Phylodynamic inference for emerging viruses using segregating sites

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

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

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

[3] Emory-UGA Center of Excellence for Influenza Research and Surveillance (CEIRS), Atlanta GA, USA

*katia.koelle@emory.edu

Key words: phylodynamic inference; segregating sites; infectious disease modeling; SARS-CoV-2

Abstract

Epidemiological models are commonly fit to case data to estimate model parameters and to infer unobserved disease dynamics. More recently, epidemiological models have also been fit to viral sequence data using phylodynamic inference approaches that generally rely on the reconstruction of viral phylogenies. However, especially early on in an expanding viral population, phylogenetic uncertainty can be substantial and methods that require integration over this uncertainty can be computationally intensive. Here, we present an alternative approach to phylodynamic inference that circumvents the need for phylogenetic tree reconstruction. Our "tree-free" approach instead relies on quantifying the number of segregating sites observed in sets of sequences over time and using this trajectory of segregating sites to infer epidemiological parameters within a Sequential Monte Carlo (SMC) framework. Using forward simulations, we first show that epidemiological parameters and processes leave characteristic signatures in segregating site trajectories, demonstrating that these trajectories have the potential to be used for phylodynamic inference. We then show using mock data that our proposed approach accurately recovers key epidemiological quantities such as the basic reproduction number and the timing of the index case. Finally, we apply our approach to SARS-CoV-2 sequence data from France, estimating a reproductive number of approximately 2.2 and an introduction time of mid-January 2021, consistent with estimates from epidemiological surveillance data. Our findings indicate that "tree-free" phylodynamic inference approaches that rely on simple population genetic summary statistics can play an important role in estimating epidemiological parameters and reconstructing infectious disease dynamics, especially early on in an epidemic.

Introduction

Phylodynamic inference methods use viral 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 (Stadler and Bonhoeffer 2013; Popinga et al. 2014; Ratmann et al. 2017; Volz et al. 2017), ebola (Stadler et al. 2014; Vaughan et al. 2017; Volz and Siveroni 2018), dengue (Rasmussen et al. 2014), influenza (Rasmussen and Stadler), and most recently severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)(Danesh et al. 2020; Miller et al. 2020; Geidelberg et al. 2021). 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 parameter estimation for structured epidemiological models, for example, susceptible-exposed-infected-recovered (SEIR) models and models with spatial compartmentalization (Volz 2012; Volz and Siveroni 2018). Birth-death approaches (Stadler 2010; Stadler et al. 2012), where a birth in the context of infectious diseases corresponds to a new infection and death corresponds to a recovery from infection, instead carry other advantages, such as incorporating the role of demographic stochasticity in disease dynamics, which may be particularly important in emerging diseases that start with low infection numbers (Boskova et al. 2014). Both of these classes of phylodynamic inference approaches rely on time-resolved phylogenies and have been incorporated into the phylogenetics software package BEAST2 (Bouckaert et al. 2014: 2) to allow joint estimation of epidemiological parameters and dynamics while integrating over phylogenetic uncertainty (Stadler et al. 2013; Volz and Siveroni 2018). 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 is 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 phylodynamic inference method that is particularly appropriate to use when viral sequences are sampled over short time periods and when phylogenetic uncertainty is considerable. This method does not rely on time-resolved phylogenies to infer epidemiological parameters or to reconstruct patterns of viral spread. Instead, the “tree-free” method we propose here fits epidemiological models to time series of the number of segregating sites observed in a viral population that is sampled over time. Like existing coalescent-based approaches, the approach we propose here allows for structured infectious disease models to be considered in a straightforward “plug-and-play” manner. Like existing birth-death process approaches, it incorporates the effect that demographic noise may have 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 phylodynamic 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.

New Approaches

Mutations occur during viral replication within infected individuals and these have the potential to be transmitted. During the epidemiological spread of an emerging virus, 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 be estimated using observed viral evolutionary trajectories. Here, we develop a phylodynamic inference approach that fits compartmental epidemiological models to times series of a low-dimensional evolutionary summary statistic. Specifically, we use trajectories of the number of segregating sites from samples of the viral population taken over time for phylodynamic inference. In Materials and Methods, we provide details on the simulation of epidemiological models that incorporate viral evolution and thus can yield simulated time series of the number of segregating sites. We further describe our phylodynamic inference approach that relies on using particle filtering (otherwise known as Sequential Monte Carlo; SMC) to infer parameters for these epidemiological models of arbitrary complexity and to reconstruct unobserved disease dynamics.

Results

Segregating site trajectories are informative of epidemiological dynamics.

Simulations of epidemiological models, as detailed in Materials and Methods, indicate that the number of segregating sites that are observed over time in a viral population are sensitive to sampling effort and are informative of epidemiological dynamics. To demonstrate this, we first simulated a susceptible-exposed-infected-recovered (SEIR) model under an epidemic scenario starting with a single infected individual (Figure 1A), further tracking the viral genotypes according to the approach outlined in Materials and Methods. The effect of sampling effort is shown in Figure 1B, which plots segregating site trajectories under dense sampling effort (40 sequences per 4-day time window) and under sparse sampling effort (20 sequences per 4-day time window). At both of these sampling efforts, the number of segregating sites first increases as the epidemic grows, as expected, with mutations accumulating in the virus population. Following the peak of the epidemic, the number of segregating sites starts to decline as viral lineages die out, reducing the amount of genetic variation present in the viral population. At lower sampling effort, less of the genetic variation present in the viral population over a given time window is likely to be sampled, resulting in a lower number of observed segregating sites during any time window.

To assess whether segregating site trajectories could be used for phylodynamic 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 Figure 1A, and a higher R_0 of 2.0. Differences in R_0 were implemented by differences in the transmission rate. The epidemic with the higher R_0 grew more rapidly (Figure

1C) and, under the same sampling effort, resulted in a more rapid increase in the number of segregating sites (Figure 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 (Lloyd-Smith et al. 2005). The extent of transmission heterogeneity is often gauged relative to the 20/80 rule (the most infectious 20% of infected individuals are responsible for 80% of the secondary cases (Woolhouse et al. 1997)), with some pathogens like SARS-CoV-2 exhibiting extreme levels of superspreading, with as low as 6-15% of infected individuals responsible for 80% of secondary cases (Althouse et al. 2020; Miller et al. 2020; Lemieux et al. 2021; Sun et al. 2021). Because transmission heterogeneity is known to impact patterns of viral genetic diversity (Koelle and Rasmussen 2012), we simulated the above SEIR model with transmission heterogeneity to ascertain its effects on segregating site trajectories. Transmission heterogeneity was implemented using a negative binomial distribution parameterized such that the most infectious 6% of infected individuals are responsible for 80% of the secondary cases (Materials and Methods). Because transmission heterogeneity has a negligible impact on epidemiological dynamics once the number of infected individuals is large (Keeling and Rohani 2008), these simulated epidemiological dynamics should be quantitatively similar to one another, with transmission heterogeneity simply expected to shorten the timing of epidemic onset in simulations with successful invasion (Lloyd-Smith et al. 2005). Our simulations confirm this pattern (Figure 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 (Figure 1E). Comparisons of segregating site trajectories between these simulations indicated that transmission heterogeneity substantially decreases the number of segregating sites during any time window (Figure 1F). These results indicate that the number of segregating sites in principle could be informative of the extent of transmission heterogeneity present in an unfolding epidemic. They also 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 (Danesh et al. 2020; Miller et al. 2020). 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 (Figure 1G). The segregating site trajectories for these three simulations indicate that reductions in R_0 over the course of an epidemic leave faint signatures in this low-dimensional summary statistic of viral diversity, with the signature being more pronounced with a more precipitous drop in R_0 (Figure 1H).

Phylodynamic inference using segregating site trajectories

To examine the extent to which phylodynamic inference based on segregating sites can be used for parameter estimation, we generated a mock segregating site trajectory by forward simulating an SEIR model with a R_0 of 1.6, sampling viral sequences from this simulation (Figure 2A), and calculating a segregating site trajectory from these sampled sequences (Figure 2B). Because the duration of the exposed period and the duration of the infectious period are generally known for viruses undergoing phylodynamic analysis, we fixed these parameters at their true values and first attempted to estimate only R_0 under the assumption that the timing of the index case t_0 is known. We estimated an R_0 value of 1.59 (95% confidence interval of 1.49 to 1.64; Materials and Methods; Figure 2C, 2C inset), demonstrating that phylodynamic inference using our segregating sites approach applied to this simulated dataset is able to recover the true R_0 value of 1.6.

Because the timing of the index case 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 Figure 2B. To do this, we first considered the parameter space ranging from an R_0 of 1.2 to 2.5 and from a t_0 of 60 days prior to the true start date of 0 to 56 days following this true start date. Considering R_0 intervals of 0.02 and t_0 intervals of 2 days, we ran 10 SMC simulations for every parameter combination. In Figure 3A, we plot the mean value of these 10 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 combinations and late t_0/high R_0 parameter combinations. However, this ridge falls off on both edges, indicating that the segregating sites approach can in principle estimate both t_0 and R_0. We therefore calculated profile likelihoods for both R_0 and t_0 (Figures 3B, 3C; Materials and Methods), arriving at an R_0 estimate of 1.50 (95% confidence = 1.34 to 1.67; Figure 3B) and a t_0 value of -13.8 (95% confidence = -27.8 to 0.3; Figure 3C) for the simulated dataset. While the maximum likelihood estimate for R_0 ran low and for t_0 ran early, the confidence intervals contained the true values of R_0 = 1.6 and t_0 = 0, respectively. Our results indicate that joint estimation of these parameters is thus possible. Using our estimates of R_0 and t_0, we reconstructed the dynamics of the segregating sites (Figure 4A) and unobserved state variables: the number of susceptible, exposed, and infected individuals over time (Figures 4B, C, D). These reconstructed state variables captured the true epidemiological dynamics, demonstrating that our segregating sites phylodynamic inference approach can be used to estimate epidemiological variables that generally go unobserved.

Phylodynamic inference for 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, when a country-wide lockdown was implemented. We decided to apply our approach to this set of sequences for several reasons. First, a large fraction of the 479 available full-genome sequences from France over this time period appear to be genetically very similar to one another (Gámbaro et al. 2020), indicating that one major lineage may have taken off in France (or at least, that most samples stemmed from one major lineage). This lineage would be the focus of our analysis. Second, an in-depth analysis previously inferred R_0 for France prior to the March 17 lock-down measures that were implemented (Salje et al. 2020). This analysis fit a compartmental infectious disease model to epidemiological data that included case, hospitalization, and death data. Because our phylodynamic inference approach can accommodate epidemiological model structures of arbitrary complexity, we can adopt the same model structure as in this previous analysis. We can also set the epidemiological parameters that are 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 can 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 n = 432 sequences without inferring a phylogeny (Materials and Methods). Using phylogenetic inference, however, we confirmed that our subset of sequences constituted a single evolutionary lineage (Figure S1). We calculated the nucleotide distance from each sequence in this subset to Wuhan/Hu-1 (Wu et al. 2020) (EPI_ISL_402125), a commonly used reference SARS-CoV-2 sequence that stemmed from a sample collected in Wuhan, China in late December 2019. Using these nucleotide distances, we estimated an evolutionary rate of 8.21 x 10^{-4} substitutions/site/yr (Figure 5A), consistent with the range of inferred evolutionary rate estimates for SARS-CoV-2 (Duchene et al. 2020; Pekar et al. 2020). This provides another confirmation that this subset of sequences is a single evolutionary lineage brought into France early on during the pandemic.

To generate a segregating site trajectory from these sequences, we established consecutive, non-overlapping 4-day time windows such that the last time window ended on March 17, 2020. Figure 5B shows the number of sequences falling into each time window. Figure 5C shows the segregating site trajectory calculated from these sequences. We jointly estimated R_0 and t_0 using this segregating site trajectory, under the assumption that the most infectious 15% of SARS-CoV-2 infected individuals are responsible for 80% of secondary infections, based on literature estimates of the extent of SARS-CoV-2 transmission heterogeneity (Sun et al. 2021) (Materials and Methods). We parameterized the model with a per genome, per transmission mutation rate of μ = 0.33 using consensus sequence data from established SARS-CoV-2 transmission pairs that were available in the literature (James et al. 2020; Popa et al. 2020; Braun et al. 2021; Lythgoe et al. 2021) (Materials and 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 (Figure 5D).

Using this approach, we estimated a μ value of 0.33 (95% confidence interval of 0.22 to 0.48), corresponding approximately to one mutation occurring every 3 transmission events.

Similar to the approach we undertook with our simulated data to jointly estimate R_0 and t_0, we first considered a broad parameter space over which to calculate log-likelihood values. Specifically, we considered R_0 values between 1.2 and 3.4 (at intervals of 0.1) and t_0 values between December 2, 2019 and February 16, 2020 (at intervals of 2 days). We ran 10 SMC simulations and calculated the mean log-likelihood for each parameter combination (Figure 6A). Similar to our findings on the simulated data set, we found evidence for a log-likelihood ridge between early t_0/low R_0 and late t_0/high R_0 parameter combinations. Profile log-likelihoods for R_0 and t_0 are shown in Figures 5B and 5C, respectively, yielding an estimate of R_0 = 2.22 (95% confidence interval = 1.5 to 2.94) and an estimate of t_0 = January 11 (95% confidence interval = December 26, 2019 to January 28, 2020). Our maximum likelihood estimate of R_0 is somewhat lower than the R_0 estimate arrived at through the epidemiological time series analysis that presented the epidemiological model structure we adopted (Salje et al. 2020). That analysis inferred an R_0 of 2.9 (95% confidence interval = 2.81 to 3.01) in France over this same time period. However, the confidence intervals of our analyses are relatively broad for R_0, and their estimate of R_0 = 2.9 falls within our 95% confidence interval. Our estimate is closer in line with estimates of the reproduction number in Wuhan prior to travel restrictions being introduced (R_0 = 2.35, with 95% CI of 1.15-4.77) (Kucharski et al. 2020) and with those estimated for Western European countries using incidence data up through March 17, 2020 (R_0 = 2.2, with 95% CI of 1.9-2.6) (Locatelli et al. 2021). Our estimate also aligns more closely with projections of R_0 made specifically for France, using outbreak data from Wuhan (Hilton and Keeling 2020): R_0 = 2.2 and R_0 = 2.7, under different assumptions related to age-dependent susceptibility and infectiousness. Finally, our R_0 estimates can be juxtaposed against results from phylodynamic analyses that used a birth-death model to infer R_0 during three distinct epochs in France using a similar set of sequence data we analyze here (Danesh et al. 2020). Their second epoch spanned February 19 through March 7, and the R_0 inferred for this time period was 2.56 (95% credible interval = 1.66 to 4.74). Our maximum likelihood estimate of t_0 in the middle of January 2020 aligns well with findings from Gámbaro et al. (2020) and is further consistent with the estimate from Salje et al.

(2020) that 58.65 (95% CI 37.85 – 88.37) individuals were present in the exposed (E_1 class) on January 22, 2020 based on fitting the epidemiological model to epidemiological data.

As we had done in our analysis of the simulated data set, we reconstructed the unobserved state variables using sampled particles from SMC simulations parameterized with R_0 and t_0 values that were sampled from the parameter space shown in Figure 6, weighted according to the log-likelihood values of the parameter combination. Plotting of reconstructed segregating site trajectories indicated a very good fit to the observed segregating site trajectory (Figure 7A). The number of individuals in the E_1, E_2, and I classes increased exponentially over the time period considered (Figure 7B), as expected for an epidemic with an $R_0 > 1$. In Figure 7C, we plot the reconstructed cumulative number of exposed individuals over time and the reconstructed cumulative number of recovered individuals over time. These cumulative dynamics indicate that by mid-March 0.004% to 0.069% of the population in France had become infected by this SARS-CoV-2 lineage and that 0.001% to 0.017% of the population in France had recovered from infection from this SARS-CoV-2 lineage. Depending on when seroconversion is assumed to occur, these cumulative predictions can be compared against findings from a serological study that was conducted over this time period in France (Le Vu et al. 2021). This study surveyed 3221 individuals, finding that 0.41% of individuals (95% confidence interval = 0.05 to 0.88) had gotten infected with SARS-CoV-2 by March 9 to 15, 2020. While these estimates fall slightly higher than our predictions, we are considering only one SARS-CoV-2 lineage (albeit likely the dominant one circulating during this time period), and would thus expect the cumulative positive proportion we predict to be lower than overall (all lineage) serology estimates. Other reasons for possible underestimation involve epidemiological model misspecification and inaccurate parameterization, for example, of the extent of transmission heterogeneity p_h.

Discussion

Here, we developed a phylodynamic 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 a “tree-free” approach for parameter estimation of emerging viral pathogens is that, early on in an epidemic or pandemic, phylogenetic uncertainty is significant, and tree-based phylodynamic inference approaches would need to integrate over this uncertainty, which is often times computationally intensive. A second benefit of using a "tree-free" approach is that parameters of the model of sequence evolution do not need to be estimated, reducing degrees of freedom considerably. Instead of viral phylogenies being the data that statistically interface with the epidemiological models, we use a low-dimensional summary statistic of the sequence data, namely the number of segregating sites present in temporally-adjacent sets of viral sequences. Beyond being a "tree-free" approach, our inference approach also benefits from being "plug-and-play" in that it can easily accommodate any arbitrarily complex (or simple) epidemiological model structure.

Based on fits to a simulated data set, we have shown that segregating site trajectories are highly informative of epidemiological parameters such as R_0 and the timing of the index case t_0. As far as we are aware, only one other peer-reviewed tree-free phylodynamic inference method exists (Kim et al. 2017), and future work should compare the approach developed here against this and potentially other phylodynamic inferences approaches.

Although there are clear benefits of the phylodynamic inference approach detailed here, it still relies on several assumptions that are also shared by other phylodynamic inference methods. Most notably, it relies on an assumption of random sampling of individuals. However, in contrast to coalescent-based models, the sampling rate does not have to be small relative to the number of infected individuals. Phylodynamic inference based on birth-death-sampling models instead requires the specification of a sampling process, such as a constant probability of an infected individual being sampled upon recovery/death (Stadler 2010). Misspecification of the sampling process can severely bias results, and much of the statistical power gained from these approaches appears to arise from the sequence of sample times rather than genealogical structure (Volz and Frost 2014). While our approach similarly requires an assumption of when individuals are sampled, our approach provides considerable flexibility in what assumptions are adopted, since the process model component of the state-space model can be easily implemented under any number of assumptions of when individuals are available for sampling. For example, in the compartmental model we used in the analysis of the France sequence data, we could in principle assume that individuals could be sampled once they became infected during a time window, rather than if they recovered during the time window.

The analysis we presented here focuses on phylodynamic 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 studies that have managed to incorporate additional data while performing phylodynamic inference have shown the value in pursing this goal (Rasmussen et al. 2011; Li et al. 2017). 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. While the process model would stay the same, another observation model can be added that relates the underlying state variables (e.g., S, E, I, R) to observed case data for instance. This proposed approach mirrors a previously described approach (Rasmussen et al. 2011), which showed that combining multiple data sources improved parameter estimation.

Our analysis focused on phylodynamic inference based on sequence data belonging to a single viral lineage, with a single index case. Our approach however can be expanded in a straightforward manner to multiple viral lineages, each with their own index case. This is especially useful in cases like SARS-CoV-2, where many regions have witnessed multiple clade introductions in fueling the start of more local epidemics (Gonzalez-Reiche et al. 2020; Miller et al. 2020). In this case, under the assumption that all lineages are phenotypically neutral and are expanding in subpopulations experiencing the same epidemiological parameters (e.g., R_0), the inference code can be expanded to estimate a single set of epidemiological parameters along with multiple index case times, one corresponding to each viral lineage. When considering multiple clades, a single segregating sites trajectory would be calculated for each clade, such that multiple segregating site trajectories could be fit to at the same time.

Our approach can also be extended in a straightforward manner to consider multiple clades that may be subject to different parameterizations for either intrinsic or extrinsic reasons. For example, clades circulating in the same region may expand at different rates due to genetic differences between the clades that confer a selective advantage of one clade over others. In this case, multiple segregating site trajectories could again be calculated – one for each clade – and phylodynamic inference would involve estimating epidemiological parameters, some of which may be assumed to be similar across clades, while others such as R_0 may differ between clades. As such, this inference method, which we initially developed for emerging pathogens with low levels of genetic diversity, may continue to be useful for endemic pathogens when questions involving emerging clades are a focus. Future work thus needs to determine when tree-free phylodynamic inference provides advantages over tree-based phylodynamic inference, and when tree-based methods provide better resolution into the dynamics of circulating virus populations.

Materials and Methods

Epidemiological model simulations and calculation of segregating site trajectories. We consider epidemiological models of arbitrary complexity that incorporate demographic stochasticity using Gillespie's τ-leap algorithm. As a concrete example of such an epidemiological model, we here use a susceptible-exposed-infected-recovered (SEIR) model whose dynamics are governed by the following equations:

$$S_{t+\Delta t} = S_t - N_{S\to E}$$

$$E_{t+\Delta t} = E_t + N_{S\to E} - N_{E\to I}$$

$$I_{t+\Delta t} = I_t + N_{E\to I} - N_{I\to R}$$

$$R_{t+\Delta t} = R_t + N_{I\to R}$$

where:

$$N_{S\to E} \sim Pois(\beta \frac{S_t}{N} I_t \Delta t)$$

$$N_{E\to I} \sim Pois(\gamma_E E_t \Delta t)$$

$$N_{I\to R} \sim Pois(\gamma_I I_t \Delta t)$$

Here, 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. While the epidemiological dynamics of this model can be simulated from the above equations alone, additional complexity is needed to incorporate virus evolution throughout the time period of the simulation. To incorporate virus evolution, we subcategorize both exposed individuals and infected individuals into genotype classes, with genotype 1 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 genotypes are numbered chronologically according to their appearance. When new mutations are generated at a transmission event, the new genotype is assumed to harbor the same mutation(s) as its infecting genotype plus any new mutations, which are similarly numbered chronologically based on appearance. We use a sparse matrix approach to store genotypes and their associated mutations to save on memory.

Given this model, during a time step Δt, $N_{E \to I}$ individuals are drawn at random from the set of individuals who are currently exposed; these will be the individuals who will transition to the infected class during this time step. Similarly, $N_{I \to R}$ individuals at drawn at random from the set of individuals who are currently infected; these will be the individuals who will transition to the recovered class during this time step. We further add $N_{S \to E}$ new individuals to the set of exposed class during time step Δt. 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 new individual enters the same genotype class of its parent. If one or more mutations occur during transmission, then this new individual enters a new genotype class, and the sparse matrix is extended to document the new genotype and its associated mutations.

We start the simulation with one infected individual carrying a viral genotype that we consider as the 'reference' genotype (genotype 1). 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 for any time window *i*, the number of segregating sites S_i. S_i is simply the number of polymorphic sites across the sampled individuals in time window *i*.

Phylodynamic inference using time series of segregating sites. Our phylodynamic inference approach relies on particle filtering, also known as Sequential Monte Carlo (SMC), to estimate model parameters and reconstruct latent state variables. The underlying forward model we use is formulated as a state-space model, with epidemiological variables (e.g., *S*, *E*, *I*, and *R*) being latent/unobserved variables in the process model. The model is simulated using Gillespie's τ-leap algorithm, as described in the section above. The evolutionary component of the model also contributes to the process model. For the observation model, we perform *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 observed in the empirical dataset in window *i*. We sample the same number of individuals as in the segregating sites dataset that the model interfaces with, since sampling effort 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 (and their associated mutations).

Between 'grabs', 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 density function parameterized with the mean S_i^{sim} value and evaluated at S_i. We use a Poisson probability density function based on our observation that a Poisson distribution with the mean number of segregating sites captures the distribution of S_i^{sim} values from the 'grabs' effectively (Figure S2). These probabilities serve as the weights for the particles. Particle weights are calculated at the end of each time window with $n_i > 0$. Particles are resampled at the end of each of these time windows according to their assigned weights. Particles with stochastic extinction of the virus prior to the end of the last time window with $n_i > 0$ have weights set to 0 in time window *i*. If the number of sampled individuals n_i in time window *i* exceeds the total number of individuals who recovered in time window *i*, the particle weight is similarly set to 0. We run 10 SMC simulations for each parameter set considered, resulting in 10 log-likelihood values.

For maximum likelihood estimation, weighted quadratic fitting is used, which is adapted from Ionides et al. (2017). First, we use local quadratic smoothing (*LOESS*) with a span of 0.75 to obtain the peak of the log-likelihood surface. The weight of each data point is determined by the distance between this peak, using the tri-cube weight function. After excluding data points with smaller weights by filtering out the smallest $\lambda \times 100$ percent, a quadratic function is fitted to data points based on weights. For Figure 2C, the λ for the quadratic fit was set to 0.5. For Figure 3B, the λ was set to 0.75, and for Figure 3C, the λ was set to 0.55. Latent state variables are reconstructed by randomly sampling a particle's $x_{0:tend}$ at the end of an SMC simulation, where t_{end} is the date at which the last sampled time window ends. All of our SMC simulations were performed with 200 particles. We used *k* = 100 'grabs' for the simulated data and, in the interest of time, *k* = 50 'grabs' for the France data.

Note that the complexity of this phylodynamic method is largely independent of the number of input sequences, in contrast to phylodynamic inference approaches that rely on integrating over phylogenetic uncertainty with BEAST.

Implementation of the transmission heterogeneity model. We implement transmission heterogeneity by subcompartmentalizing the infected classes into a high-transmission and a low-transmission class, as has been done elsewhere (Volz and Siveroni 2018; Miller et al. 2020). For an SEIR model, the model extended to incorporate transmission heterogeneity becomes:

$$S_{t+\Delta t} = S_t - N_{S \to E}$$

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

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

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

$$R_{t+\Delta t} = R_t + N_{I_h\to R} + N_{I_l\to R}$$

where:

$$N_{S\to E} \sim Pois(\beta_h \frac{S_t}{N} I_{h,t}\Delta t) + Pois(\beta_l \frac{S_t}{N} I_{l,t}\Delta t)$$

$$N_{E\to I} \sim Pois(\gamma_E E_t \Delta t)$$

$$N_{E\to I_h} \sim Bin(N_{E\to I}, p_h)$$

$$N_{E\to I_l} = N_{E\to I} - N_{E\to I_h}$$

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

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

The parameter p_h quantifies the proportion of exposed individuals who transition to the highly infectious 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 (Volz and Siveroni 2018; Miller et al. 2020), 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 resulted 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 (Miller et al. 2020), to $\frac{\left[\frac{1-p_h}{p_h}\right]}{\left[\frac{1}{P}-1\right]}$. With c defined in this way, p_h is interpreted as the proportion of most infectious individuals that result in P = 80% of secondary infections. 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 $\beta_h = c\beta_l$.

Epidemiological model structure and parameterization used for the France analysis.

The process model we use in our phylodynamic inference of the France sequence data is based on a previously published epidemiological model for SARS-COV-2 in France (Salje et al. 2020). 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 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, implemented using Gillespie's τ-leap algorithm, is given by:

$$S_{t+\Delta t} = S_t - N_{S\rightarrow E1}$$

$$E_{1,t+\Delta t} = E_{1,t} + N_{S\rightarrow E1} - N_{E1\rightarrow E2}$$

$$E_{2,t+\Delta t} = E_{2,t} + N_{E1\rightarrow E2} - N_{E2\rightarrow I}$$

$$I_{t+\Delta t} = I_t + N_{E2\rightarrow I} - N_{I\rightarrow R}$$

$$R_{t+\Delta t} = R_t + N_{I\rightarrow R}$$

where:

$$N_{S\rightarrow E1} \sim Pois(\beta \frac{S_t}{N} I_t \Delta t) + Pois(\beta \frac{S_t}{N} E_{2,t} \Delta t)$$

$$N_{E1\rightarrow E2} \sim Pois(\gamma_{E1} E_{1,t} \Delta t)$$

$$N_{E2\rightarrow I} \sim Pois(\gamma_{E2} E_{2,t} \Delta t)$$

$$N_{I\rightarrow\ R} \sim Pois(\gamma_I I_t \Delta t)$$

with β being the transmission rate, 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. While exposed class 2 (E_2) and infected class I both transmit as efficiently, their model contains this level of detail to more effectively interface with the case data, where symptoms do not appear before an individual is infected (in class I). We keep with this model, rather than reducing it to having only a single exposed class and a single infectious class to keep the same distribution of infected times as in their model.

Because SARS-CoV-2 dynamics are characterized by substantial levels of transmission heterogeneity (Adam et al. 2020; Miller et al. 2020; Sun et al. 2021) and we have shown in Figure 1 that transmission heterogeneity impacts segregating site trajectories, we expanded the compartmental epidemiological model described above to include transmission heterogeneity in a manner similar to the one we used in Figures 1E, F. Based specifically on the analysis by Sun and coauthors (Sun et al. 2021), we set p_h to 0.15, such that 15% of infections are responsible for 80% of secondary infections.

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 between 87 transmission pairs of SARS-CoV-2 from four studies (James et al. 2020; Popa et al. 2020; Braun et al. 2021; Lythgoe et al. 2021). 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 (Lythgoe et al. 2021) Sequence data were aligned to Wuhan/Hu-1 (MN908947.3) (Wu et al. 2020) using MAFFT v.7.464 (Katoh 2002). 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 (Harris et al. 2020). Ambiguous nucleotides were considered in the identification of *de novo* substitutions (i.e. 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 λ parameter of the Poisson distribution. The 95% confidence intervals were calculated using the exact method using SciPy v.1.5.4 (SciPy 1.0 Contributors et al. 2020) such that the lower bound was $\frac{(\mathrm{X}^2_{2Y,0.025})/2}{87}$ and the upper bound was $\frac{(\mathrm{X}^2_{2(Y+1),0.975})/2}{87}$ where Y is the total number of observed substitutions.

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 (Duchene et al. 2020; Pekar et al. 2020). With a genome length of SARS-CoV-2 of approximately 30,000 nucleotides and a generation interval of approximately 4.5 days (Griffin et al. 2020), 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://www.france24.com/en/20200316-live-france-s-macron-addresses-nation-amid-worsening-coronavirus-outbreak) in France and uploaded through April 29th, 2021 from GISAID (Shu and McCauley 2017). 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 122 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 is similar to the set of sequences analyzed in Danesh et al. (2020). Sequences were binned into four-day windows, aligned such that the last window ended on the latest sampling date, and the number of segregating sites in each window calculated in Python using NumPy. Ambiguous nucleotides were considered in the calculation of segregating 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, collected date complete 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 (Minh et al. 2020) was used to construct a maximum likelihood phylogeny, and ModelFinder (Kalyaanamoorthy et al. 2017) was used to find the best fit nucleotide substitution model (GTR+F+I). Small branches were collapsed. TreeTime v. 0.8.0 (Sagulenko et al. 2017) 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} with a standard deviation of 5×10^{-4}, 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 (Hunter 2007) and Baltic (https://github.com/evogytis/baltic).

Availability of code.

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

FIGURES

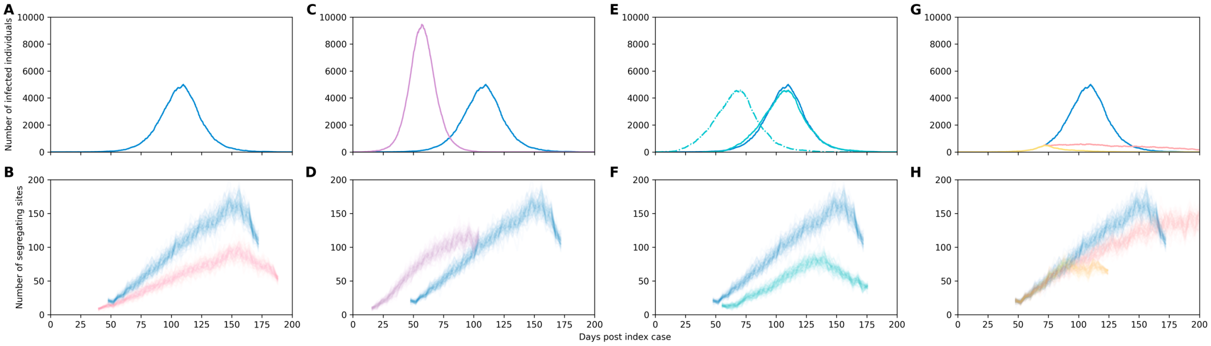


Figure 1. Segregating site trajectories under simulated epidemiological dynamics. (A) Simulated dynamics of infected individuals (*I*) under an SEIR model simulated with an R_0 of 1.6. (B) Segregating site trajectories under dense and sparse sampling. Dense sampling (blue lines) corresponds to 40 sequences sampled per time window. Sparse sampling (red lines) corresponds to 20 sequences sampled per time window. (C) Simulated dynamics of infected individuals (*I*) under an SEIR model simulated with an R_0 of 2.0 (purple line) compared to those of the R_0 = 1.6 simulation (blue line). A higher transmission rate was used to generate the higher R_0 value of 2.0. (D) Segregating site trajectories for the R_0 = 2.0 simulation (purple lines) and the R_0 = 1.6 simulation (blue lines). Both simulations are densely sampled (40 sequences sampled per time window). (E) Simulated dynamics of infected individuals (*I*) under an SEIR model with an R_0 of 1.6 and incorporating transmission heterogeneity (teal, dashed line) compared to those of the original R_0 = 1.6 simulation (blue line) without transmission heterogeneity. Transmission heterogeneity was included by setting p_h = 0.06, resulting in 6% of the most infectious individuals being responsible for 80% of secondary infections. For ease of comparing segregating site trajectories, the transmission heterogeneity simulation was shifted later in time such that its epidemic peak aligned with the simulation without transmission heterogeneity (teal, solid line). (F) Segregating site trajectories for the shifted transmission heterogeneity simulation (teal lines) and the simulation without transmission heterogeneity (blue line). Both simulations are densely sampled (40 sequences sampled per time window). (G) Simulated dynamics of infected individuals (*I*) under an SEIR model simulated with changing R_0. Changes in R_0 occurred when the number of infected individuals reached 400. 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 blue has R_0 remaining at 1.6. (H) Segregating site trajectories for the three simulations shown in Figure 1G. All three simulations are

densely sampled (40 sequences sampled per time window). 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 μ = 0.2. Initial conditions are $S(t_0)$ = N-1, $E(t_0)$ = 0, $I(t_0)$ = 1, and $R(t_0)$ = 0. For the transmission heterogeneity simulation (subplot E), initial conditions are $S(t_0)$ = N-1, $E(t_0)$ = 0, $I_\mathrm{h}(t_0)$ = 1, $I_\mathrm{l}(t_0)$ = 0, and $R(t_0)$ = 0. A time step of τ = 0.1 days was used in the Gillespie τ-leap algorithm. Time windows of T = 4 days were used to bin sequences for the segregating sites calculation. 100 different segregating site trajectories are shown for each simulation.

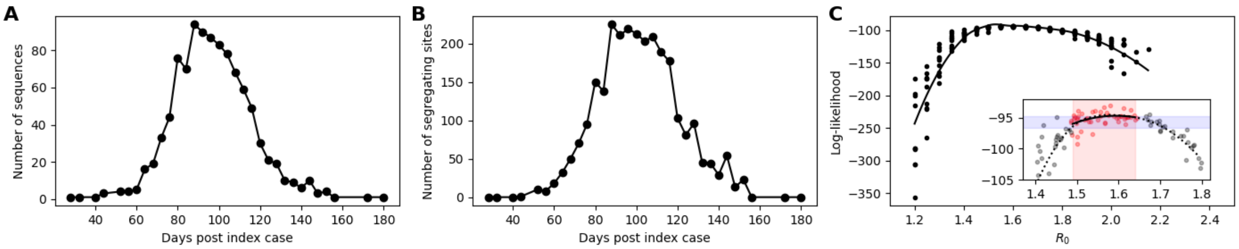


Figure 2. Phylodynamic inference on a simulated trajectory of segregating sites. (A) The number of sampled sequences over time, by time window. Sampling was done in proportion to the number of individuals recovering in a time window. In all, 1000 sequences were sampled over the course of the simulated epidemic. The number of samples in a given time window was constrained to be ≤100. (B) Simulated segregating site trajectory from the sampled sequences. (C) Estimation of R_0 using SMC. Points show log-likelihood values from different SMC simulations across a range of R_0 values between 1.2 and 2.5, in 0.05 increments. Smoothed likelihood surface was obtained by *LOESS* smoothing with a span of 0.75. Inset: Maximum likelihood estimation of R_0 using quadratic fitting. Black points in inset show log-likelihood values from different SMC simulations across a range of R_0 values between 1.4 and 1.8. The vertical black dashed line shows the maximum likelihood estimate (MLE) of R_0 (1.59). The red band shows the 95% confidence interval of R_0 (1.49 – 1.64). MLE and 95% CI were obtained from fitting a quadratic function to the log-likelihood values shown in the inset, using a similar approach to the one outlined in Ionides et al. (2017) with a λ value of 0.5. 95% CI were set at the values of R_0 corresponding to the maximum likelihood value at the peak of the quadratic curve minus 1.92 log-likelihood units. Model parameters for the simulated data set are: R_0= 1.6, γ_E= 1/2 days^{-1}, γ_I = 1/3 days^{-1}, population size N = 10^5, t_0 = 0, and the per genome, per transmission mutation rate μ = 0.2. Initial conditions are $S(t_0)$ = N-1, $E(t_0)$ = 0, $I(t_0)$ = 1, and $R(t_0)$ = 0. A time step of τ = 0.1 days was used in the Gillespie τ-leap algorithm. A time window of T = 4 days was used to bin sequences for the segregating sites calculation.

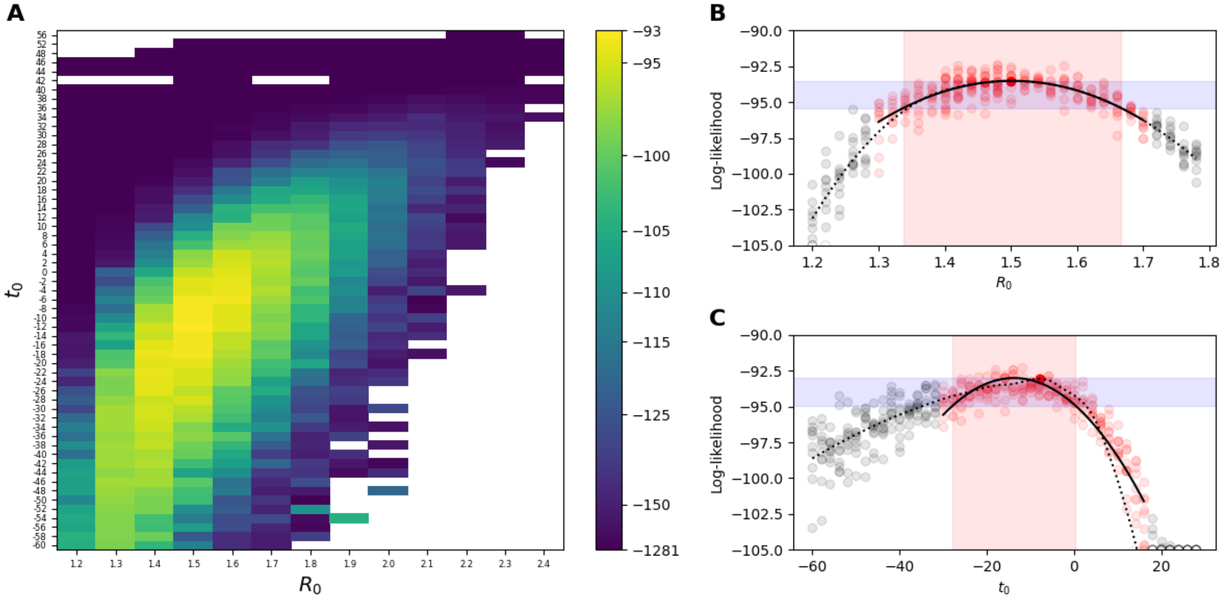


Figure 3. Joint estimation of the basic reproduction number (R_0) and the timing of the index case (t_0) using simulated data. (A) The likelihood surface based on the segregating site trajectory shown in Figure 2B is shown over a broad range of R_0 values (1.2 to 2.4, in 0.1 increments) and t_0 values (from 60 days prior to 56 days following the true t_0 of 0 in 2-day increments). Blank cells yielded log-likelihood values of <-1281. Log-likelihood values shown in each cell across this broad range of R_0 and t_0 are mean log-likelihood values calculated from 10 SMC simulations at each parameterization. (B-C) Profile likelihood for R_0 (B) and t_0 (C). Profile likelihoods were calculated using an approach similar to the one outlined in Ionides et al. (2017). The *LOESS* fit is shown with a dotted black line. The quadratic fit is shown with a solid black line. Points included in the quadratic fit are shown in red; points excluded from the quadratic fit are shown in gray. The shaded red area is the 95% confidence interval for the focal parameter. The shaded blue area shows the range of log-likelihood values that fall within 1.92 log-likelihood values of the quadratic fit's maximum value.

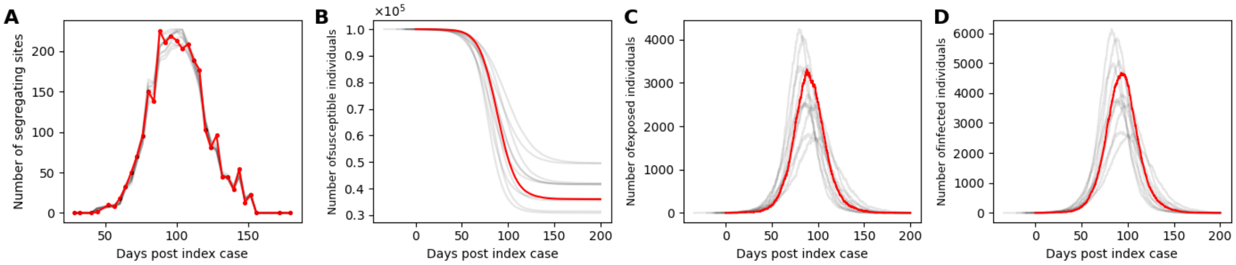


Figure 4. Trajectories of reconstructed unobserved state variables for the simulated dataset. (A) Simulated trajectory of the number of segregating sites (red), alongside reconstructed trajectory of the number of segregating sites from 10 sampled SMC particles (gray). For each SMC particle, a combination of t_0 and R_0 values of 10 SMC iterations were randomly chosen based on their log-likelihood values. (B) Simulated dynamics of susceptible individuals (red), alongside reconstructed dynamics of susceptible individuals from these SMC simulations (gray). (E) Simulated dynamics of exposed individuals (red), alongside reconstructed dynamics of exposed individuals (gray). (F) Simulated dynamics of infected individuals (red), alongside reconstructed dynamics of infected individuals (gray).

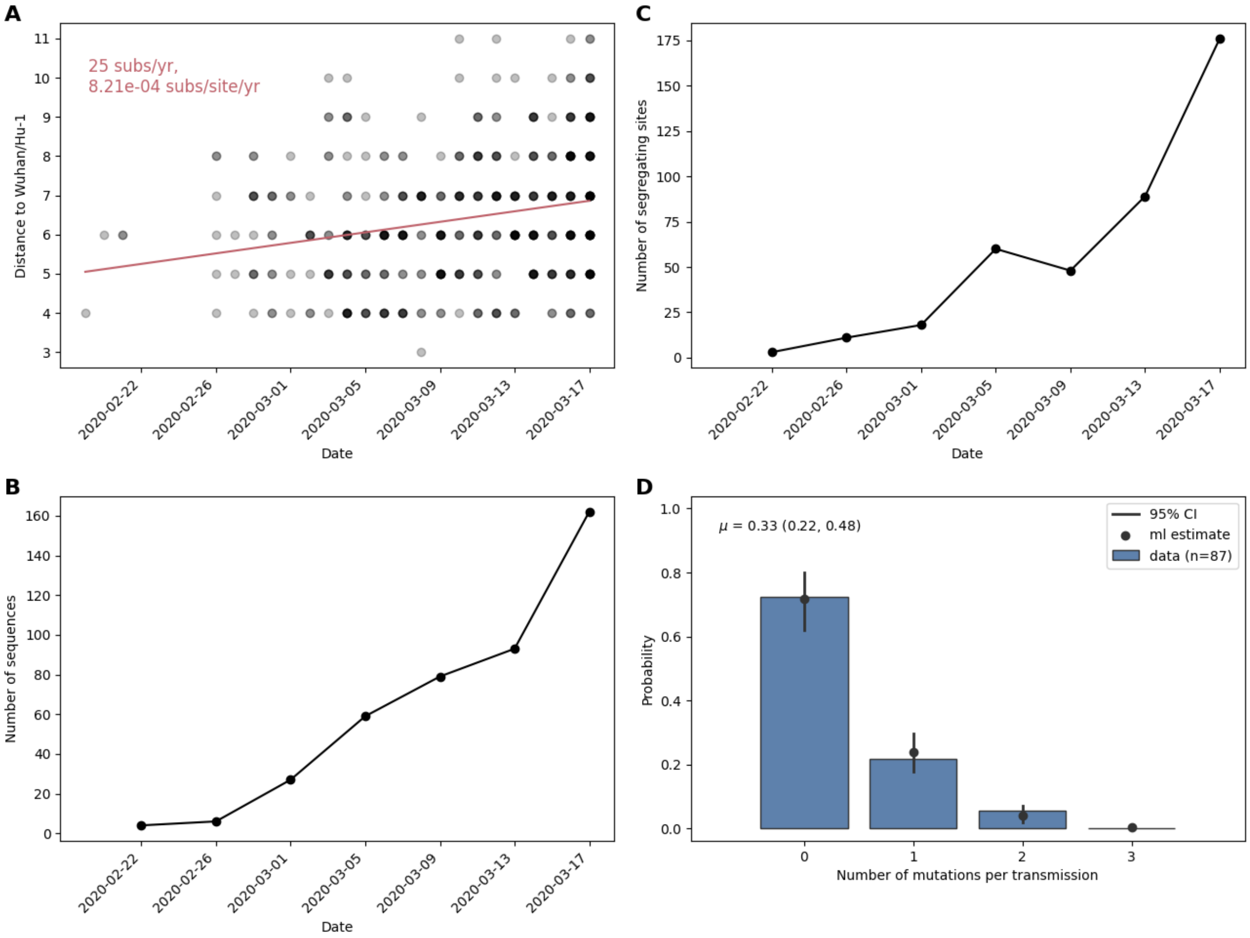


Figure 5. Sequences and parameters used in the estimation of R_0 and t_0 for the France data. (A) Sequences used in the phylodynamic analysis, plotted by their collection date and their nucleotide divergence from the Wuhan/Hu-1 reference sequence. (B) The number of sampled sequences over time, calculated using a T = 4 day time window. (C) The segregating site trajectory calculated from the sampled sequences, using the same T = 4 day time window shown in (B). (D) Estimation of the per genome, per transmission mutation rate μ. Blue histogram plots the fraction of transmission pairs with consensus sequences that differed from one another by the number of mutations shown on the x-axis. The Poisson estimate from these data, shown in black, was μ = 0.33 (95% CI = 0.22-0.48).

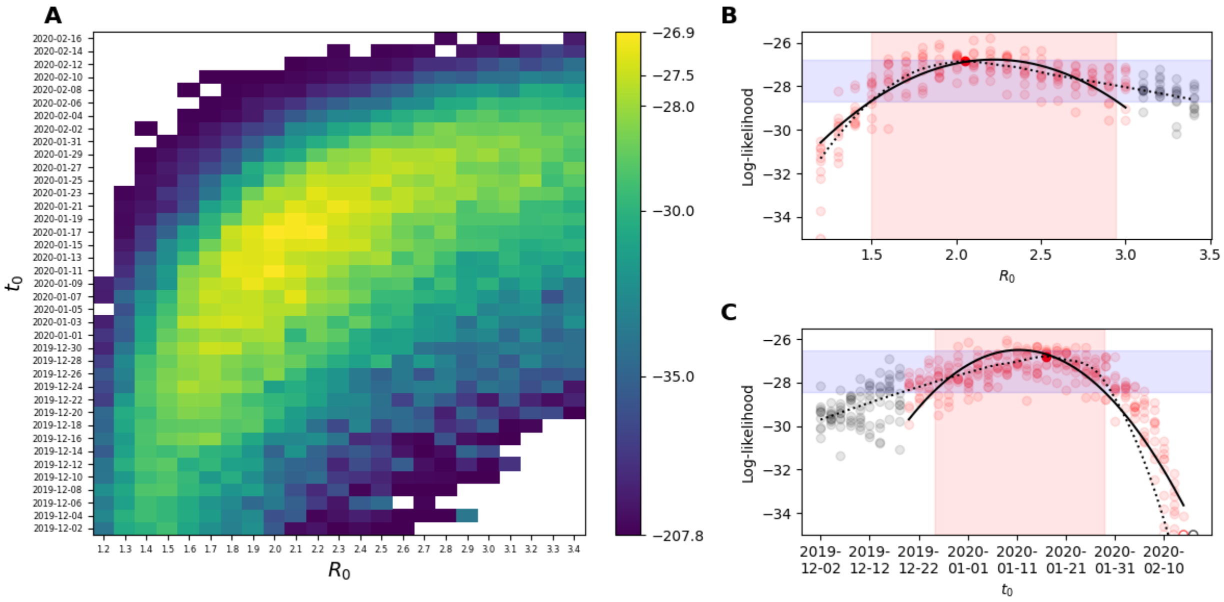


Figure 6. Joint estimation of the basic reproduction number R_0 and the timing of the index case t_0 using the France data. (A) The joint log-likelihood surface based on the estimated segregating site trajectory for the France data. Each cell is colored according to the mean of log-likelihood for a t_0, R_0 combination obtained from 10 SMC simulations. (B-C) Profile likelihood for R_0 (B) and t_0 (C). Profile likelihoods were calculated using an approach similar to the one outlined in Ionides et al. (2017). The *LOESS* fit is shown with a dotted black line. The quadratic fit is shown with a solid black line. Points included in the quadratic fit are shown in red; points excluded from the quadratic fit are shown in black. The shaded red and blue areas are, as in Figures 3B and 3C, the 95% confidence interval for the focal parameter and the range of log-likelihood values that fall within 1.92 log-likelihood values of the quadratic fit's maximum value.

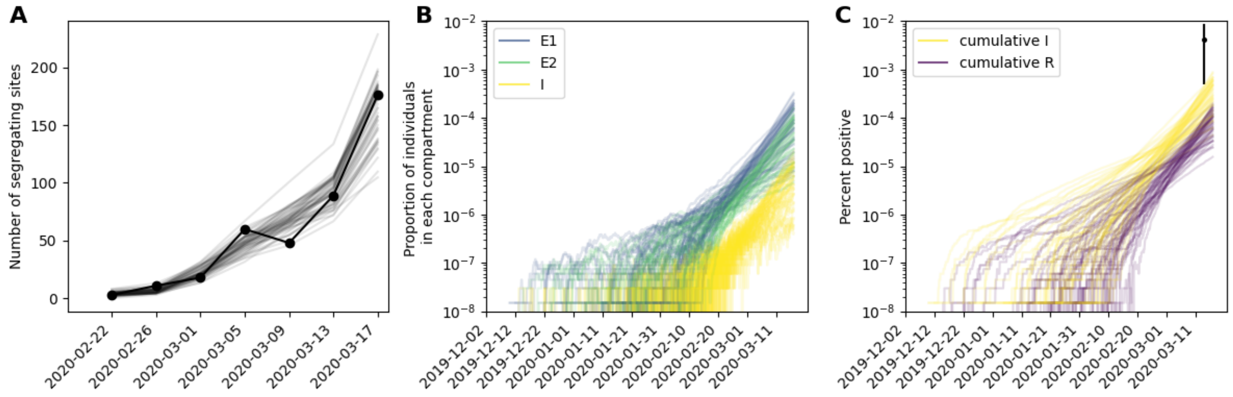
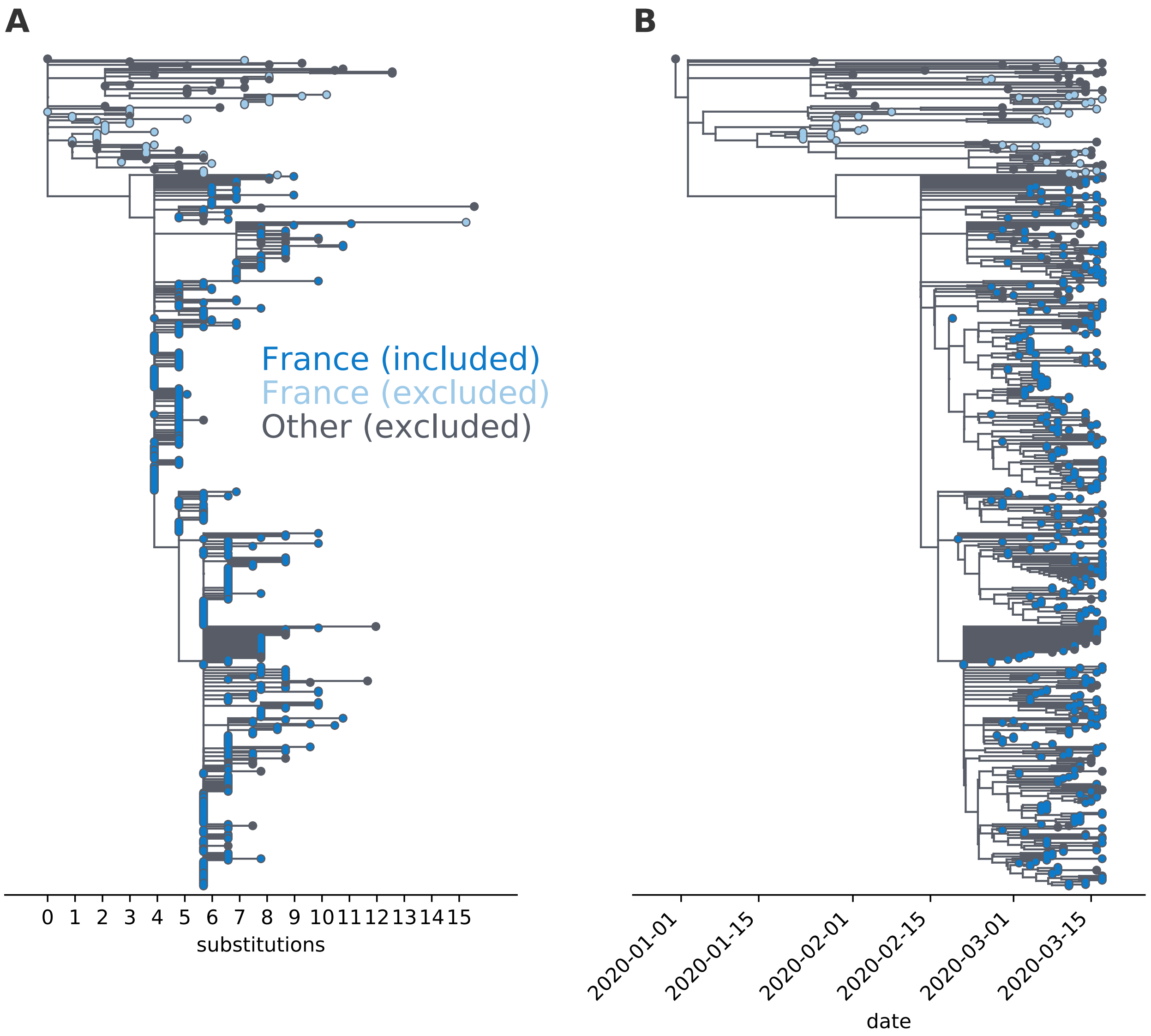


Figure 7. Trajectories of reconstructed unobserved state variable for the France data. For the reconstruction of all state variables shown, a combination of two parameters, R_0 and t_0, are sampled based on their log-likelihood values from 10 SMC simulations. (A) Segregating site trajectory for the France data, alongside segregating site trajectories from 10 sampled SMC particles. (B) Reconstructed dynamics of the number of individuals in the first exposed class (E_1), the second exposed class (E_2), and the infected class (I). (C) Cumulative number of exposed individuals (yellow) and cumulative number of recovered individuals (purple) over time. The maximum likelihood estimate of the fraction of the population that had been infected with SARS-CoV-2 by mid-March, and the 95% confidence interval of this estimate, are shown in black. Estimates are from a serological study conducted during the time window March 9-15, 2020 (Le Vu et al. 2021).



683

684 **Figure S1.** Inferred phylogenies for the sequences sampled from France, January 23-March 17, 2020. (A)
685 Divergence tree, showing the number of nucleotide substitutions from Wuhan/Hu-1. Sequences from
686 France are colored in blue, with dark blue coloring indicating sequences that were included in our single-
687 lineage analysis and light blue coloring indicating sequences that were excluded from our analysis. Tips
688 colored in gray denote genetically similar sequences sampled from outside of France during this time
689 period. (B) Time-aligned maximum likelihood phylogeny, with coloring of sequences as in (A).

690

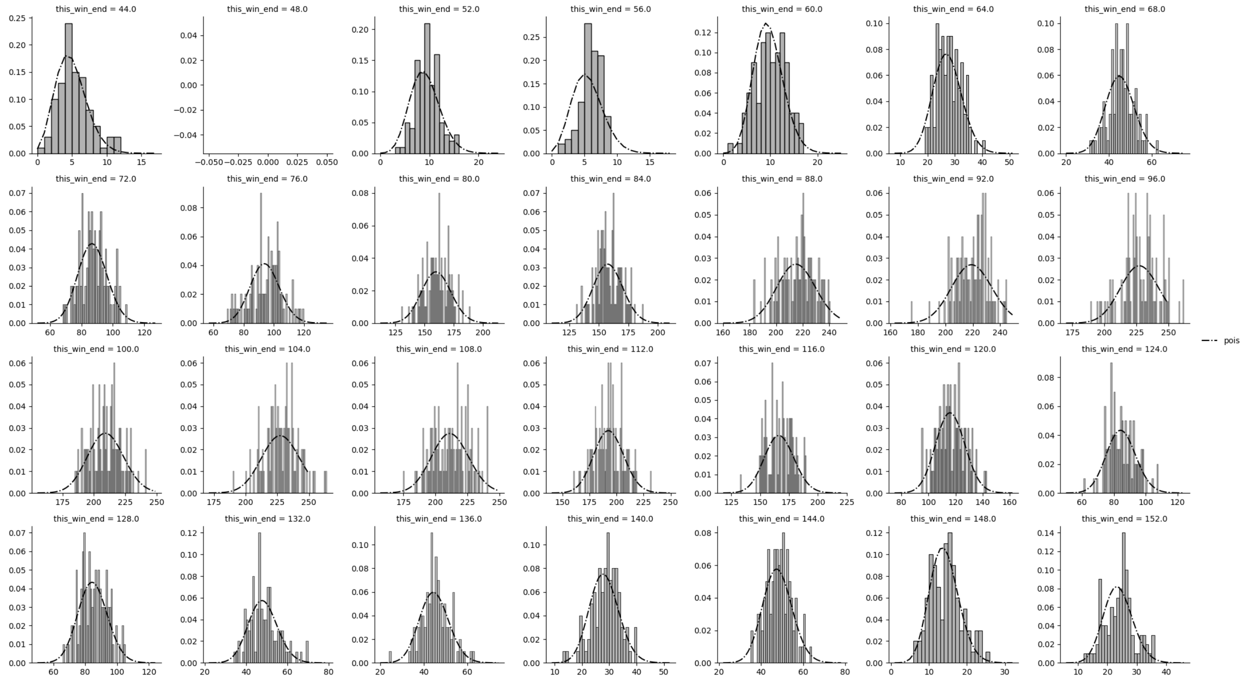


Figure S2. Appropriateness of the Poisson distribution in the observation model. Each subplot shows a time window *i*, with the blue vertical line indicating the observed value in that time window, S_i. Each time window further shows a histogram of S_i^{sim} values from 100 'grabs' from a single randomly sampled particle. The dash-dotted black curves show Poisson probability mass functions, parameterized with the average of the S_i^{sim} values.

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.

References

Adam DC, Wu P, Wong JY, Lau EHY, Tsang TK, Cauchemez S, Leung GM, Cowling BJ. 2020. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat Med* 26:1714–1719.

Althouse BM, Wenger EA, Miller JC, Scarpino SV, Allard A, Hébert-Dufresne L, Hu H. 2020. Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLoS Biol* 18:e3000897.

Boskova V, Bonhoeffer S, Stadler T. 2014. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models.Koelle K, editor. *PLoS Computational Biology* 10:e1003913.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis.Prlic A, editor. *PLoS Computational Biology* 10:e1003537.

Braun K, Moreno G, Wagner C, Accola MA, Rehrauer WM, Baker D, Koelle K, O'Connor DH, Bedford T, Friedrich TC, et al. 2021. Limited within-host diversity and tight transmission bottlenecks limit SARS-CoV-2 evolution in acutely infected individuals. Evolutionary Biology Available from: http://biorxiv.org/lookup/doi/10.1101/2021.04.30.440988

Danesh G, Elie B, Michalakis Y, Sofonea MT, Bal A, Behillil S, Destras G, Boutolleau D, Burrel S, Marcelin A-G, et al. 2020. Early phylodynamics analysis of the COVID-19 epidemic in France. Epidemiology Available from: http://medrxiv.org/lookup/doi/10.1101/2020.06.03.20119925

Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. 2020. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution* 6:veaa061.

Gámbaro F, Behillil S, Baidaliuk A, Donati F, Albert M, Alexandru A, Vanpeene M, Bizard M, Brisebarre A, Barbet M, et al. 2020. Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020. *Eurosurveillance* [Internet] 25. Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.26.2001200

Geidelberg L, Boyd O, Jorgensen D, Siveroni I, Nascimento FF, Johnson R, Ragonnet-Cronin M, Fu H, Wang H, Xi X, et al. 2021. Genomic epidemiology of a densely sampled COVID-19 outbreak in China. *Virus Evolution* 7:veaa102.

Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre S, Kleiner G, Polanco J, Khan Z, et al. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science*:eabc1917.

Griffin JM, Collins AB, Hunt K, McEvoy D, Casey M, Byrne AW, McAloon CG, Barber A, Lane EA, More SJ. 2020. A rapid review of available evidence on the serial interval and generation time of COVID-19. Epidemiology Available from: http://medrxiv.org/lookup/doi/10.1101/2020.05.08.20095075

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature* 585:357–362.

Hilton J, Keeling MJ. 2020. Estimation of country-level basic reproductive ratios for novel Coronavirus (SARS-CoV-2/COVID-19) using synthetic contact matrices. *PLoS Comput Biol* 16:e1008031.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9:90–95.

Ionides EL, Breto C, Park J, Smith RA, King AA. 2017. Monte Carlo profile confidence intervals for dynamic systems. *Journal of The Royal Society Interface* 14:20170126.

James SE, Ngcapu S, Kanzi AM, Tegally H, Fonseca V, Giandhari J, Wilkinson E, Chimukangara B, Pillay S, Singh L, et al. 2020. High Resolution analysis of Transmission Dynamics of Sars-Cov-2 in Two Major Hospital Outbreaks in South Africa Leveraging Intrahost Diversity. Infectious Diseases (except HIV/AIDS) Available from: http://medrxiv.org/lookup/doi/10.1101/2020.11.15.20231993

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589.

Katoh K. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30:3059–3066.

Keeling MJ, Rohani P. 2008. Modeling Infectious Diseases in Humans and Animals. Princeton University Press

Kim K, Omori R, Ito K. 2017. Inferring epidemiological dynamics of infectious diseases using Tajima's D statistic on nucleotide sequences of pathogens. *Epidemics* 21:21–29.

Koelle K, Rasmussen DA. 2012. Rates of coalescence for common epidemiological models at equilibrium. *J. R. Soc. Interface* 9:997–1007.

Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Sun F, Jit M, Munday JD, et al. 2020. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases* 20:553–558.

Le Vu S, Jones G, Anna F, Rose T, Richard J-B, Bernard-Stoecklin S, Goyard S, Demeret C, Helynck O, Escriou N, et al. 2021. Prevalence of SARS-CoV-2 antibodies in France: results from nationwide serological surveillance. *Nat Commun* 12:3025.

Lemieux JE, Siddle KJ, Shaw BM, Loreth C, Schaffner SF, Gladden-Young A, Adams G, Fink T, Tomkins-Tinch CH, Krasilnikova LA, et al. 2021. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* 371.

Li LM, Grassly NC, Fraser C. 2017. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Molecular Biology and Evolution* 34:2982–2995.

Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005. Superspreading and the effect of individual variation on disease emergence. *Nature* 438:355–359.

Locatelli I, Trächsel B, Rousson V. 2021. Estimating the basic reproduction number for COVID-19 in Western Europe.Khudyakov YE, editor. *PLoS ONE* 16:e0248731.

Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Andersson M, Otecko N, Wise EL, Moore N, et al. 2021. SARS-CoV-2 within-host diversity and transmission. *Science* 372:eabg0821.

Miller D, Martin MA, Harel N, Tirosh O, Kustin T, Meir M, Sorek N, Gefen-Halevi S, Amit S, Vorontsov O, et al. 2020. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat Commun* 11:5518.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.Teeling E, editor. *Molecular Biology and Evolution* 37:1530–1534.

Pekar J, Worobey M, Moshiri N, Scheffler K, Wertheim JO. 2020. Timing the SARS-CoV-2 Index Case in Hubei Province. Evolutionary Biology Available from: http://biorxiv.org/lookup/doi/10.1101/2020.11.20.392126

Popa A, Genger J-W, Nicholson MD, Penz T, Schmid D, Aberle SW, Agerer B, Lercher A, Endler L, Colaço H, et al. 2020. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci. Transl. Med.* 12:eabe2555.

Popinga A, Vaughan T, Stadler T, Drummond AJ. 2014. Inferring epidemiological dynamics with Bayesian coalescent inference: the merits of deterministic and stochastic models. *Genetics*.

Rasmussen DA, Boni MF, Koelle K. 2014. Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. *Molecular Biology and Evolution* 31:258–271.

Rasmussen DA, Ratmann O, Koelle K. 2011. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol* 7:e1002136.

Rasmussen DA, Stadler T. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. *Evolutionary Biology*:24.

Ratmann O, Hodcroft EB, Pickles M, Cori A, Hall M, Lycett S, Colijn C, Dearlove B, Didelot X, Frost S, et al. 2017. Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. *Molecular Biology and Evolution* 34:185–203.

Sagulenko P, Puller V, Neher R. 2017. TreeTime: maximum likelihood phylodynamic analysis.

Salje H, Tran Kiem C, Lefrancq N, Courtejoie N, Bosetti P, Paireau J, Andronico A, Hozé N, Richet J, Dubost C-L, et al. 2020. Estimating the burden of SARS-CoV-2 in France. *Science* 369:208–211.

SciPy 1.0 Contributors, Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17:261–272.

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* [Internet] 22. Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2017.22.13.30494

Stadler T. 2010. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology* 267:396–404.

Stadler T, Bonhoeffer S. 2013. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. R. Soc. B* [Internet] 368. Available from: http://rstb.royalsocietypublishing.org/content/368/1614/20120198

Stadler T, Kouyos R, Wyl V von, Yerly S, Böni J, Bürgisser P, Klimkait T, Joos B, Rieder P, Xie D, et al. 2012. Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol* 29:347–357.

Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences* 110:228–233.

Stadler T, Kühnert D, Rasmussen DA, du Plessis L. 2014. Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data. *PLoS Curr* [Internet]. Available from: https://currents.plos.org/outbreaks/article/insights-into-the-early-epidemic-spread-of-ebola-in-sierra-leone-provided-by-viral-sequence-data/

Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, Zhan Z, Chen X, Zhao S, Huang Y, et al. 2021. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science* 371:eabe2424.

Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T. 2017. Directly Estimating Epidemic Curves From Genomic Data. Available from: http://biorxiv.org/lookup/doi/10.1101/142570

Volz EM. 2012. Complex population dynamics and the coalescent under neutrality. *Genetics* 190:187–201.

Volz EM, Frost SDW. 2014. Sampling through time and phylodynamic inference with coalescent and birth-death models. *Journal of The Royal Society Interface* 11:20140945–20140945.

Volz EM, Ndembi N, Nowak R, Kijak GH, Idoko J, Dakum P, Royal W, Baral S, Dybul M, Blattner WA, et al. 2017. Phylodynamic analysis to inform prevention efforts in mixed HIV epidemics. *Virus Evolution* [Internet] 3. Available from: https://academic.oup.com/ve/article-lookup/doi/10.1093/ve/vex014

Volz EM, Siveroni I. 2018. Bayesian phylodynamic inference with complex models.Darling AE, editor. *PLoS Comput Biol* 14:e1006546.

Woolhouse MEJ, Dye C, Etard J-F, Smith T, Charlwood JD, Garnett GP, Hagan P, Hii JLK, Ndhlovu PD, Quinnell RJ, et al. 1997. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proceedings of the National Academy of Sciences* 94:338–342.

Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579:265–269.