J. Finello, A. Diallo, R. T. C. Lee, Y. N. Xu, W. Yeo, S. Tiruvayipati, S. Yadahalli, Network for Genomic Surveillance in South Africa (NGS-SA), E. Wilkinson, A. Iranzadeh, J. Giandhari, D. Doolabh, S. Pillay, U. Ramphal, J. E. San, N. Msomi, K. Mlisana, A. Von Gottberg, S. Walaza, A. Ismail, T. Mohale, S. Engelbrecht, G. Van Zyl, W. Preiser, A. Sigal, D. Hardie, G. Marais, M. Hsiao, S. Korsman, M.-A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O. Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C. K. Wibmer, D. Martin, R. J. Lessells, J. N. Bhiman, C. Williamson, T. De Oliveira, Swiss SARS-CoV-2 Sequencing Consortium, C. Chen, S. Nadeau, L. Du Plessis, C. Beckmann, M. Redondo, O. Kobel, C. Noppen, S. Seidel, N. S. De Souza, N. Beerenwinkel, I. Topolsky, P. Jablonski, L. Fuhrmann, D. Dreifuss, K. Jahn, P. Ferreira, S. Posada-Céspedes, C. Beisel, R. Denes, M. Feldkamp, I. Nissen, N. Santacroce, E. Burcklen, C. Aquino, A. C. De Gouvea, M. D. Moccia, S. Grüter, T. Sykes, L. Opitz, G. White, L. Neff, D. Popovic, A. Patrignani, J. Tracy, R. Schlapbach, E. Dermitzakis, K. Harsh-

man, I. Xenarios, H. Pegeot, L. Cerutti, D. Penet, T. Stadler, B. P. Howden, V. Sintchenko, N. S. Zuckerman, O. Mor, H. M. Blankenship, T. De Oliveira, R. T. P. Lin, M. M. Siqueira, P. C. Resende, A. T. R. Vasconcelos, F. R. Spilki, R. S. Aguiar, I. Alexiev, I. N. Ivanov, I. Philipova, C. V. F. Carrington, N. S. D. Sahadeo, B. Branda, C. Gurry, S. Maurer-Stroh, D. Naidoo, K. J. Von Eije, M. D. Perkins, M. Van Kerkhove, S. C. Hill, E. C. Sabino, O. G. Pybus, C. Dye, S. Bhatt, S. Flaxman, M. A. Suchard, N. D. Grubaugh, G. Baele, and N. R. Faria. Global disparities in SARS-CoV-2 genomic surveillance. *Nature Communications*, 13 (1):7003, Nov. 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-33713-y. URL `https://www.nature.com/articles/s41467-022-33713-y`.

Z. Chen, A. S. Azman, X. Chen, J. Zou, Y. Tian, R. Sun, X. Xu, Y. Wu, W. Lu, S. Ge, Z. Zhao, J. Yang, D. T. Leung, D. B. Domman, and H. Yu. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nature Genetics*, 54(4): 499–507, Mar. 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01033-y. URL `http://dx.doi.org/10.1038/s41588-022-01033-y`.

N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, K. Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit, K. E. Atkins, and W. J. Edmunds. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, 372(6538), Apr. 2021. ISSN 1095-9203. doi: 10.1126/science.abg3055. URL `http://dx.doi.org/10.1126/science.abg3055`.

E. de Silva, N. M. Ferguson, and C. Fraser. Inferring pandemic growth rates from sequence data. *Journal of The Royal Society Interface*, 9(73):1797–1808, Feb. 2012. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2011.0850. URL `https://royalsocietypublishing.org/doi/10.1098/rsif.2011.0850`.

B. L. Dearlove, F. Xiang, and S. D. W. Frost. Biased phylodynamic inferences from analysing clusters of viral sequences. *Virus Evolution*, 3(2), July 2017. ISSN 2057-1577. doi: 10.1093/ve/vex020. URL https://academic.oup.com/ve/article/doi/10.1093/ve/vex020/4061302.

S. Elbe and G. Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46, Jan. 2017. ISSN 2056-6646. doi: 10.1002/gch2.1018. URL http://dx.doi.org/10.1002/gch2.1018.

C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut, O. G. Pybus, H. Lopez-Gatell, C. M. Alpuche-Aranda, I. B. Chapela, E. P. Zavala, D. M. E. Guevara, F. Checchi, E. Garcia, S. Hugonnet, C. Roth, and The WHO Rapid Pandemic Assessment Collaboration. Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings. *Science*, 324(5934):1557–1561, June 2009. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1176062. URL https://www.science.org/doi/10.1126/science.1176062.

S. D. Frost, O. G. Pybus, J. R. Gog, C. Viboud, S. Bonhoeffer, and T. Bedford. Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92, Mar. 2015. ISSN 17554365. doi: 10.1016/j.epidem.2014.09.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1755436514000437.

Y. Furuse. Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. *International Journal of Infectious Diseases*, 103:305–307, Feb. 2021. ISSN 1201-9712. doi: 10.1016/j.ijid.2020.12.034. URL http://dx.doi.org/10.1016/j.ijid.2020.12.034.

N. D. Grubaugh, J. T. Ladner, M. U. G. Kraemer, G. Dudas, A. L. Tan, K. Gan-

gavarapu, M. R. Wiley, S. White, J. Thézé, D. M. Magnani, K. Prieto, D. Reyes, A. M. Bingham, L. M. Paul, R. Robles-Sikisaka, G. Oliveira, D. Pronty, C. M. Barcellona, H. C. Metsky, M. L. Baniecki, K. G. Barnes, B. Chak, C. A. Freije, A. Gladden-Young, A. Gnirke, C. Luo, B. MacInnis, C. B. Matranga, D. J. Park, J. Qu, S. F. Schaffner, C. Tomkins-Tinch, K. L. West, S. M. Winnicki, S. Wohl, N. L. Yozwiak, J. Quick, J. R. Fauver, K. Khan, S. E. Brent, R. C. Reiner, P. N. Lichtenberger, M. J. Ricciardi, V. K. Bailey, D. I. Watkins, M. R. Cone, E. W. Kopp, K. N. Hogan, A. C. Cannons, R. Jean, A. J. Monaghan, R. F. Garry, N. J. Loman, N. R. Faria, M. C. Porcelli, C. Vasquez, E. R. Nagle, D. A. T. Cummings, D. Stanek, A. Rambaut, M. Sanchez-Lockhart, P. C. Sabeti, L. D. Gillis, S. F. Michael, T. Bedford, O. G. Pybus, S. Isern, G. Palacios, and K. G. Andersen. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature*, 546 (7658):401–405, May 2017. ISSN 1476-4687. doi: 10.1038/nature22400. URL `http://dx.doi.org/10.1038/nature22400`.

M. D. Hall, M. E. J. Woolhouse, and A. Rambaut. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: A simulation study. *Virus Evolution*, 2(1), Jan. 2016. ISSN 2057-1577. doi: 10.1093/ve/vew003. URL `http://dx.doi.org/10.1093/ve/vew003`.

W. W. Hannon, P. Roychoudhury, H. Xie, L. Shrestha, A. Addetia, K. R. Jerome, A. L. Greninger, and J. D. Bloom. Narrow transmission bottlenecks and limited within-host viral diversity during a SARS-CoV-2 outbreak on a fishing boat. *Virus Evol.*, 8(2):veac052, June 2022.

D. Hare, G. Gonzalez, J. Dean, K. McDonnell, M. J. Carr, and C. F. De Gascun. Genomic epidemiological analysis of SARS-CoV-2 household transmission. *Access*

Microbiology, 3(7), July 2021. ISSN 2516-8290. doi: 10.1099/acmi.0.000252. URL http://dx.doi.org/10.1099/acmi.0.000252.

J. Hedge, S. J. Lycett, and A. Rambaut. Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biology Letters*, 9(5):20130331, Oct. 2013. ISSN 1744-957x. doi: 10.1098/rsbl.2013.0331. URL http://dx.doi.org/10.1098/rsbl.2013.0331.

M. D. Karcher, J. A. Palacios, T. Bedford, M. A. Suchard, and V. N. Minin. Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference. *PLOS Computational Biology*, 12(3):e1004789, Mar. 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004789. URL http://dx.doi.org/10.1371/journal.pcbi.1004789.

B. Li, A. Deng, K. Li, Y. Hu, Z. Li, Y. Shi, Q. Xiong, Z. Liu, Q. Guo, L. Zou, H. Zhang, M. Zhang, F. Ouyang, J. Su, W. Su, J. Xu, H. Lin, J. Sun, J. Peng, H. Jiang, P. Zhou, T. Hu, M. Luo, Y. Zhang, H. Zheng, J. Xiao, T. Liu, M. Tan, R. Che, H. Zeng, Z. Zheng, Y. Huang, J. Yu, L. Yi, J. Wu, J. Chen, H. Zhong, X. Deng, M. Kang, O. G. Pybus, M. Hall, K. A. Lythgoe, Y. Li, J. Yuan, J. He, and J. Lu. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *Nat. Commun.*, 13(1):460, Jan. 2022.

M. A. Martin and K. Koelle. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". *Sci. Transl. Med.*, 13(617):eabh1803, Oct. 2021.

M. A. Martin, D. VanInsberghe, and K. Koelle. Insights from SARS-CoV-2 sequences. *Science*, 371(6528):466–467, Jan. 2021. ISSN 1095-9203. doi: 10.1126/science.abf3995. URL http://dx.doi.org/10.1126/science.abf3995.

J. T. McCrone, R. J. Woods, E. T. Martin, R. E. Malosh, A. S. Monto, and A. S.

Lauring. Stochastic processes constrain the within and between host evolution of influenza virus. *Elife*, 7, May 2018.

B. North, D. Curtis, and P. Sham. A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. *The American Journal of Human Genetics*, 71(2):439–441, Aug. 2002. ISSN 0002-9297. doi: 10.1086/341527. URL http://dx.doi.org/10.1086/341527.

K. V. Parag, L. du Plessis, and O. G. Pybus. Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences. *Molecular Biology and Evolution*, 37(8):2414–2429, Jan. 2020. ISSN 1537-1719. doi: 10.1093/molbev/msaa016. URL http://dx.doi.org/10.1093/molbev/msaa016.

Y. Park, M. A. Martin, and K. Koelle. Epidemiological inference for emerging viruses using segregating sites. *Nature Communications*, 14(1):3105, 2023.

A. F. Y. Poon. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evolution*, 2(2):vew031, July 2016. ISSN 2057-1577. doi: 10.1093/ve/vew031. URL https://academic.oup.com/ve/article-lookup/doi/10.1093/ve/vew031.

A. Rambaut and E. Holmes. The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Currents*, 1:Rrn1003, Aug. 2009. ISSN 2157-3999. doi: 10.1371/currents.rrn1003. URL http://dx.doi.org/10.1371/currents.RRN1003.

L. Song, H. Liu, F. S. L. Brinkman, E. Gill, E. J. Griffiths, W. W. L. Hsiao, S. Savić-Kallesøe, S. Moreira, G. Van Domselaar, M. H. Zawati, and Y. Joly. Addressing Privacy Concerns in Sharing Viral Sequences and Minimum Contextual Data in a Public Repository During the COVID-19 Pandemic. *Frontiers in Genetics*,

12, Mar. 2022. ISSN 1664-8021. doi: 10.3389/fgene.2021.716541. URL http://dx.doi.org/10.3389/fgene.2021.716541.

R. Spott, M. W. Pletz, C. Fleischmann-Struzek, A. Kimmig, C. Hadlich, M. Hauert, M. Lohde, M. Jundzill, M. Marquet, P. Dickmann, R. Schüchner, M. Hölzer, D. Kühnert, and C. Brandt. Exploring the Spatial Distribution of Persistent SARS-CoV-2 Mutations - Leveraging mobility data for targeted sampling. Nov. 2024. doi: 10.7554/elife.94045.2. URL http://dx.doi.org/10.7554/eLife.94045.2.

F. Tajima. Effect of non-random sampling on the estimation of parameters in population genetics. *Genetical Research*, 66(3):267–276, Dec. 1995. ISSN 0016-6723, 1469-5073. doi: 10.1017/s0016672300034704. URL https://www.cambridge.org/core/product/identifier/S0016672300034704/type/journal%5Farticle.

R. Viana, S. Moyo, D. G. Amoako, H. Tegally, C. Scheepers, C. L. Althaus, U. J. Anyaneji, P. A. Bester, M. F. Boni, M. Chand, W. T. Choga, R. Colquhoun, M. Davids, K. Deforche, D. Doolabh, L. du Plessis, S. Engelbrecht, J. Everatt, J. Giandhari, M. Giovanetti, D. Hardie, V. Hill, N.-Y. Hsiao, A. Iranzadeh, A. Ismail, C. Joseph, R. Joseph, L. Koopile, S. L. Kosakovsky Pond, M. U. G. Kraemer, L. Kuate-Lere, O. Laguda-Akingba, O. Lesetedi-Mafoko, R. J. Lessells, S. Lockman, A. G. Lucaci, A. Maharaj, B. Mahlangu, T. Maponga, K. Mahlakwane, Z. Makatini, G. Marais, D. Maruapula, K. Masupu, M. Matshaba, S. Mayaphi, N. Mbhele, M. B. Mbulawa, A. Mendes, K. Mlisana, A. Mnguni, T. Mohale, M. Moir, K. Moruisi, M. Mosepele, G. Motsatsi, M. S. Motswaledi, T. Mphoyakgosi, N. Msomi, P. N. Mwangi, Y. Naidoo, N. Ntuli, M. Nyaga, L. Olubayo, S. Pillay, B. Radibe, Y. Ramphal, U. Ramphal, J. E. San, L. Scott, R. Shapiro, L. Singh, P. Smith-Lawrence, W. Stevens, A. Strydom, K. Subramoney, N. Tebeila, D. Tshiabuila, J. Tsui, S. van Wyk, S. Weaver, C. K. Wibmer, E. Wilkinson, N. Wolter, A. E. Zarebski, B. Zuze, D. Goedhals, W. Preiser, F. Treurnicht,

M. Venter, C. Williamson, O. G. Pybus, J. Bhiman, A. Glass, D. P. Martin, A. Rambaut, S. Gaseitsiwe, A. von Gottberg, and T. de Oliveira. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*, 603 (7902):679–686, Jan. 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04411-y. URL http://dx.doi.org/10.1038/s41586-022-04411-y.

R. S. Waples and E. C. Anderson. Purging putative siblings from population genetic data sets: a cautionary view. *Molecular Ecology*, 26(5):1211–1224, Feb. 2017. ISSN 1365-294x. doi: 10.1111/mec.14022. URL http://dx.doi.org/10.1111/mec.14022.

J. O. Wertheim, A. J. Leigh Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D. M. Smith, and S. L. Kosakovsky Pond. The Global Transmission Network of HIV-1. *The Journal of Infectious Diseases*, 209(2):304–313, Oct. 2013. ISSN 0022-1899. doi: 10.1093/infdis/jit524. URL http://dx.doi.org/10.1093/infdis/jit524.

S. Wohl, J. R. Giles, and J. Lessler. Sample size calculation for phylogenetic case linkage. *PLOS Computational Biology*, 17(7):e1009182, July 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009182. URL http://dx.doi.org/10.1371/journal.pcbi.1009182.

Chapter 5

Transmission history reconstruction using phylogenies

The following *commentary* was written in March 2022 as part of the PBEE qualifying exam. I begin by examining early approaches that used phylogenies to infer transmission direction. After recognizing that transmission trees and phylogenies are distinct entities, I review methods for reconstructing 'who-infected-whom' transmission histories. As existing approaches are based on diverse assumptions, data types, and methodologies, I conclude that the choice of approach should be guided by the characteristics of the available data and that systematic comparisons are needed to better guide users.

5.1 Introduction

Understanding transmission dynamics is key to the effective control and management of infectious diseases. These dynamics can be studied at different scales, from transmission chains between individuals in an outbreak to global patterns of disease spread during a pandemic or endemic circulation of a pathogen. Characterizing transmission chains between individual hosts (or, more generally, infectious units) is fundamental for

understanding why and where outbreaks occur. Reconstructed transmission histories of 'who-infected-whom' can be used to identify the source of an outbreak and the key attributes of the infectious units that fuel the local spread of a pathogen (Ypma et al., 2012). Transmission histories are often depicted using a tree representation (see Section 5.7.1).

The process of transmission tree reconstruction aims to identify all relevant infectious units and the transmission process between them, including transmission timing. As such, accurate tree reconstruction often requires extensive epidemiological data, including contact histories of individuals and the timing of symptom development of infected cases (Cauchemez and Ferguson, 2011). While case investigations and contact tracing can, at times, provide these data, they are often incomplete and involve uncertainty in observation. Pathogen genetic data have the potential to complement these more traditional epidemiological data sources, providing the possibility of further resolving who-infected-whom during the transmission process. Genetic data may prove useful because pathogens, with their short generation times of minutes to days, can accumulate mutations over the course of an outbreak. Patterns of shared (and unshared) genetic variation can thus be informative of transmission patterns and be used to reconstruct the history of pathogen spread (Holmes et al., 1995).

Over the last 15 years, quantitative approaches for inferring transmission histories have undergone significant development. Approaches that integrate pathogen genetic data to infer these histories are broadly reviewed in HALL et al. (2016). In this review, existing approaches up to the time of publication were categorized according to whether and how they integrated within-host genetic variation. An alternative way to classify transmission history inference approaches is by whether an approach adopts a 'pairwise approach' or a 'phylogenetic approach.' Pairwise approaches rely on calculations of pairwise genetic distances between samples isolated from each infectious unit to infer transmission histories. In contrast, phylogenetic approaches rely on the

use of pathogen phylogenies, which reconstruct the ancestral relationships between pathogens. In this commentary, I focus specifically on reviewing and synthesizing quantitative approaches that have relied on phylogenies reconstructed from pathogen genetic data to infer transmission histories. For a brief overview of phylogenetic trees and phylogenetic inference methods, see 5.7.2.

In the following, I first discuss early work showing how phylogenetic trees can provide insight into the structure of transmission chains even in the absence of rigorous quantitative integration with transmission tree inference. Based on points made in the literature, I then argue why phylogenetic trees should be more formally and systematically incorporated into transmission tree inference rather than simply being used as a proxy for transmission trees. Finally, I discuss statistical phylogenetic-based transmission tree inference approaches that have been developed as phylogenetic inference methods have become more heavily used by the community of infectious disease modelers. I end by discussing the applicability of phylogenetic-based transmission tree inferences to real-world transmission chains and by providing my perspective on future work that is needed to systematically compare across different inference approaches.

5.2 Early analyses using pathogen phylogenies to infer transmission history

Early approaches that relied on pathogen phylogenies to infer transmission histories focused on identifying the sources of infection in epidemiological investigations. In these early studies (Esteban et al., 1996; Heinsen et al., 2000; Holmes et al., 1993; Metzker et al., 2002), tree topologies describing the evolutionary relationships between pathogen samples were used to ascertain the strength of evidence for a direct epidemiological link between a putative donor and a recipient of infection. Evidence for a direct transmission link between a donor and a recipient was provided by sample isolates

(tips) from a putative donor and a recipient being evolutionarily more closely related to one another than between the recipient and sampled isolates from other individuals who were thought to be epidemiologically unlinked to the recipient. The reasoning behind this expectation is that, since pathogens within a recipient host originated from those in a donor host, fewer mutations (and a greater degree of evolutionary relatedness) are expected between donor and recipient sample isolates compared to ones between a recipient and an unlinked host. This will lead to the clustering of pathogen isolates from donor and recipient, forming a monophyletic clade (Figure 1). When isolates from two individuals form a monophyletic clade, phylogenetic relationships within the observed monophyletic clade can be further investigated to ascertain the strength of evidence for a donor-to-recipient transmission chain. Within the monophyletic clade, expected phylogenetic relationships between two individuals linked by direct transmission can be categorized into three classes (Leitner, 2019; Romero-Severson et al., 2016): monophyletic recipient isolates that are nested within paraphyletic recipient isolates (PM relationship, Figure 5.1A), polyphyletic recipient isolates that are nested within paraphyletic recipient isolates (PP relationship, Figure 5.1B), and monophyletic recipient isolates alongside monophyletic recipient isolates that together form a sister clade relationship (MM relationship, Figure 5.1C).

PM relationships are expected to be observed when a small number of pathogen particles transmit from a donor to a recipient through a tight transmission bottleneck. (Transmission bottleneck size is defined as the number of pathogen particles founding an infection, and many analyses on various pathogens have found transmission bottleneck sizes to be small, particularly for respiratory viruses (Lythgoe et al., 2021; Martin and Koelle, 2021; McCrone and Lauring, 2017). Due to the presence of these tight transmission bottlenecks, this nested PM structure is expected for many pathogens. The PM relationship can also arise from indirect transmission from a donor to a recipient through an intermediate, unsampled host. The PP relationship indicates

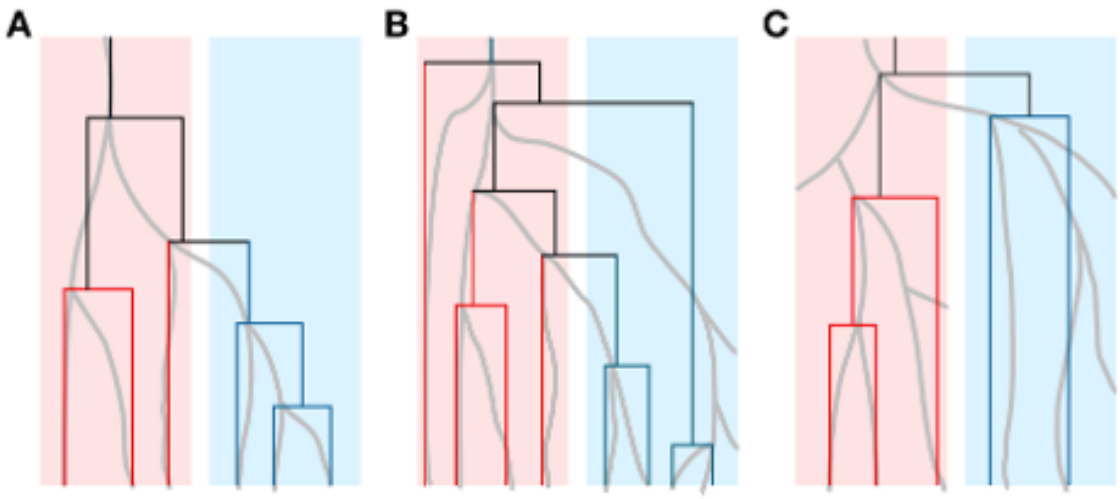

Figure 5.1: **Possible tree topologies under a scenario of direct transmission.** Pathogen population histories within hosts are indicated with grey curves. The host in whom each pathogen lineage resides is indicated with colored boxes, with red denoting the donor and blue denoting the recipient. Phylogenies of extant pathogen lineages are shown as a tree with straight lines. Branches of phylogenetic trees are colored according to the sampled host of each tip. (A) A hypothetical PM phylogeny resulting from the direct transmission of a pathogen from a donor to a recipient in the case of a stringent transmission bottleneck and little pathogen turnover. Pathogens sampled from the recipient form a monophyletic clade. Pathogens sampled from the donor form a paraphyletic clade. (B) A hypothetical PP phylogeny resulting from the direct transmission of a pathogen from a donor to a recipient in the case of a loose transmission bottleneck. Pathogens sampled from the recipient form a paraphyletic clade. Pathogens sampled from the donor also form a paraphyletic clade. (C) A hypothetical MM phylogeny resulting from the direct transmission of a pathogen from a donor to a recipient in the case of a strict bottleneck and lineage turnover in both hosts. Pathogens sampled from the recipient and the donor each form a monophyletic clade, such that the donor-derived clade and the recipient-derived clade are sister clades.

that multiple pathogen lineages are transmitted from a donor to a recipient. The transmission of multiple pathogen lineages requires transmission bottlenecks to be loose. With the transmission of multiple pathogen lineages, the PP relationship provides the strongest evidence for direct transmission from a donor to a recipient. The MM relationship can occur when direct transmission from a donor to a recipient is followed by lineage turnover within both individuals. However, MM relationships can also arise in cases of indirect transmission from a donor to a recipient via an unsampled host or when the two sampled individuals instead share a common source of infection rather than being in a donor-recipient relationship. Thus, caution is required when interpreting this relationship as evidence in support of direct transmission. These relationships between underlying transmission scenarios and topological phylogenetic patterns are systemically discussed and investigated in Romero-Severson et al. (2016).

Based on these theoretically expected patterns between tree topology and transmission between a donor and a recipient, phylogenetic analyses of pathogen genes have been used to contribute to epidemiological investigations. One notable real-world instance of a phylogenetic analysis that was used as criminal evidence focused on a gastroenterologist who was accused of attempted murder of his ex-girlfriend (hereafter, "the victim") through deliberately infecting her with blood or blood products obtained from one of his HIV+ patients (Metzker et al., 2002). To investigate this accusation, viral samples were sequenced from the patient, the victim, and HIV-positive individuals from the local community. Phylogenetic trees were reconstructed from a portion of the envelope gene (*gp120*) and from the reverse transcriptase (RT) gene region of the HIV genome. In the *gp120* phylogeny, the patient's sequences and the victim's sequences formed a monophyletic clade. However, the relationship was an MM relationship, with the victim's sequences and the patient's sequences each being monophyletic (Figure 5.2A). This sister clade relationship likely arose as a result of rapid lineage turnover in the victim's and the patient's *gp120* gene, as a result of positive selection acting on the

envelope glycoprotein. While the *gp120* gene region phylogeny thus points towards a possible donor-recipient relationship between the patient and the victim, it alone does not provide sufficient evidence for the suspected directionality of virus transmission. Phylogenetic reconstruction from the RT gene, however, considerably strengthened the criminal evidence. Here, the patient's sequences formed a paraphyletic clade, with the victim's sequences nested within the patient's sequences, corresponding to a PM relationship (Figure 5.2B). This relationship appeared in 100% of bootstrap replicates, indicating strong phylogenetic support for this clustering of patient and victim sequences. Taken together, these phylogenetic analyses provided strong evidence for the prosecution's argument that the victim's source of infection was the gastroenterologist's patient. Similar approaches were used to determine the sources of hepatitis C virus infections in nosocomial transmission settings (Esteban et al., 1996; Heinsen et al., 2000).

Phylogenetic relationships between sampled pathogen sequences have also been used to determine between several possible sources of infection. For example, after a surgeon in Baltimore tested positive for HIV, the surgeon's patients were recalled and tested for HIV infection. One of the identified patients had two known risk factors for HIV infection, namely, having undergone a surgical procedure by an HIV-positive surgeon and having received a blood transfusion. To investigate the source of infection for this patient, a phylogenetic tree was reconstructed from viral sequences of the patient, the surgeon, the blood donor who was found to be HIV-positive, and other published HIV sequences. Phylogenetic reconstruction indicated that the blood donor's sequences and the patient's sequences formed a monophyletic clade, with the surgeon's sequences forming a distinct, genetically more distant clade (Figure 5.2C). This indicates that the patient's and the blood donor's viral populations were more closely related to one another than the patient's and the surgeon's, such that the source of infection was likely the blood donor rather than the surgeon (Holmes et al.,

1993).

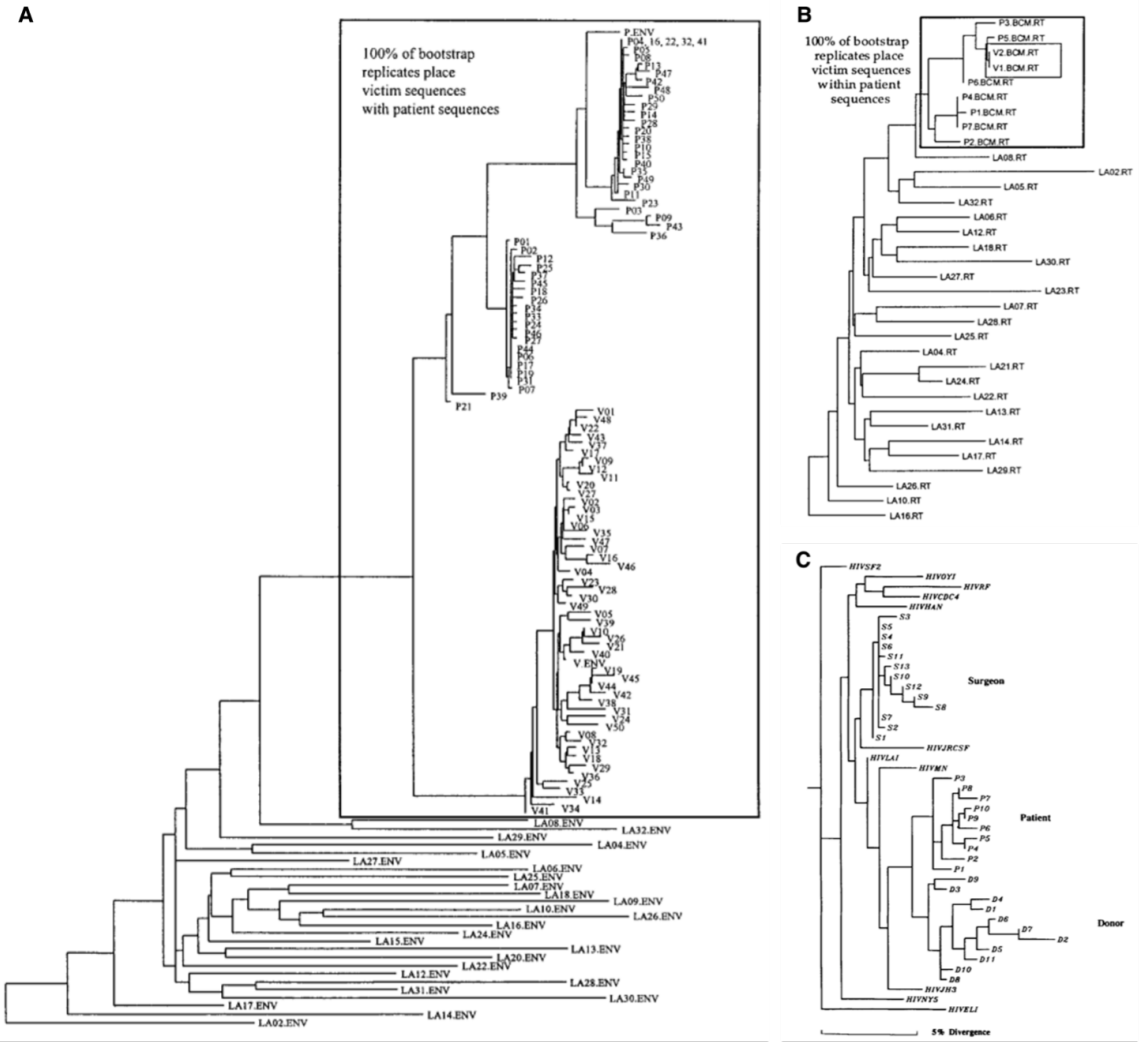

Figure 5.2: **Reconstructed phylogenies including samples from potentially epidemiologically-linked individuals.** (A) A phylogenetic tree reconstructed from the *gp120* region of the HIV genome. Samples from the gastroenterologist's HIV-positive patient are labeled with a P, those from the victim are labeled with a V, and those from HIV-positive individuals from the local community are labeled with an LA. (B) A phylogenetic tree reconstructed from the RT region of the HIV genome. Labeling is as in panel (A). Phylogenies in (A) and (B) are adapted from Metzker et al. (2002) with permission (Copyright (2002) National Academy of Sciences, U.S.A.). (C) A phylogenetic tree reconstructed from the gag gene region of the HIV genome. Samples from the surgeon are labeled S1-S13, those from his patients are labeled P1-P10, and those from the blood donor of the transfusion received by the patient are labeled D1-D11. Reproduced from Holmes et al. (1993) with permission (Copyright (1993) The Journal of infectious diseases); permission conveyed through Copyright Clearance Center, Inc.

5.3 Recognizing differences between transmission trees and phylogenetic trees

Inspired by early successes in using pathogen phylogenies to identify sources of infection, Leitner and colleagues argued that phylogenetic trees had the potential to be used more quantitatively to infer transmission histories of viral pathogens, including HIV (Leitner et al., 1996). More specifically, in this study, the authors focused on an HIV transmission cluster that had full information on who-infected-whom. Inferred phylogenies using different regions of the HIV genome were compared to the topology of this known transmission tree. Phylogenetic trees were reconstructed using five different inference approaches: neighbor-joining, minimum evolution, maximum likelihood, maximum parsimony, and the unweighted pair group method using arithmetic averages (UPGMA). A range of sequence evolution models were also considered. Based on topological comparisons between the transmission tree and each of these inferred phylogenetic trees, Leitner and colleagues concluded that the majority of the viral phylogenies accurately recovered the true transmission tree.

While this early study did not make a distinction between transmission trees and phylogenetic trees, later studies pointed out that these two types of trees are conceptually different entities and might also differ from one another topologically for one or more reasons (Jombart et al., 2010; Pybus and Rambaut, 2009; Romero-Severson et al., 2014; Ypma et al., 2013). Jombart and colleagues more specifically argued that using phylogenetic trees to identify who-infected-whom may be problematic as the internal nodes of a phylogeny represent unobserved common ancestors and all sampled isolates occur at the tips of the tree. In contrast, in a transmission tree, internal nodes are commonly donor individuals, and tips are infected individuals from whom secondary transmission did not occur (Jombart et al., 2010).

Another study pointed out that transmission trees and phylogenetic trees could

differ from one another in terms of tree topology because of incomplete lineage sorting (Pybus and Rambaut, 2009; Ypma et al., 2013). Incomplete lineage sorting is commonly observed across broad taxonomic groups and has given rise to the distinction between gene trees and species trees (Maddison and Knowles, 2006; Rosenberg and Nordborg, 2002). In the case of pathogen phylogenies, incomplete lineage sorting occurs when the coalescence of isolates from two recipients precedes the coalescence of isolates from the donor and either of the recipients (Figure 5.3A). This could result in a disagreement in the topology of the transmission tree and the pathogen phylogeny: in the phylogeny, the two recipients would be more closely related to one another than either one is to the donor when the true transmission tree would have the donor as the source of both recipient infections.

Finally, another study pointed out that even when a transmission tree and a phylogeny share the same topology, their branch lengths could differ (Figure 5.3B) (Romero-Severson et al., 2014; Ypma et al., 2013). This is especially the case with dense sampling, as the coalescent times in time-resolved phylogenetic trees precede the transmission times that are depicted in transmission trees. This is because, with genetic diversity present in an infection, the time of the most recent common ancestor of a pathogen sampled in a recipient and one sampled in a donor should be prior to transmission occurrence. This discrepancy between the time of lineage coalescence and the time of transmission has been termed the "pre-transmission interval" (Leitner and Albert, 1999), and it is this interval that results in differences in the branch lengths of the transmission tree and the phylogenetic tree.

.

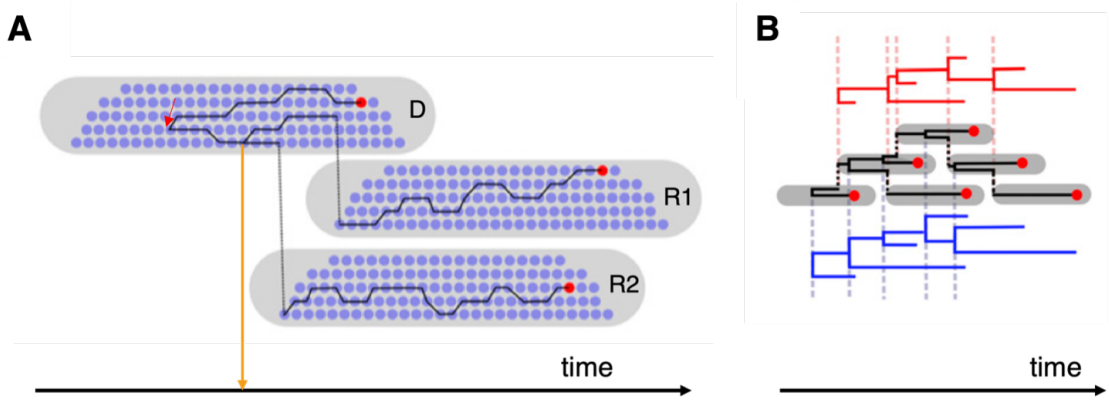

Figure 5.3: **Diagrams depicting scenarios by which phylogenetic trees and transmission trees become inconsistent with, or different from, one another.** Time progresses from left to right in both panels. Grey ovals denote hosts. (A) A depiction of incomplete lineage sorting. The genealogical relationship between pathogen particles sampled from an infection donor (top oval) and two recipients (middle and bottom oval) is shown. Blue and red dots in the grey ovals denote unsampled and sampled pathogen particles, respectively. The lineage of the sampled pathogen particles in recipients (middle and bottom ovals marked as R1 and R2) coalesces before it coalesces with the donor (top oval marked as D), the coalescence of two lineages (indicated with an orange arrow) precedes the coalescence of the donor and recipient lineages (indicated with a red arrow). Reproduced with modification from Ypma et al. (2013) with permission (Copyright (2013) Genetics). (B) A depiction of the pre-transmission interval. In the tree with grey ovals and red dots (tree in the middle) connected by black lines, grey ovals, and red dots denote hosts and sampled pathogen particles. Solid black lines indicate the pathogen lineages and dotted lines indicate the transmission of a pathogen from one host to another. Transmission tree representation (in red) and phylogenetic tree representation of the tree in the middle are shown above and below. The transmission tree branches at the transmission event, and the phylogenetic tree branches at the coalescent event. Reproduced with modification from Ypma et al. (2013) with permission (Copyright (2013) Genetics); permission conveyed through Copyright Clearance Center, Inc.

5.4 Reconstructing transmission trees using phylogenies

Given the above studies that have highlighted the conceptual, topological, and quantitative differences between transmission trees and phylogenetic trees, how could phylogenetic trees be used to infer transmission histories within an outbreak? An early study by Cottam and colleagues acknowledged the difference between these two tree types and then reconciled their interpretation by assigning an infectious unit to internal nodes of the pathogen phylogeny (Cottam et al., 2008). The overarching aim of this study was to reconstruct the transmission history of the United Kingdom foot-and-mouth disease outbreak that occurred in 2001. During this outbreak, a total of 20 farms were infected with the foot-and-mouth-disease virus (FMDV). While the sources of infection for 5 of these farms had been established, the sources of infection for the remaining 15 farms remained unknown. As the number of transmission trees compatible with epidemiological data is large, Cottam and colleagues first used the reconstructed phylogenetic tree to constrain the number of candidates for the reconstructed transmission tree to a plausible set. Instead of equating the reconstructed phylogenetic tree with the transmission tree, they then inferred the transmission tree by assigning infected farms to the internal nodes of the reconstructed phylogenetic tree (Figure 5.4A). The transmission tree candidates compatible with the farm-assigned phylogenetic tree were selected, and the likelihood of each of these transmission trees was calculated to identify the maximum-likelihood transmission tree.

Although this study showed how a phylogenetic tree can be used to infer the transmission tree through the host assignment, it assumed the absence of within-farm viral diversity. However, within-farm - or more generally, within-infectious unit genetic diversity - can arise through *de novo* mutation that occurs within the unit and/or by the unit receiving multiple lineages from one or more infection sources. As such,

this assumption may or may not be an appropriate one to adopt, depending on which pathogen is considered, the extent and duration of pathogen circulation within the unit, the scale of the unit considered, and the contact patterns between a unit and other units. For example, at the scale of individual hosts being the infectious unit, adopting the assumption of no within-host genetic diversity may be reasonable for acutely-infecting respiratory viruses. This is because within-host viral diversity in these infections appears to be low (McCrone et al., 2018; Valesano et al., 2021). In contrast, pathogens with longer infection durations or ones causing chronic infection within a host may accrue genetic diversity, violating this assumption. Within-host genetic diversity can also come about when multiple pathogen lineages initiate infection in a host, as might be the case when the transmission bottleneck size is large (resulting in a genetic bottleneck size that exceeds one) or when superinfection occurs (that is, when an individual experiences infection from one donor on top of an earlier infection from another donor) (Wymant et al., 2017).

Ignoring pathogen genetic diversity within an infectious unit, when it occurs, could result in systematic errors in the reconstruction of transmission histories (Romero-Severson et al., 2014; Worby et al., 2014; Worby and Read, 2015). Thus, recent methodological developments in transmission tree inference have focused on statistically accommodating within-unit pathogen genetic diversity. Below, I review these developments in the context of the assumptions these approaches make on the source or sources of within-unit genetic diversity.

5.4.1 Within-unit genetic diversity stemming from *de novo* mutation

A subset of developed phylogeny-based approaches for reconstructing transmission trees in the presence of within-unit genetic diversity assumes that *de novo* mutations occurring within-unit are the source of diversity. Specifically, they assume that the

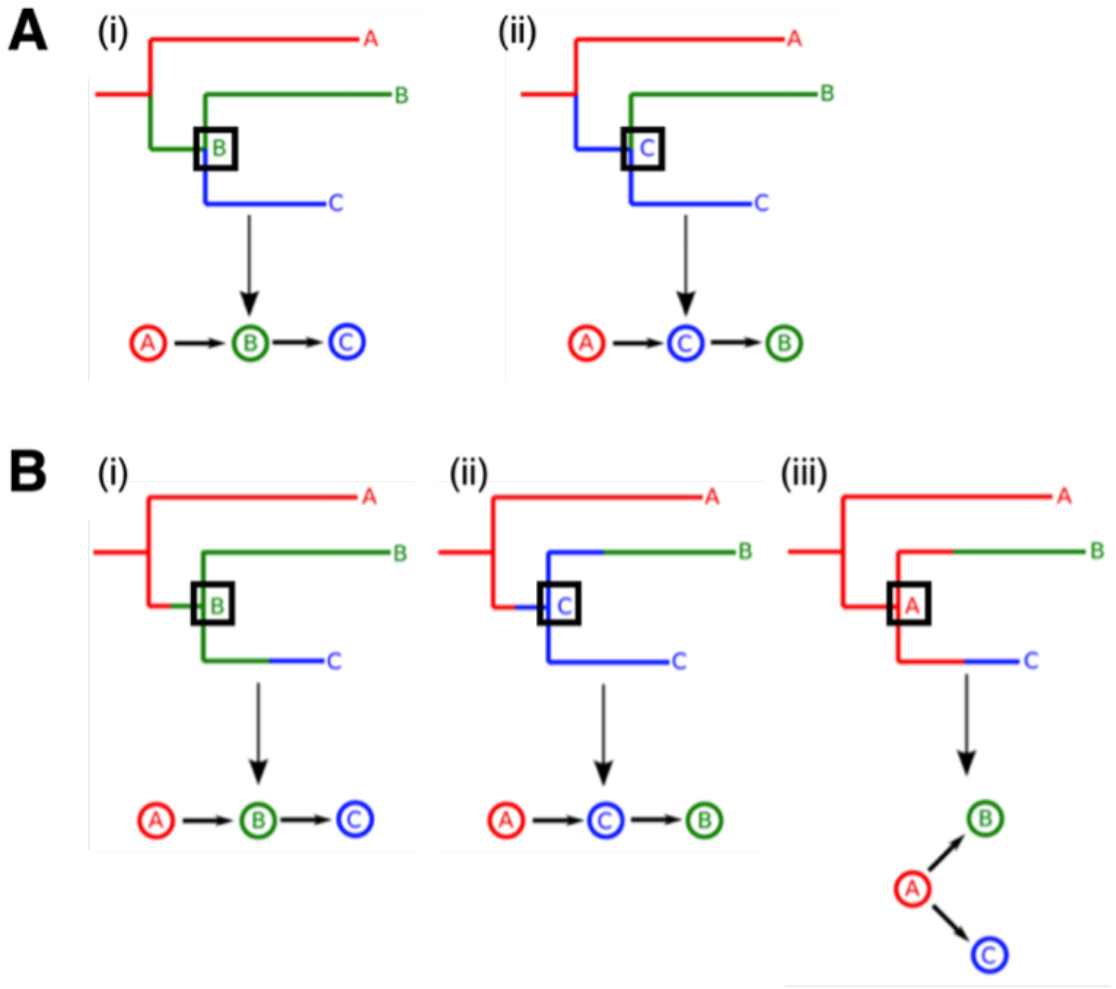

Figure 5.4: **Assignment of hosts to internal nodes of a given phylogenetic tree using likelihood-based approaches.** In the phylogenetic trees shown, tips are labeled according to the host the pathogen was sampled from. Internal nodes are assigned to a possible host in which the associated ancestral pathogen resides. Branches are colored based on the host in which the pathogen lineage is circulating. Once the host assignment is complete, each phylogenetic tree is translated into a transmission tree, shown in the 'bean-bag' representation below each phylogeny. (A) The two possible host assignments (and resulting transmission histories) in the case of no within-host pathogen diversity. Without within-host diversity, there is only one (non-evolving) lineage present in a given host. Thus, the transmission scenario compatible with the given phylogeny is where the ancestral pathogen is located in one of the two hosts that are present at the tips associated with the internal node. This limits the number of possible hosts for an internal node to two. (B) The three possible host assignments (and resulting transmission histories) in the case of within-host diversity. When within-host diversity (stemming from *de novo* mutation) is considered, the given phylogenetic tree is consistent with all three of these different transmission history scenarios. Reproduced with modification from Hall et al. (2015), published in PLoS Computational Biology under a Creative Commons Attribution 4.0 International (CC BY) license.

genetic bottleneck during the transmission event is complete (that is, that the genetic bottleneck size is one) and that each unit is infected only once. When within-host diversity arising from *de novo* mutation is incorporated, transmission events no longer coincide with the coalescence of pathogen lineages in the pathogen phylogeny, as the mutation occurs during replication within a host. Thus, the coalescent event associated with an internal node of a phylogenetic tree could have happened prior to the transmission of a pathogen by a donor or following the transmission of a pathogen in a recipient.

To allow for this phenomenon, Didelot and colleagues extended the conceptual approach developed by Cottam and colleagues (Cottam et al., 2008) to explicitly incorporate within-unit evolution arising from *de novo* mutation (Didelot et al., 2014). In the development of their approach, they considered the infectious unit to be individual hosts with relatively long-term infections. Given a time-resolved phylogenetic tree and infection recovery times, their approach searches for a transmission tree with the highest posterior probability while estimating additional parameters that govern the epidemiological and within-host dynamics. Two central components for the posterior probability calculation are the probability of observing a transmission tree given the epidemiological parameters and the probability of observing the phylogenetic tree given the within-host dynamics and transmission tree. The calculation for these two probabilities is based on the model for within-host dynamics and transmission dynamics. In their approach, the transmission dynamics are modeled by a susceptible-infected-removed (SIR) model, and this model is used in calculating the probability of a transmission tree given epidemiological parameters for transmission and recovery. Within-host dynamics are modeled as a neutral coalescent process with a single parameter that quantifies the within-host effective population size. The probability of observing the phylogenetic tree is obtained by multiplying the probability of observing each of the tree's subtrees, where each subtree represents evolution that has occurred

in a single host. Subtrees are, therefore, delimited by transmission events. The solution space of the transmission trees is explored by Markov Chain Monte Carlo (MCMC). Similar to the original work by Cottam and colleagues (Cottam et al., 2008), the approach developed by Didelot and colleagues is likelihood-based and relies on a single reconstructed phylogenetic tree as input data. However, as pointed out by Hall and coauthors 2015, inference based on a single phylogeny does not account for phylogenetic uncertainty. The authors, however, suggest that their approach can account for phylogenetic uncertainty by applying their method to the posterior samples of phylogenetic trees, which are commonly provided by the software package BEAST (Drummond and Rambaut, 2007).

While the approaches developed by Cottam and colleagues (Cottam et al., 2008) and Didelot and colleagues (Didelot et al., 2014) both relied on single, "pre-generated" phylogenetic trees to infer transmission trees, other inference approaches jointly reconstruct transmission trees and phylogenies based on the same underlying model. The earliest approach that both considered within-host diversity and co-estimated phylogenetic trees alongside transmission trees was the approach developed by Ypma and colleagues (Ypma et al., 2013). This approach relies on both epidemiological and pathogen genetic data to co-estimate the phylogenetic tree and the transmission tree. Their approach searches for the combination of transmission tree and phylogenetic tree, along with other parameters, that have the highest probability using MCMC. Distinct from Didelot and colleagues' MCMC approach to searching the space of transmission trees, Ypma and colleagues' approach has three components in the likelihood calculation. The first and second components are similar to the ones in the approach by Didelot and colleagues (Didelot et al., 2014), corresponding to the probability of observing the transmission tree and the probability of observing the phylogenetic tree. However, different models were used for the underlying transmission dynamics and within-host dynamics. Rather than subdividing a single phylogeny into

subtrees, each representing a single individual's within-host dynamics, Ypma and colleagues concatenated consecutive within-host genealogies by connecting a tip of the donor's tree to the root of the recipient tree. With this concatenation approach, there is an assumption of a complete transmission bottleneck. Along with these two likelihood components, the third component in Ypma and colleagues' approach is given by the probability of observing the observed sequences under a given phylogenetic tree. Calculation of this probability involves specifying a model of sequence evolution and estimating its parameters, including nucleotide substitution rates.

A second approach that co-estimates the phylogenetic tree alongside the transmission tree is an approach developed by Hall and colleagues (Hall et al., 2015). Similar to the approach taken by Ypma and colleagues, the phylogenetic tree, transmission tree, and other model parameters are estimated using three likelihood components. To calculate the probability of a phylogenetic tree given a transmission tree and other parameters, this approach subdivides the phylogenetic tree into subtrees, similar to the approach taken by Didelot and colleagues (Didelot et al., 2014). Hall and colleagues' approach, however, allows for multiple samples per host as long as they form a single monophyletic clade in the phylogeny. For the transmission dynamics, Hall and colleagues' approach uses individual-based modeling instead of compartmental epidemiological models. While using individual-based modeling allows for more realism and a straightforward way by which to incorporate host heterogeneity and non-random mixing of the host population, relying on this type of modeling introduces additional parameters and complexity. This, in turn, increases the computational effort to calculate likelihoods and, therefore, may limit the scalability of the approach.

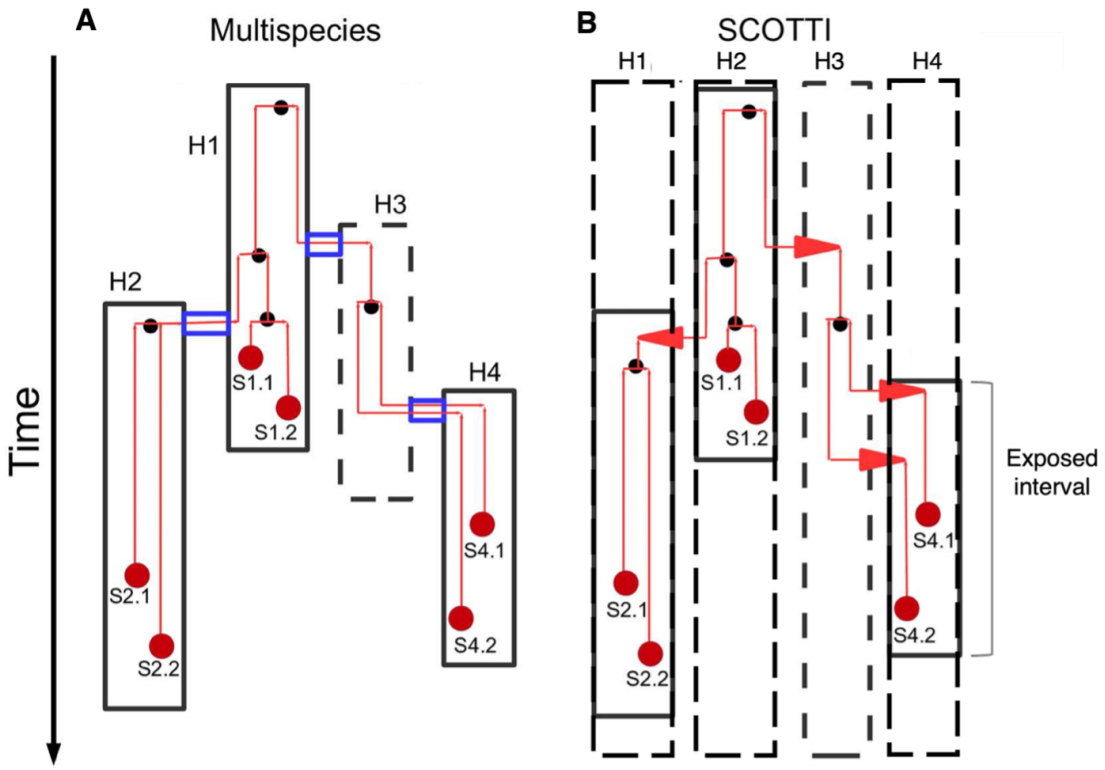

Figure 5.5: **Genealogy of pathogens sampled from different hosts under the multi-species coalescent model (A) versus the structured coalescent model (SCOTTI; B).** Red dots denote pathogen samples. Black dots denote coalescent events. Rectangles boxing off lineages represent hosts. For each host, the introduction and removal times (defining the exposed interval) are indicated with solid lines. Dashed lines indicate an unsampled host with an unknown exposed interval. (A) In the multi-species coalescent model, pathogens are assumed to be transmitted under a complete transmission bottleneck, shown with blue rectangles connecting hosts. (B) In a structured coalescent model, each host is modeled as a subpopulation that is present at the same time, as indicated in the dashed rectangle. However, migration to and from a host is restricted to the host's exposed interval. The migration between hosts is marked with red arrows. Multiple lineages can be introduced into a host, thus making this model more appropriate for pathogens with loose transmission bottlenecks and for scenarios of superinfection. Reproduced with modification from De Maio et al. (2016), published in PLoS Computational Biology under a Creative Commons Attribution 4.0 International (CC BY) license.

5.4.2 Within-unit diversity stemming from multiple infection and *de novo* mutation

When the source of within-unit diversity is limited to *de novo* mutations, the set of possible infectious units in which a pair of pathogens diverged is limited. This limited set is what permits the calculation of the likelihood with explicitly modeled within-unit dynamics. However, when within-unit diversity is present and does not stem from *de novo* evolution alone, the application of the approaches discussed in the previous section may exclude the true transmission history. The incorporation of multiple infections (either by superinfection or incomplete genetic bottlenecks) will expand the set of possible infectious units (generally, hosts) in which the coalescence of two pathogen lineages occurred. This expansion results from the possibility that lineage coalescence can occur in a host that is upstream in a transmission chain from the host that is the most recent infectious ancestor that is common to the pair of individuals that are sampled. Thus, to allow for within-unit diversity that may stem from multiple infections, approaches that are computationally less demanding are needed.

To allow multiple infections and, more generally, to allow more complex scenarios of transmission history, De Maio and colleagues developed a transmission tree inference approach based on a structured coalescent model (De Maio et al., 2016). This stands in contrast to the approaches detailed above that allow only for *de novo*-generated within-unit diversity (Didelot et al., 2014; Hall et al., 2015; Ypma et al., 2013). Those approaches essentially use a multi-species coalescent model (Rannala and Yang, 2003), where the pathogen population within a host is a separate population established by the transmission of a single pathogen lineage (Figure 5.5A). Following infection, an infectious unit can no longer receive additional lineages; it can only infect previously uninfected units. In contrast, a structured coalescent model considers infectious units as subpopulations, with migration that can repeatedly occur between

them. As such, the set of infectious units can be considered a meta-population. Under a structured coalescent model framework, transmission events are migration events between subpopulations. In their approach, De Maio and colleagues assume that all infectious units (subpopulations) have equal and constant population sizes (Figure 5.5B). Each host has an exposure interval that starts with an introduction time and ends with a removal time. Within this exposure interval, the pathogen population within a host migrates to other hosts at equal rates. Consideration of this process backward in time, migration (that is, transmission) of a pathogen provides an opportunity for a lineage in a host to coalesce with another lineage in another host. The way by which migration occurs between subpopulations allows for scenarios such as the transmission of multiple lineages from one host to another host (that is, an incomplete genetic bottleneck) and superinfection (infection of a host from multiple donors) that were not considered within the realm of possibilities under previous approaches.

Both the structured coalescent transmission tree reconstruction approach and the approaches before it rely on likelihood calculations for inference. Further, they are based on only a few (if not one) samples per infected host. With advances in sequencing technology and decreases in the cost of sequencing, more samples per host are now often available. More recently developed approaches have, therefore, by design, focused on more effectively exploiting this higher within-individual sampling effort (Dhar et al., 2022; Sashittal and El-Kebir, 2019, 2020; Wymant et al., 2017). The underlying idea of these approaches is similar to earlier approaches (Metzker et al., 2002; Esteban et al., 1996; Heinsen et al., 2000; Holmes et al., 1993) to infer the epidemiological relationship between two infected hosts. Based on the work by Romero-Severson and colleagues that shows how cladistic relationships (the PM, PP, and MM relationships described above) can be used to infer the transmission direction between two individuals (Romero-Severson et al., 2016), these recent methods use a

parsimony-based approach to reconstruct the transmission tree.

The approach by Wymant and colleagues relies on data sets with multiple samples per host generated either by deep-sequencing technology or the longitudinal sampling of a single host (Wymant et al., 2017). Their approach reconstructs the transmission tree by assigning hosts to internal nodes of the phylogenetic tree. The assignment of a host is done by minimizing the number of infections with the Sankoff algorithm (Sankoff, 1975), which is typically used for ancestor reconstruction for a phylogenetic tree. However, they modified the algorithm so that it can handle the 'unassigned' state where the host is suspected to be outside of the samples or the topological signal for a host assignment is ambiguous. In addition, to address the limitation of the Sankoff algorithm, which may generate an unrealistic scenario of a single introduction where multiple introductions are plausible, they introduced penalties for high within-host diversity. This results in the assignment of two lineage introductions (instead of a single introduction) when the genetic distance between two samples from a host exceeds a given threshold. In this approach, multiple phylogenetic trees that are generated from different regions of the pathogen genome can be assigned to a host and integrated over to reconstruct a transmission tree.

Several other recently developed approaches also rely on parsimony for transmission tree reconstruction. While Wymant and colleagues' approach reconstructs a single host assignment for each phylogenetic tree, an approach developed by Dhar and colleagues (Dhar et al., 2022) acknowledges that there could be many optimal host assignments that result in the same minimal number of infections. Thus, among those optimal host assignments, they select the trees with the minimum number of back-transmissions (that is, reinfections). A similar approach developed by Sashittal and El-Kebir (Sashittal and El-Kebir, 2019) specifically aims to account for a loose transmission bottleneck. While allowing for co-transmission of multiple lineages via a loose transmission bottleneck, they search for the most parsimonious tree. In contrast

to the Dhar and colleagues' approach (which searches for trees with the minimum number of back-transmissions among trees with the minimum number of infections), Sashittal and El-Kebir's approach minimizes the number of co-transmissions from trees with the minimum number of infections.

5.5 Perspectives

5.5.1 Inference methods rely on different assumptions, approaches, and data

Although transmission trees and phylogenetic trees differ from one another conceptually, phylogenetic trees contain valuable information and could be used in transmission tree reconstruction. Specifically, these two trees can be reconciled with one another through host assignment to internal nodes of a phylogenetic tree. While some approaches assign hosts to internal nodes based on a single, 'pre-generated' phylogenetic tree, other approaches jointly infer the transmission tree and the phylogenetic tree. Above, I reviewed phylogeny-based approaches for inferring transmission histories according to the assumptions these approaches adopt regarding the source of genetic variation within a host (or infectious unit). Early approaches did not consider within-host genetic variation, assuming that pathogen genetic differences arose during the process of transmission (Cottam et al., 2008). Later approaches allowed for within-host pathogen genetic variation generated by *de novo* mutation and/or infection with multiple lineages (De Maio et al., 2016; Dhar et al., 2022; Didelot et al., 2014; Hall et al., 2015; Sashittal and El-Kebir, 2019; Wymant et al., 2017; Ypma et al., 2013). Where within-host genetic diversity can stem from is a central assumption for the inference approaches, as it determines the set of host assignments that are compatible with a given phylogenetic tree.

The assumptions of simpler within-host dynamics limit the set of potential trans-

mission histories by limiting the potential host (or infectious unit) locations of ancestral pathogens. In Cottam et al. (2008), which does not consider within-host evolution, pathogens sampled from two different hosts must find their most recent common ancestor in one of the two hosts (Figure 5.4A). However, when within-host variation generated from *de novo* mutations is allowed, this opens a new possibility where the most recent common ancestor existed in another host that transmitted to both of the sampled hosts (Figure 5.4B). As such, this increases the potential location of the ancestor pathogen from two to three hosts, expanding the space of the transmission tree. The incorporation of unsampled hosts and multiple infections further adds complexities to the relationship between a phylogenetic tree and a transmission tree.

In the likelihood-based approaches reviewed (De Maio et al., 2016; Didelot et al., 2014; Hall et al., 2015; Ypma et al., 2013), the likelihood of a transmission tree is calculated based on explicitly modeled within-host dynamics. This makes the inference computationally expensive. In addition, these approaches must, therefore, specify assumptions regarding the underlying within-host dynamics. However, these dynamics are not always well characterized, and simplifications are adopted for computational ease. Commonly, these approaches assume that the within-host evolutionary dynamics are governed by the Kingman coalescent (Kingman, 1982), which is clearly an oversimplification for pathogens with complex within-host life cycles and dynamics. Further, because these likelihood-based approaches model transmission dynamics given the epidemiological data, they require epidemiological data that may, at times, be missing or inaccurate.

More recent parsimony-based inference approaches (Dhar et al., 2022; Sashittal and El-Kebir, 2019; Wymant et al., 2017), however, have other limitations. First, because they infer transmission histories from the within-host diversity of hosts, they require multiple samples per host, either by longitudinal sampling from a host or by deep-sequencing. Although this type of data is becoming more readily available due

to advances in deep-sequencing technology, it is not always available. In addition, although different parsimony algorithms are used by different approaches, there are no clear criteria to determine the most reasonable parsimony algorithms.

5.5.2 Choice of inference method to use should be based on data characteristics

Which of the above-described approaches should be used for transmission tree inference based on several factors? The first is the type of genetic data required for the inference method. For instance, all of the parsimony-based approaches reviewed here (Dhar et al., 2022; Sashittal and El-Kebir, 2019; Wymant et al., 2017) require multiple samples from an infected host, and the presence of within-host pathogen genetic diversity. Although these types of data are becoming increasingly available thanks to the falling cost of sequencing, they might not be available in all cases. In addition, within-host diversity might require different approaches to be appropriately captured. For instance, while short reads in deep sequencing data could capture the within-host diversity in HIV-1 infections due to the high mutation rate of this virus (Wymant et al., 2017), short reads may not be able to capture the within-host diversity of a Staphylococcus aureus infection, resulting from a bacterial pathogen with a low mutation rate. This is because the read length may be too short to provide enough opportunity for mutations to occur (Hall et al., 2019). In this case, the within-host diversity needs to be captured by sequencing multiple colony picks (Hall et al., 2019).

In addition to the type of sequence data, another factor that should be considered is the completeness of the sequence and the presence of epidemiological data. Each inference method has different assumptions regarding the observation and sampling of hosts. Most of the inference methods reviewed here (Cottam et al., 2008; Dhar et al., 2022; Didelot et al., 2014; Sashittal and El-Kebir, 2019; Ypma et al., 2013) assume that all hosts are observed and sequenced. Without knowing that there is a missing

host, algorithms used in these approaches may assign an incorrect host to an internal node. Thus, violation of these assumptions may lead to incorrect host assignment and reconstruction of the transmission history.

Biological characteristics, including the life cycles of a pathogen, also need to be considered. For instance, pathogens with frequent co-infection and weak transmission bottlenecks may violate the assumptions regarding the source of within-host diversity. When pathogens are located in different body compartments, the sampling sites and the compartment that is primarily responsible for the transmission of pathogens should also be considered so that within-host diversity can be appropriately captured.

5.5.3 Systematic comparisons are needed to evaluate the performance of inference approaches

Due to differences in the types of data used and differences in the underlying transmission assumptions, systematic comparisons between different methods are challenging. This is, in part, may be the reason why there are so few systematic comparisons that exist that have evaluated the relative performance of transmission tree inference approaches. Of the comparisons that have been conducted, one examined both phylogeny-based and genetic distance-based approaches (Firestone et al., 2019). In this study, they used simulated datasets for a foot-and-mouth disease outbreak to compare the accuracy of different approaches for the inference of transmission networks between farms and substitution rates. To assess the robustness of the inference methods to unsampled premises, comparisons were made with all genomic data available and again with only 50% of the genomic data available. When the genomic data were available for all premises, the genetic distance-based approach by Lau and colleagues (Lau et al., 2015) showed the highest accuracy among the nine compared approaches, correctly identifying the source of 73% of the infected premises (Firestone et al., 2019). A modified version of Cottam and colleagues' phylogeny-based approach came in second

in terms of performance. Lau and colleagues' approach showed the best performance when only 50% of infected farms were sampled.

A second study that aimed to conduct a systematic comparison across inference methods examined the performance of approaches, including both phylogeny-based and distance-based approaches. In the context of a low genetic diversity outbreak of Mycobacterium tuberculosis (Sobkowiak et al., 2022). In this study, both simulated and real-world data were used in the comparison. In this analysis, Outbreaker2 (Campbell et al., 2018) performed best, with the highest sensitivity, defined as the proportion of true transmission links that were correctly identified.

Although these two studies provided valuable insight regarding the performance of different inference approaches, more systematic comparisons are needed. In particular, it is not well studied how uncertainty and incompleteness of genetic and epidemiological data affect transmission tree inference. Among inference methods that assume complete observation and sampling of infected hosts (Cottam et al., 2008; Dhar et al., 2022; Didelot et al., 2014; Sashittal and El-Kebir, 2019; Ypma et al., 2013), the method suggested by Ypma and colleagues is the only study that included assessment for how the approach performs in the case of unobserved or unsampled data. This lack of knowledge regarding the potential biases due to violation of assumptions limits the applicability of inference methods to very limited circumstances. More comparative studies would facilitate the appropriate application of inference methods on empirical datasets. Further, because different pathogens have different characteristics, any systematic comparison should consider various datasets as one benchmarking test based on a pathogen with some characteristics that may not be reflective of performance on data from another pathogen with different characteristics.

5.6 Conclusion

Although transmission trees and phylogenetic trees are conceptually different from one another, phylogenetic trees have the potential to inform transmission histories, complementing traditional epidemiological data. With more studies that systematically assess the performance of inference approaches under various conditions, researchers will be better guided to choose an inference method that is most appropriate for the type of data that is available to them and for the specific characteristics of their pathogen of study.

5.7 Supplementary information

5.7.1 Tree representations of transmission history

During the process of pathogens spreading through a population, pathogens transmit from infected hosts to susceptible hosts. Each transmission can be considered a transmission link between hosts, with sequential transmissions forming a chain of transmission. These chains of transmission can be represented by a rooted tree structure when every infected host is infected by only a single infectious contact and when re-infection does not occur (Welch, 2011). A complete transmission tree exhaustively summarizes an outbreak, including the identity of every host infected during the outbreak and the time of infection and recovery of each infected host (Welch, 2011). In a complete transmission tree (Figure 5.6A), the timeline of infection is shown for each infected individual, with a connecting horizontal line between the time of infection and the time of recovery of a single individual. Transmission of the infection is shown as a branching event, stemming from the donor of the infection to the recipient of the infection, which is represented as a new horizontal line with its timeline of infection.

Depicted transmission trees can also be incomplete, omitting some degree of detail, for example, the timing of transmission events or the timing of infection and recovery (Figure 5.6B). This simplified depiction of transmission dynamics occurring during an outbreak has been termed a "beanbag" tree (De Maio et al., 2016). In beanbag trees, each node represents an infectious unit, and each directional edge between a pair of connected nodes represents a transmission event. Depending on the scale of the study, an infectious unit can be an individual or a premise (e.g., a household or a farm).

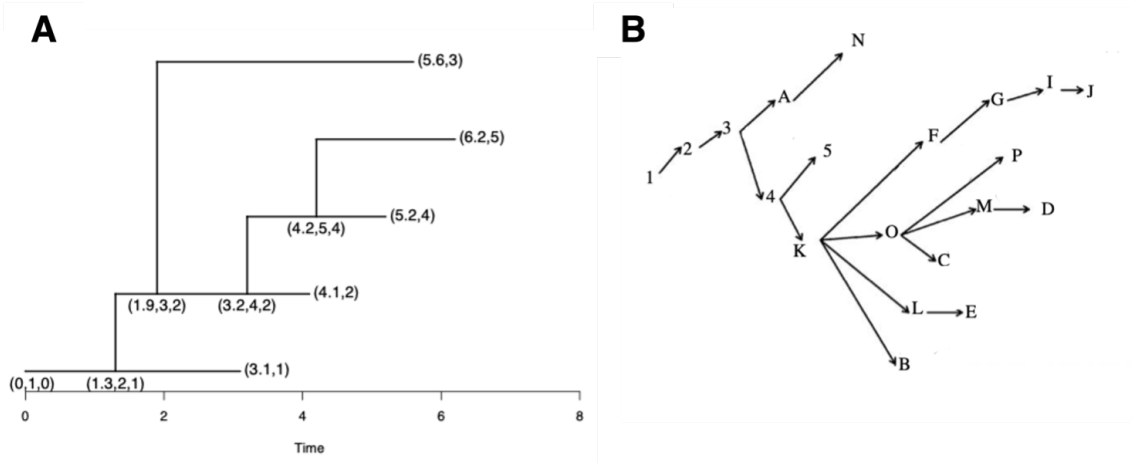

Figure 5.6: **Transmission tree depictions.** (A) A complete transmission tree, showing infection times and recovery times of all individuals infected during an outbreak, along with information on who-infected-whom. Each horizontal line indicates the timeline of infection, from the time of infection (start of the horizontal line) to the time of recovery (end of the horizontal line). Here, recovery events are labeled with the time of recovery and the identity number of the host, respectively. The branching of the tree indicates the transmission event, and the associated internal node is labeled with the transmission time, the recipient of the transmission, and the donor of the transmission, respectively. Reproduced from Welch (2011) with permission (Copyright (2011) Welch). (B) An example of an incomplete ("bean-bag") transmission tree. Hosts are labeled with numbers or letters. Reproduced from Cottam et al. (2008) with permission (Copyright (2008) Royal Society B: Biological Sciences); permission conveyed through Copyright Clearance Center, Inc.

5.7.2 Reconstructing and dating phylogenetic trees

What is a phylogenetic tree?

Phylogenetic trees are a key concept in the field of evolutionary biology. They describe the evolutionary relationships between different sampled taxa (e.g., species, individuals). The structure of phylogenetic trees includes nodes and edges. The external nodes are also called 'tips' or 'leaves', and the edges are also called 'branches.' Sampled taxa, located at the tips of a tree, are related to one another through evolutionary descent from common ancestors. These common ancestors are depicted as internal nodes. The length of branches connecting nodes reflects the amount of evolutionary change between them.

Reconstructing a phylogenetic tree from sequence data

While a phylogeny can be reconstructed based on many different types of character data, one of the most commonly used data types is nucleotide sequences. Each site in a nucleotide sequence serves as a character, and the four nucleotides are the possible character states that provide information to infer the evolutionary relationships between taxa. Reconstruction of a phylogenetic tree from nucleotide sequence data starts with the alignment of viral sequences to allow for comparison across homologous sites (Figure 5.7). For the phylogenetic tree reconstruction, two different approaches are available: distance-based approaches and optimality approaches.

Distance-based approaches use calculated distances between pairs of sequences to reconstruct a phylogeny. The distances between sequences are calculated based on a sequence evolution model that describes the probability of substitution from one nucleotide to another. The simplest model of sequence evolution is JC69 (Jukes and Cantor, 1969), where all substitutions occur with equal probability. One example of a distance-based approach is the least-squares method, which searches for a tree

that has the minimum difference between the expected difference of the tree and the observed difference in sequence pairs. Other distance-based approaches include minimum evolution, UPGMA, and neighbor-joining approaches.

Optimality approaches instead consider each nucleotide site as a separate character, and all nucleotide sites are considered for each of the sampled sequences. Optimality approaches include maximum parsimony, maximum likelihood, and Bayesian methods. The maximum parsimony approach searches for a tree that minimizes the number of evolutionary changes needed to explain the observed sequences. This approach does not explicitly incorporate a model of sequence evolution. In contrast, maximum likelihood and Bayesian methods reconstruct phylogenies using likelihood calculations that involve models of sequence evolution. More specifically, the likelihood is given by the probability of observing a set of sampled sequences, which can be calculated under a specified phylogeny (including topology and branch lengths) and a parameterized model of sequence evolution.

Reconstructing a phylogenetic tree from sequence data

In reconstructed phylogenetic trees, branch lengths often represent the genetic distance between nodes and thus cannot be directly used to infer the temporal ordering of branching events. Branch lengths, however, can be converted into units of time, providing a common frame of reference to compare the branching events in a phylogenetic tree to epidemiological events of interest, such as transmission events. The conversion relies on molecular clock models, which are based on the assumption that the genetic difference between two taxa is proportional to the time since their divergence. Using the genetic differences between taxa and their sampling times, the substitution rate can be estimated, and this estimated substitution rate is used to convert branch lengths from units of genetic differences to units of time. This is done by dividing genetic differences by the substitution rate. Before time-resolving (otherwise known

as "dating") a phylogenetic tree, it is recommended to test whether enough temporal signal is present in the dataset through approaches such as root-to-tip regression (Korber et al., 2000; Rambaut et al., 2016).

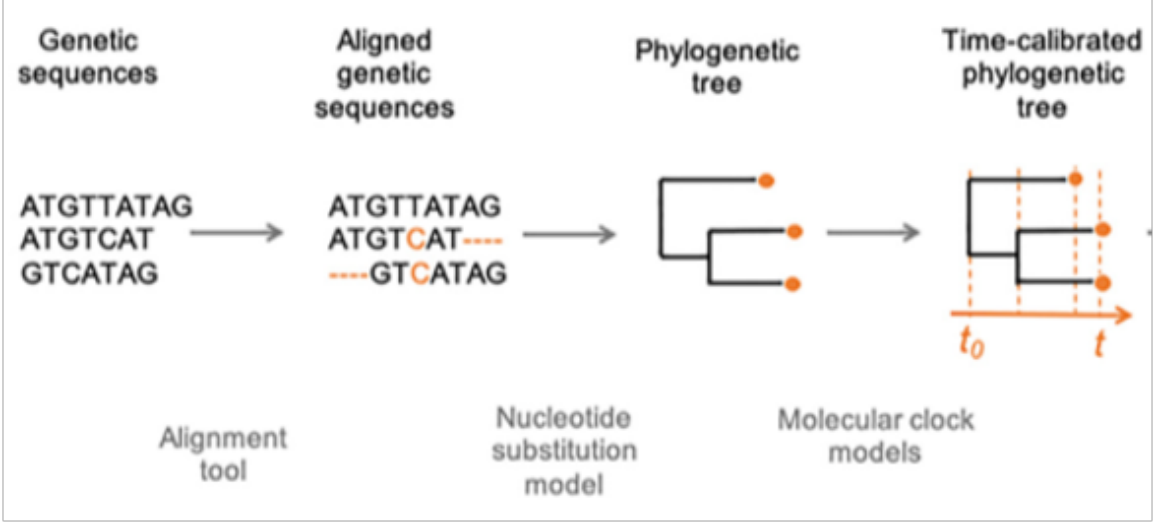

Figure 5.7: **Workflow for phylogenetic tree reconstruction and time-calibration of a phylogenetic tree.** Reproduced with modification from Guinat et al. (2021) with permission (Copyright (2021) Trends in ecology & evolution); permission conveyed through Copyright Clearance Center, Inc.

5.7.3 Supplementary Table

Methods	Category	Assumptions regarding within-host diversity	Features
Cottam et al. (2008)	Two-step, likelihood	No within-host diversity	Does not explicitly model within-host dynamics
Ypma et al. (2013)	One-step, likelihood	De novo mutation	Concatenates within-host genealogy according to transmission history
Didelot et al. (2014) (Transphylo)	Two-step, likelihood	De novo mutation	Uses a multispecies coalescent model for within-host dynamics Labels tips of time-resolved phylogenetic tree removal time
Hall et al. (2015) (beastlier)	One-step, likelihood	De novo mutation	Allows for unsampled hosts Limited to infection started by a single pathogen lineage
De Maio et al. (2016) (SCOTTI)	One-step, likelihood	De novo mutation & multiple infections	Allows for unsampled, unobserved hosts Uses a structured coalescent model
Wymant et al. (2017) (Phyloscanner)	Two-step, parsimony	De novo mutation & multiple infections	Does not explicitly model within-host dynamics Penalizes within-host diversity to identify multiple infections
Dhar et al. (2022) (Tnet)	Two-step, parsimony	De novo mutation & multiple infections	Does not explicitly model within-host dynamics Minimizes the number of back-transmissions
Sashittal and El-Kebir (2019) (SharpTNI)	Two-step, parsimony	De novo mutation & multiple infections	Does not explicitly model within-host dynamics Aims to address weak bottleneck Minimize the number of co-transmission

Table 5.1: Statistical tests for one-dimensional summary statistics

Chapter 5 References

F. Campbell, X. Didelot, R. Fitzjohn, N. Ferguson, A. Cori, and T. Jombart. outbreaker2: a modular platform for outbreak reconstruction. *BMC Bioinformatics*, 19(S11), Oct. 2018. doi: 10.1186/s12859-018-2330-z. URL https://doi.org/10.1186/s12859-018-2330-z.

S. Cauchemez and N. M. Ferguson. Methods to infer transmission risk factors in complex outbreak data. *Journal of The Royal Society Interface*, 9(68):456–469, Aug. 2011. doi: 10.1098/rsif.2011.0379. URL https://doi.org/10.1098/rsif.2011.0379.

E. M. Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, and D. T. Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, 275(1637):887–895, Jan. 2008. ISSN 1471-2954. doi: 10.1098/rspb.2007.1442. URL http://dx.doi.org/10.1098/rspb.2007.1442.

N. De Maio, C.-H. Wu, and D. J. Wilson. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLOS Computational Biology*, 12(9):e1005130, Sept. 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005130. URL http://dx.doi.org/10.1371/journal.pcbi.1005130.

S. Dhar, C. Zhang, I. I. Mandoiu, and M. S. Bansal. TNet: Transmission Network Inference Using Within-Host Strain Diversity and its Application to Geographical

Tracking of COVID-19 Spread. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):230–242, Jan. 2022. ISSN 2374-0043. doi: 10.1109/tcbb.2021.3096455. URL http://dx.doi.org/10.1109/TCBB.2021.3096455.

X. Didelot, J. Gardy, and C. Colijn. Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. *Molecular Biology and Evolution*, 31 (7):1869–1879, Apr. 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu121. URL http://dx.doi.org/10.1093/molbev/msu121.

A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214, 2007. doi: 10.1186/1471-2148-7-214. URL https://doi.org/10.1186/1471-2148-7-214.

J. I. Esteban, J. Gómez, M. Martell, B. Cabot, J. Quer, J. Camps, A. González, T. Otero, A. Moya, R. Esteban, and J. Guardia. Transmission of Hepatitis C Virus by a Cardiac Surgeon. *New England Journal of Medicine*, 334(9):555–561, Feb. 1996. ISSN 1533-4406. doi: 10.1056/nejm199602293340902. URL http://dx.doi.org/10.1056/NEJM199602293340902.

S. M. Firestone, Y. Hayama, R. Bradhurst, T. Yamamoto, T. Tsutsui, and M. A. Stevenson. Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. *Scientific Reports*, 9(1), Mar. 2019. doi: 10.1038/s41598-019-41103-6. URL https://doi.org/10.1038/s41598-019-41103-6.

C. Guinat, T. Vergne, A. Kocher, D. Chakraborty, M. C. Paul, M. Ducatez, and T. Stadler. What can phylodynamics bring to animal health research? *Trends in Ecology & Evolution*, 36(9):837–847, Sept. 2021. ISSN 0169-5347. doi: 10.1016/j.tree.2021.04.013. URL http://dx.doi.org/10.1016/j.tree.2021.04.013.

M. Hall, M. Woolhouse, and A. Rambaut. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Computational*

Biology, 11(12):e1004613, Dec. 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004613. URL http://dx.doi.org/10.1371/journal.pcbi.1004613.

M. HALL, M. WOOLHOUSE, and A. RAMBAUT. Using genomics data to reconstruct transmission trees during disease outbreaks. *Revue Scientifique et Technique de l'OIE*, 35(1):287–296, Apr. 2016. doi: 10.20506/rst.35.1.2433. URL https://doi.org/10.20506/rst.35.1.2433.

M. D. Hall, M. T. Holden, P. Srisomang, W. Mahavanakul, V. Wuthiekanun, D. Limmathurotsakul, K. Fountain, J. Parkhill, E. K. Nickerson, S. J. Peacock, and C. Fraser. Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *eLife*, 8, Oct. 2019. doi: 10.7554/elife.46402. URL https://doi.org/10.7554/elife.46402.

A. Heinsen, F. Bendtsen, and A. Fomsgaard. A phylogenetic analysis elucidating a case of patient-to-patient transmission of hepatitis C virus during surgery. *Journal of Hospital Infection*, 46(4):309–313, Dec. 2000. ISSN 0195-6701. doi: 10.1053/jhin.2000.0842. URL http://dx.doi.org/10.1053/jhin.2000.0842.

E. C. Holmes, L. Q. Zhang, P. Simmonds, A. S. Rogers, and A. J. L. Brown. Molecular Investigation of Human Immunodeficiency Virus (HIV) Infection in a Patient of an HIV-Infected Surgeon. *Journal of Infectious Diseases*, 167(6):1411–1414, June 1993. ISSN 1537-6613. doi: 10.1093/infdis/167.6.1411. URL http://dx.doi.org/10.1093/infdis/167.6.1411.

E. C. Holmes, S. Nee, A. Rambaut, G. P. Garnett, and P. H. Harvey. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 349 (1327):33–40, July 1995. doi: 10.1098/rstb.1995.0088. URL https://doi.org/10.1098/rstb.1995.0088.

T. Jombart, R. M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, June 2010. ISSN 1365-2540. doi: 10.1038/hdy.2010.78. URL http://dx.doi.org/10.1038/hdy.2010.78.

T. H. Jukes and C. R. Cantor. Evolution of Protein Molecules. *Mammalian Protein Metabolism*, pages 21–132, 1969. doi: 10.1016/b978-1-4832-3211-9.50009-7.

J. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248, Sept. 1982. doi: 10.1016/0304-4149(82)90011-4. URL https://doi.org/10.1016/0304-4149(82)90011-4.

B. Korber, M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. Timing the Ancestor of the HIV-1 Pandemic Strains. *Science*, 288(5472):1789–1796, June 2000. doi: 10.1126/science.288.5472.1789. URL https://doi.org/10.1126/science.288.5472.1789.

M. S. Lau, G. Marion, G. Streftaris, and G. Gibson. A systematic Bayesian integration of epidemiological and genetic data. *PLoS computational biology*, 11(11):e1004633, 2015.

T. Leitner. Phylogenetics in HIV transmission. *Current Opinion in HIV and AIDS*, 14(3):181–187, May 2019. doi: 10.1097/coh.0000000000000536. URL https://doi.org/10.1097/coh.0000000000000536.

T. Leitner and J. Albert. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proceedings of the National Academy of Sciences*, 96(19):10752–10757, Sept. 1999. doi: 10.1073/pnas.96.19.10752. URL https://doi.org/10.1073/pnas.96.19.10752.

T. Leitner, D. Escanilla, C. Franzén, M. Uhlén, and J. Albert. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings*

of the National Academy of Sciences, 93(20):10864–10869, Oct. 1996. doi: 10.1073/pnas.93.20.10864. URL https://doi.org/10.1073/pnas.93.20.10864.

K. A. Lythgoe, M. Hall, L. Ferretti, M. De Cesare, G. MacIntyre-Cockett, A. Trebes, M. Andersson, N. Otecko, E. L. Wise, N. Moore, J. Lynch, S. Kidd, N. Cortes, M. Mori, R. Williams, G. Vernet, A. Justice, A. Green, S. M. Nicholls, M. A. Ansari, L. Abeler-Dörner, C. E. Moore, T. E. A. Peto, D. W. Eyre, R. Shaw, P. Simmonds, D. Buck, J. A. Todd, T. R. Connor, S. Ashraf, A. Da Silva Filipe, J. Shepherd, E. C. Thomson, D. Bonsall, C. Fraser, and T. Golubchik. SARS-CoV-2 within-host diversity and transmission. *Science*, 372(6539), Mar. 2021. doi: 10.1126/science.abg0821. URL https://doi.org/10.1126/science.abg0821.

W. P. Maddison and L. L. Knowles. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55(1):21–30, Feb. 2006. doi: 10.1080/10635150500354928. URL https://doi.org/10.1080/10635150500354928.

M. A. Martin and K. Koelle. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". *Sci. Transl. Med.*, 13(617):eabh1803, Oct. 2021.

J. T. McCrone and A. S. Lauring. Genetic bottlenecks in intraspecies virus transmission. *Current Opinion in Virology*, 28:20–25, Nov. 2017. doi: 10.1016/j.coviro.2017.10.008. URL https://doi.org/10.1016/j.coviro.2017.10.008.

J. T. McCrone, R. J. Woods, E. T. Martin, R. E. Malosh, A. S. Monto, and A. S. Lauring. Stochastic processes constrain the within and between host evolution of influenza virus. *Elife*, 7, May 2018.

M. L. Metzker, D. P. Mindell, X.-M. Liu, R. G. Ptak, R. A. Gibbs, and D. M. Hillis. Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the*

National Academy of Sciences, 99(22):14292–14297, Oct. 2002. ISSN 1091-6490. doi: 10.1073/pnas.222522599. URL http://dx.doi.org/10.1073/pnas.222522599.

O. G. Pybus and A. Rambaut. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10(8):540–550, Aug. 2009. ISSN 1471-0064. doi: 10.1038/nrg2583. URL http://dx.doi.org/10.1038/nrg2583.

A. Rambaut, T. T. Lam, L. M. Carvalho, and O. G. Pybus. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, 2(1):vew007, Jan. 2016. doi: 10.1093/ve/vew007. URL https://doi.org/10.1093/ve/vew007.

B. Rannala and Z. Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple LOCI. *Genetics*, 164(4):1645–1656, Aug. 2003. doi: 10.1093/genetics/164.4.1645. URL https://doi.org/10.1093/genetics/164.4.1645.

E. Romero-Severson, H. Skar, I. Bulla, J. Albert, and T. Leitner. Timing and Order of Transmission Events Is Not Directly Reflected in a Pathogen Phylogeny. *Molecular Biology and Evolution*, 31(9):2472–2482, May 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu179. URL http://dx.doi.org/10.1093/molbev/msu179.

E. O. Romero-Severson, I. Bulla, and T. Leitner. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, 113(10):2690–2695, Feb. 2016. doi: 10.1073/pnas.1522930113. URL https://doi.org/10.1073/pnas.1522930113.

N. A. Rosenberg and M. Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5):380–390, May 2002. doi: 10.1038/nrg795. URL https://doi.org/10.1038/nrg795.

D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, Jan. 1975. doi: 10.1137/0128004. URL https://doi.org/10.1137/0128004.

P. Sashittal and M. El-Kebir. SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck. Nov. 2019. doi: 10.1101/842237. URL http://dx.doi.org/10.1101/842237.

P. Sashittal and M. El-Kebir. Sampling and summarizing transmission trees with multi-strain infections. *Bioinformatics*, 36(Supplement_1):i362–i370, May 2020. doi: 10.1093/bioinformatics/btaa438. URL https://doi.org/10.1093/bioinformatics/btaa438.

B. Sobkowiak, K. Romanowski, I. Sekirov, J. L. Gardy, and J. Johnston. Comparing transmission reconstruction models with Mycobacterium tuberculosis whole genome sequence data. *bioRxiv (Cold Spring Harbor Laboratory)*, Jan. 2022. doi: 10.1101/2022.01.07.475333. URL https://doi.org/10.1101/2022.01.07.475333.

A. L. Valesano, K. E. Rumfelt, D. E. Dimcheff, C. N. Blair, W. J. Fitzsimmons, J. G. Petrie, E. T. Martin, and A. S. Lauring. Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *bioRxiv (Cold Spring Harbor Laboratory)*, Jan. 2021. doi: 10.1101/2021.01.19.427330. URL https://doi.org/10.1101/2021.01.19.427330.

D. Welch. Is network clustering detectable in transmission trees? *Viruses*, 3(6):659–676, June 2011. doi: 10.3390/v3060659. URL https://doi.org/10.3390/v3060659.

C. J. Worby and T. D. Read. "SEEDY" (Simulation of Evolutionary and Epidemiological Dynamics): An R Package to Follow Accumulation of Within-Host Mutation in Pathogens. *Plos One*, 10(6):e0129745, June 2015. ISSN 1932-6203. doi:

10.1371/journal.pone.0129745. URL http://dx.doi.org/10.1371/journal.pone.0129745.

C. J. Worby, M. Lipsitch, and W. P. Hanage. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLoS Computational Biology*, 10(3):e1003549, Mar. 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003549. URL http://dx.doi.org/10.1371/journal.pcbi.1003549.

C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, and C. Fraser. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Molecular Biology and Evolution*, 35(3):719–733, Nov. 2017. ISSN 1537-1719. doi: 10.1093/molbev/msx304. URL http://dx.doi.org/10.1093/molbev/msx304.

R. J. Ypma, M. Jonges, A. Bataille, A. Stegeman, G. Koch, M. Van Boven, M. Koopmans, W. M. Van Ballegooijen, and J. Wallinga. Genetic data provide evidence for Wind-Mediated transmission of highly pathogenic avian influenza. *The Journal of Infectious Diseases*, 207(5):730–735, Dec. 2012. doi: 10.1093/infdis/jis757. URL https://doi.org/10.1093/infdis/jis757.

R. J. F. Ypma, W. M. van Ballegooijen, and J. Wallinga. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. *Genetics*, 195(3):1055–1062, Nov. 2013. ISSN 1943-2631. doi: 10.1534/genetics.113.154856. URL http://dx.doi.org/10.1534/genetics.113.154856.

Chapter 6

Conclusion

Traditionally, infectious disease surveillance has relied primarily on case data. However, the increasing availability of genome sequences has opened new opportunities to better understand infectious disease dynamics, especially when case detection is incomplete due to asymptomatic infections or when reporting rates are unstable due to limited testing capacity. This is possible because genome sequences contain the footprints of past events in the form of genetic variations. By analyzing patterns of genetic variation in sampled genomes, phylodynamics, and other genome-based approaches have uncovered diverse aspects of viral spread and evolution. Furthermore, the unprecedented scale of sampling and rapid sharing of genome sequences during the COVID-19 pandemic set the foundation for both pandemic-scale phylodynamic inferences and genome-based surveillance during very early spread. These new applications present unique challenges that require further methodological development and evaluation. This thesis aimed to further understanding of genome-based approaches, especially during the early spread of newly emerged viruses.

In **Chapter 2**, I focused on the challenges due to the low level of genetic variation during the early spread of the virus and proposed a novel tree-free approach that estimates epidemiological parameters from the segregation site trajectory. Since

this approach does not rely on phylogenetic reconstruction, it circumvents the need to integrate over phylogenetic uncertainty, which can be computationally intensive when genetic diversity is low. An additional benefit is that this approach uses a particle-filtering algorithm. Due to the 'plug-and-play' property of particle filtering, the underlying simulation model used here could be replaced with other models of interest.

In **Chapter 3**, I investigated how generation interval misspecification affects phylodynamic estimation of the reproduction number during early exponential growth. While the impact of misspecification on case-based inference was recognized earlier by Wallinga and Lipsitch (2007), less attention has been paid to phylodynamic approaches. I compared phylodynamic estimates under three generation interval distributions and demonstrated that ignoring the distribution's shape and focusing solely on the mean generation interval can lead to an underestimation of the basic reproduction number. Notably, this underestimation was not observed when the growth rate was matched, even under an exponential distribution. However, currently, few phylodynamic approaches can incorporate flexible generation interval distributions.

In **Chapter 4**, I examined epidemiologically clustered sequences. I demonstrated that clustered sequences in a dataset can lead to an underestimation of the epidemic growth rate. I then showed that one-dimensional summary statistics cannot effectively capture differences between randomly sampled datasets and non-randomly sampled datasets with clustered sequences. However, the distributions of pairwise tMRCA and pairwise nucleotide diversity that have higher dimensions could capture signatures from pairs within the same clusters. Further work is needed to develop an approach that uses these summary statistics to identify non-randomness in datasets. Additionally, since this work relies on a discrete-time model, expansion to a continuous-time model is necessary for better applicability to real-world data.

In **Chapter 5**, I reviewed the relationship between the phylogenetic tree and the

transmission tree. Even long before the phylodynamic approaches were established, pathogen phylogenies have been used to better understand the transmission dynamics. The earlier approaches relied on the phylogenetic relationships between sampled pathogen sequences to identify the source of infection (Esteban et al., 1996; Heinsen et al., 2000; Holmes et al., 1993; Metzker et al., 2002). However, it was later that recognized that the transmission tree and the phylogenies are conceptually different entities (Jombart et al., 2010; Pybus and Rambaut, 2009; Romero-Severson et al., 2014; Ypma et al., 2013). Based on this recognition, a number of approaches have been suggested (Cottam et al., 2008; Ypma et al., 2013; Didelot et al., 2014; Hall et al., 2015; De Maio et al., 2016; Wymant et al., 2017; Dhar et al., 2022; Sashittal and El-Kebir, 2019), which rely different type of data, assumptions and approaches. As such, I conclude that a systematic comparison, especially under the missing sample, is needed to facilitate the appropriate application of inference methods on empirical datasets. This chapter focuses on the 'who-infected-whom' level dynamics. However, in order to understand how pathogen phylogeny can be used to estimate the epidemiological dynamics, further investigation at the host population level is needed.

Viral genome sequencing is increasingly becoming routine in infectious disease surveillance and monitoring protocols worldwide, and viral genome sequences have become an important source of complementary information that enhances and extends traditional epidemiological case data. As such, these genomic data now serve as a critical component in public health-related decision-making, thus the need for robust and reliable inference approaches is increasing. Future work in this field should consider the potential biases introduced by violations of assumptions in genome-based inferences when applying existing approaches and interpreting the results. Additionally, more methodological developments are needed to better accommodate more realistic assumptions, which may become more feasible with the growing availability of genome sequence data.

Chapter 6 References

E. M. Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, and D. T. Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, 275(1637):887–895, Jan. 2008. ISSN 1471-2954. doi: 10.1098/rspb.2007.1442. URL http://dx.doi.org/10.1098/rspb.2007.1442.

N. De Maio, C.-H. Wu, and D. J. Wilson. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLOS Computational Biology*, 12(9):e1005130, Sept. 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005130. URL http://dx.doi.org/10.1371/journal.pcbi.1005130.

S. Dhar, C. Zhang, I. I. Mandoiu, and M. S. Bansal. TNet: Transmission Network Inference Using Within-Host Strain Diversity and its Application to Geographical Tracking of COVID-19 Spread. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):230–242, Jan. 2022. ISSN 2374-0043. doi: 10.1109/tcbb.2021.3096455. URL http://dx.doi.org/10.1109/TCBB.2021.3096455.

X. Didelot, J. Gardy, and C. Colijn. Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. *Molecular Biology and Evolution*, 31 (7):1869–1879, Apr. 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu121. URL http://dx.doi.org/10.1093/molbev/msu121.

J. I. Esteban, J. Gómez, M. Martell, B. Cabot, J. Quer, J. Camps, A. González,

T. Otero, A. Moya, R. Esteban, and J. Guardia. Transmission of Hepatitis C Virus by a Cardiac Surgeon. *New England Journal of Medicine*, 334(9):555–561, Feb. 1996. ISSN 1533-4406. doi: 10.1056/nejm199602293340902. URL http://dx.doi.org/10.1056/NEJM199602293340902.

M. Hall, M. Woolhouse, and A. Rambaut. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Computational Biology*, 11(12):e1004613, Dec. 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004613. URL http://dx.doi.org/10.1371/journal.pcbi.1004613.

A. Heinsen, F. Bendtsen, and A. Fomsgaard. A phylogenetic analysis elucidating a case of patient-to-patient transmission of hepatitis C virus during surgery. *Journal of Hospital Infection*, 46(4):309–313, Dec. 2000. ISSN 0195-6701. doi: 10.1053/jhin.2000.0842. URL http://dx.doi.org/10.1053/jhin.2000.0842.

E. C. Holmes, L. Q. Zhang, P. Simmonds, A. S. Rogers, and A. J. L. Brown. Molecular Investigation of Human Immunodeficiency Virus (HIV) Infection in a Patient of an HIV-Infected Surgeon. *Journal of Infectious Diseases*, 167(6):1411–1414, June 1993. ISSN 1537-6613. doi: 10.1093/infdis/167.6.1411. URL http://dx.doi.org/10.1093/infdis/167.6.1411.

T. Jombart, R. M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, June 2010. ISSN 1365-2540. doi: 10.1038/hdy.2010.78. URL http://dx.doi.org/10.1038/hdy.2010.78.

M. L. Metzker, D. P. Mindell, X.-M. Liu, R. G. Ptak, R. A. Gibbs, and D. M. Hillis. Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences*, 99(22):14292–14297, Oct. 2002. ISSN 1091-6490. doi: 10.1073/pnas.222522599. URL http://dx.doi.org/10.1073/pnas.222522599.

O. G. Pybus and A. Rambaut. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10(8):540–550, Aug. 2009. ISSN 1471-0064. doi: 10.1038/nrg2583. URL http://dx.doi.org/10.1038/nrg2583.

E. Romero-Severson, H. Skar, I. Bulla, J. Albert, and T. Leitner. Timing and Order of Transmission Events Is Not Directly Reflected in a Pathogen Phylogeny. *Molecular Biology and Evolution*, 31(9):2472–2482, May 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu179. URL http://dx.doi.org/10.1093/molbev/msu179.

P. Sashittal and M. El-Kebir. SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck. Nov. 2019. doi: 10.1101/842237. URL http://dx.doi.org/10.1101/842237.

J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2007.

C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, and C. Fraser. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Molecular Biology and Evolution*, 35(3):719–733, Nov. 2017. ISSN 1537-1719. doi: 10.1093/molbev/msx304. URL http://dx.doi.org/10.1093/molbev/msx304.

R. J. F. Ypma, W. M. van Ballegooijen, and J. Wallinga. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. *Genetics*, 195(3): 1055–1062, Nov. 2013. ISSN 1943-2631. doi: 10.1534/genetics.113.154856. URL http://dx.doi.org/10.1534/genetics.113.154856.