Epidemiological 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 Center of Excellence for Influenza Research and Response (CEIRR), Atlanta GA, USA

*katia.koelle@emory.edu

Key words: statistical inference; segregating sites; viral sequence data; infectious disease modeling; phylodynamics; SARS-CoV-2

Abstract

Epidemiological models are commonly fit to case data to estimate model parameters and to infer unobserved disease dynamics. Epidemiological models have also been fit to viral sequence data using phylodynamic inference approaches that 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. Moreover, these approaches require estimation of parameters associated with models of sequence evolution in addition to the estimation of epidemiological parameters that are of primary interest. Here, we present an alternative approach to phylodynamic inference that can be used early on during the expansion of a viral lineage 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 interfacing epidemiological models with sequence data. We then show that our proposed approach accurately recovers key epidemiological quantities such as the basic reproduction number and the timing of the index case from mock data simulated under a single-introduction scenario. Finally, we apply our approach to SARS-CoV-2 sequence data from France, estimating a reproductive number of approximately 2.5 to 2.7 under an epidemiological model structure that allows for multiple introductions, consistent with estimates from epidemiological surveillance data. Our findings indicate that "tree-free" statistical 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 early on in an epidemic or during the expansion of a viral lineage.

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 structured epidemiological models, for example, susceptible-exposed-infected-recovered (SEIR) models and models with spatial subdivision (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, 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 (Boskova et al. 2014). Both of these types of phylodynamic inference approaches rely on time-resolved phylogenies and have been incorporated into the phylogenetics software package BEAST2 (Bouckaert et al. 2014) to allow for 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 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 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.

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 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, we develop a statistical inference approach that fits compartmental epidemiological models to times 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. In Materials and Methods, we

84 provide details on the simulation of epidemiological models that incorporate viral evolution and
85 thus can yield simulated time series of the number of segregating sites. In that section, we
86 further describe our statistical inference approach that relies on using particle filtering
87 (otherwise known as Sequential Monte Carlo; SMC) to infer parameters for these
88 epidemiological models and to reconstruct unobserved disease dynamics. Because we propose
89 the use of our method early on in an epidemic (or during the early expansion of a viral lineage),
90 we focus primarily on estimating the basic reproduction number R_0 using this inference
91 approach.

92 Results

93 Segregating site trajectories are informative of epidemiological dynamics.

94 The number of segregating sites present in a set of sampled viral sequences is defined as the
95 number of nucleotide sites at which genetic variation is present in the sample set. To determine
96 whether the number of segregating sites that are observed over time in a viral population may
97 be informative of underlying epidemiological dynamics, we forward-simulated a classic
98 susceptible-exposed-infected-recovered (SEIR) epidemiological model, augmented with viral
99 evolution, under various sampling efforts and parameterizations (Figure 1; Materials and
100 Methods). Simulations of this augmented SEIR model, initialized with a single infected
101 individual, first indicate that segregating site trajectories are sensitive to sampling effort, as
102 expected (Figure 1A,B). More specifically, we considered three different sampling strategies,
103 each with sequences binned in non-overlapping 4-day time windows to calculate segregating
104 site trajectories. These three sampling strategies consisted of a strategy with full sampling
105 effort (all sequences per 4-day time window), one with dense sampling effort (40 sequences
106 per 4-day time window) and one with sparse sampling effort (20 sequences per 4-day time
107 window). With all three of these sampling efforts, the number of segregating sites first
108 increases as the epidemic grows, with mutations accumulating in the virus population.
109 Following the peak of the epidemic, the number of segregating sites starts to decline as viral
110 sublineages die out, reducing the amount of genetic variation present in the viral population. A
111 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 Figure 1A, and a higher R_0 of 2.0. 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 10-15% 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 (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).

141 To compare segregating site trajectories between these simulations, we therefore shifted the
142 simulation with transmission heterogeneity later in time such that the two simulated epidemics
143 peaked at similar times (Figure 1E). Comparisons of segregating site trajectories between these
144 simulations indicated that transmission heterogeneity decreases the number of segregating
145 sites during any time window (Figure 1F). As expected, higher levels of transmission
146 heterogeneity result in a greater decrease in the number of segregating sites (Figure S2). These
147 results indicate that transmission heterogeneity needs to be taken into consideration when
148 estimating epidemiological parameters using segregating site trajectories.

149 Finally, we wanted to assess whether changes in R_0 over the course of an epidemic would leave
150 signatures in segregating site trajectories. We considered this scenario because phylodynamic
151 inference has often been used to quantify the effect of public health interventions on R_0, most
152 recently in the context of SARS-CoV-2 (Danesh et al. 2020; Miller et al. 2020). We thus
153 implemented simulations with R_0 starting at 1.6 and then either remaining at 1.6 or reduced to
154 either 1.1 or 0.75 when the number of infected individuals reached 400 (Figure 1G). The
155 segregating site trajectories for these three simulations indicate that reductions in R_0 over the
156 course of an epidemic leave signatures in this summary statistic of viral diversity (Figure 1H).
157 The signatures left in the segregating site trajectories reflect the epidemiological dynamics that
158 result from the reductions in R_0. Reducing R_0 to 1.1 results in a slower increase in the number of
159 cases, a delayed epidemic peak, and a delayed decline in cases; as such, the number of
160 segregating sites increases more slowly and peaks later. Reducing R_0 to 0.75 results in an
161 immediate decline in cases, with an observed drop in the number of segregating sites due to
162 the stochastic loss of sublineages. Similar magnitude reductions in R_0 that occurred later on in
163 the simulated epidemic resulted in fainter signatures of this effect in the segregating site
164 trajectories (Figure S3).

165

166 **Epidemiological inference using segregating site trajectories**

167 To examine the extent to which inference based on segregating sites can be used for
168 epidemiological parameter estimation, we generated a mock segregating site trajectory by

169 forward simulating an SEIR model with a R_0 of 1.6. From this simulation, we randomly sampled
170 500 viral sequences and binned these sequences into 4-day time windows based on their
171 sampling times (Figure 2A). Figure 2B shows the segregating site trajectory from these binned
172 sequences. From this trajectory, we first attempted to estimate only R_0 under the assumption
173 that the timing of the index case t_0 is known. We estimated an R_0 value of 1.54 (95% confidence
174 interval of 1.37 to 1.81; Materials and Methods; Figure 2C), demonstrating that our segregating
175 sites inference approach applied to this simulated dataset is able to recover the true R_0 value of
176 1.6. Lower levels of sampling effort decreased the lower bound of the R_0 estimate to 1.17
177 (Figure S4); the true R_0 value of 1.6, however, still fell within the estimated 95% confidence
178 interval. Instead of random sampling of infected individuals, adopting a more uniformly
179 distributed sampling strategy acted to reduce the uncertainty in the R_0 estimate (Figure S5). In
180 Figure S6, we present results for the same set of sequences as those used in Figure 2, with the
181 sequence data now binned in time windows of 1 day, 2 days, 6 days, and 10 days, instead of 4
182 days. These results show that R_0 estimates are not biased by the use of different time window
183 lengths.

184 Because the timing of the index case (in cases with a single introduction) is almost certainly not
185 known for an emerging epidemic, we further attempted to estimate both R_0 and t_0 using the
186 segregating site trajectory shown in Figure 2B. To get an idea of the likelihood landscape over
187 this parameter space, we first considered a range of R_0 values between 1.0 to 2.5 and a range of
188 t_0 within 50 days of the true start date of 0. We then divided this parameter space into fine
189 resolution parameter combinations (R_0 intervals of 0.1 and t_0 intervals of 2 days), and ran 20
190 SMC simulations for every parameter combination. In Figure 3A, we plot the mean value of the
191 20 SMC log-likelihoods for every parameter combination in the considered parameter space.
192 Examination of this plot indicates that there is a log-likelihood ridge that runs between early
193 t_0/low R_0 parameter combinations and late t_0/high R_0 parameter combinations. However, this
194 ridge falls off on both sides, indicating that inference using segregating site trajectories can in
195 principle estimate both t_0 and R_0. The maximum likelihood estimate for R_0 was 1.7 (95%
196 confidence = 1.5 to 2.0; true value = 1.6). The maximum likelihood estimate for t_0 was 14 days
197 (95% confidence = -12 to 32 days post the true t_0 date of 0). Our results indicate that joint

198 estimation of these parameters is thus possible in cases where a single introduction is
199 responsible for igniting local circulation. Using our estimates of R_0 and t_0, we reconstructed the
200 dynamics of the segregating sites (Figure 4A) and unobserved state variables: the number of
201 susceptible, exposed, and infected individuals over time (Figures 4B, C, D). These reconstructed
202 state variables captured the true epidemiological dynamics, demonstrating that our segregating
203 sites approach can be used to infer epidemiological variables that generally go unobserved.

204 As mentioned in the Introduction, there are existing phylodynamic inference approaches
205 available that can estimate epidemiological model parameters using viral phylogenies that have
206 been reconstructed from sequence data. Of particular note is the coalescent-based inference
207 approach developed by Volz (Volz 2012) that has been implemented as PhyDyn (Volz and
208 Siveroni 2018) in BEAST2. To compare our results using the segregating sites approach to
209 results using PhyDyn, we generated mock viral nucleotide sequences from our set of 500
210 sampled sequences (Materials and Methods) and used these nucleotide sequences as input
211 into PhyDyn. Assuming the same epidemiological model structure and using uninformative
212 priors, PhyDyn was similarly able to recover the true R_0 value of 1.6 used in the forward
213 simulation (Figure 3B; 95% credible interval = 1.43 to 1.61). Because PhyDyn infers
214 epidemiological parameters using a tree-based method, the program does not estimate the
215 time of the index case t_0. Instead, it estimates the time of the most recent common ancestor
216 (tMRCA) of the viral phylogeny. This tMRCA is necessarily later than the time of the index case
217 (Pekar et al. 2020). The tMRCA estimate's credible interval spanned from -28.22 to 0.96 days
218 post the true time of the index case (t_0 = 0). As such, interpretation of the PhyDyn results would
219 almost certainly result in timing the index case t_0 as less than 0 (too early). The early estimate of
220 t_0 may be due to the "push-of-the-past" effect described by Nee and coauthors, which results
221 from the assumption of deterministic dynamics in the inference process when the underlying
222 population dynamics are stochastic (and conditioned on the persistence of a lineage)(Nee et al.
223 1994). This "push-of-the-past" effect is usually reflected in an overestimate of the growth rate
224 (or an overestimate in R_0) in coalescent-based inference approaches that are applied to
225 datasets with small population sizes (here, of infected individuals). Here, because R_0 controls
226 not only the rate of increase in the number of infected individuals at the start of the simulated

227 pandemic but also the time at which the simulated pandemic starts to decline, the "push-of-
228 the-past" effect may instead be reflected in a tMRCA estimate that occurs too early.

229 Because the impetus for developing the segregating sites inference approach was based on the
230 extent of phylogenetic uncertainty present early on in an epidemic, we re-applied the inference
231 approach to sequences sampled between day 32 and day 52 of the epidemic. Our results on
232 this subset of simulated data indicate that R_0 and t_0 could again be jointly estimated, although
233 the confidence intervals for R_0 and t_0 were both broader, as expected with a shorter time series
234 (Figure 5A). Using this same set of sequences, PhyDyn estimated the credible interval for R_0 to
235 be 1.54 to 5.41, such that it tended to overestimate this parameter. In contrast to the PhyDyn
236 results shown in Figure 3B, the tMRCA credible interval estimated by PhyDyn on this shorter
237 time series was 14.5 to 31.0 days post the true t_0 of 0 (Figure 5B), indicating potential
238 consistency with this parameter. The tendency for PhyDyn to overestimate the basic
239 reproduction number in this shorter time series that includes only the exponential growth
240 phase of the simulated pandemic is consistent with expectations based on the "push-of-the-
241 past" effect.

242 **Epidemiological inference using SARS-CoV-2 sequences from France**

243 We applied the segregating sites inference approach to a set of SARS-CoV-2 sequences sampled
244 from France between January 23, 2020 and March 17, 2020, when a country-wide lockdown
245 was implemented. We decided to apply our approach to this set of sequences for several
246 reasons. First, a large fraction of the 479 available full-genome sequences from France over this
247 time period appear to be genetically very similar to one another (Gámbaro et al. 2020),
248 indicating that one major lineage may have taken off in France (or at least, that most samples
249 stemmed from one major lineage). This lineage would be the focus of our analysis. Second, an
250 in-depth epidemiological analysis previously inferred R_0 for France prior to the March 17 lock-
251 down measures that were implemented (Salje et al. 2020). This analysis fit a compartmental
252 infectious disease model to epidemiological data that included case, hospitalization, and death
253 data. Because our segregating sites inference approach can accommodate epidemiological
254 model structures of arbitrary complexity, we can adopt the same model structure as in this

255 previous analysis. We can also set the epidemiological parameters that are assumed fixed in
256 this previous analysis to their same values. By controlling for model structure and the set of
257 model parameters assumed as given, we can ask to what extent sequence data corroborate the
258 R_0 estimates arrived at from detailed fits to epidemiological data.

259 To apply our segregating sites approach to the viral sequences from France, we first identified
260 the subset of the 479 sequences that constituted a single, large lineage. To keep with the "tree-
261 free" emphasis of our approach, we identified this subset of n = 432 sequences without
262 inferring a phylogeny (Materials and Methods). Using phylogenetic inference, however, we
263 confirmed that our subset of sequences constituted a single clade, with excluded sequences
264 from France falling outside of this clade (Figure S7). To generate a segregating site trajectory
265 from these sequences, we defined 4-day time windows such that the last time window ended
266 on March 17, 2020. Figure 6A shows the number of sequences falling into each time window.
267 Figure 6B shows the segregating site trajectory calculated from these sequences.

268 We parameterized the model with a per genome, per transmission mutation rate of μ = 0.33
269 using consensus sequence data from established SARS-CoV-2 transmission pairs that were
270 available in the literature (James et al. 2020; Popa et al. 2020; Braun et al. 2021; Lythgoe et al.
271 2021) (Materials and Methods). Specifically, for each of the 87 transmission pairs we had access
272 to, we calculated the nucleotide distance between the consensus sequence of the donor
273 sample and that of the recipient sample and fit a Poisson distribution to these data (Figure 6C).
274 Using this approach, we estimated a μ value of 0.33 (95% confidence interval of 0.22 to 0.48),
275 corresponding approximately to one mutation occurring every 3 transmission events.

276 Similar to the approach we undertook with our simulated data, we first attempted to estimate
277 R_0 and the timing of the index case t_0 for this segregating site trajectory. We considered a broad
278 parameter space over which to calculate log-likelihood values. Specifically, we considered R_0
279 values between 1.0 and 4.5 and t_0 values of December 1st, 2019 and February 14th, 2020. We
280 ran 20 SMC simulations and calculated the mean log-likelihood for each parameter combination
281 (Figure 7). We estimated R_0 to be 2.3 (95% confidence interval = 1.7 to 3.6), consistent with the
282 R_0 estimate of 2.9 (95% confidence interval = 2.81 to 3.01) arrived at through epidemiological

time series analysis (Salje et al. 2020). The maximum likelihood estimated of t_0 was January 18th, 2020 (95% confidence interval = December 22nd 2019 to February 8th, 2020). Our maximum likelihood estimate of t_0 in the middle of January 2020 aligns well with findings from Gámbaro et al. (2020).

Despite the reasonable estimate for R_0 we obtained under this model, we decided to also consider an alternative model that allowed for multiple introductions of the focal lineage into France (Materials and Methods). This decision was based on evolutionary analyses that have shown that regional SARS-CoV-2 epidemics in Europe (as well as in the US) were initiated through multiple introductions rather than only a single one (Worobey et al. 2020). Instead of attempting to 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 in the outside reservoir, 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^{-2}) and three different t_e values: December 24th, 2019, January 1st, 2020, and January 8th, 2020 (Materials and Methods). At each of these t_e values, we ran 20 SMC simulations and calculated the mean log-likelihood for each parameter combination (Figure 8A-C). The maximum likelihood estimates of R_0 were 2.7 (95% confidence interval = 2.1 to 3.6), 2.5 (95% confidence interval = 2.0 to 3.7), and 2.7 (95% confidence interval = 1.9 to 3.7), respectively, under t_e = December 24th, 2019, January 1st, 2020 and January 8th, 2020. These results indicate that the inferred R_0 values are relatively insensitive to the assumed emergence time of the basal genotype in the outside reservoir. At later emergence times, the estimates for η were higher, indicating that a later emergence time was compensated for by a higher transmission rate between the outside reservoir and individuals in 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 space shown in Figure 8, according to the same approach as the one we used for reconstructing

state variables on the mock segregating sites trajectory (Figure 9). As expected for an epidemic with an $R_0 > 1$, the total number of infected individuals increased exponentially over the time period considered (Figures 9D-F). In Figures 9G-I, we plot the reconstructed cumulative number of recovered individuals over time. These cumulative trajectories indicate that by mid-March 2020, approximately 0.02% to 0.15% of individuals in France had recovered from infection from this SARS-CoV-2 lineage. 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). Based on a survey of 3221 individuals, this study found 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 (Figures 9G-I). Our estimates fall in line with these independent estimates. Finally, in Figures 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), the sampled particles showed evidence of inferred repeated introductions of this lineage into France, with the majority of particles pointing towards 5-100 introductions of these lineage into France prior to February 22.

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 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, 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 important epidemiological parameters such as 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 2.5-2.7 (with a 95% confidence interval of approximately 2.0 to 3.7). 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. Most notably, it relies on an assumption of random sampling of individuals and on an assumption that all genetic variation that is observed is phenotypically neutral. These assumptions are shared by existing phylodynamic inference methods, although some newer statistical developments have relaxed the assumption of neutrality by incorporating adaptive evolution (Rasmussen and Stadler 2019). 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.

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 studies that have managed to incorporate additional data while performing phylodynamic inference have shown the value in pursuing 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.

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, each fueling the start of more local epidemics (Gonzalez-Reiche et al. 2020; Miller et al. 2020). In this case, 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, under specified constraints such as the basic reproduction number being the same across all clades or specified sets of clades. When distinct sets of clades are allowed to differ in their reproductive numbers, questions relating to the selective advantage of some clades over others can also 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 when questions involving emerging clades are a focus.

Materials and Methods

Epidemiological model simulations and calculation of segregating site trajectories. To generate simulated (mock) data, as well as for inference purposes, we specify a compartmental epidemiological model (of arbitrary complexity) and simulate the model under demographic stochasticity using Gillespie's τ-leap algorithm. As a concrete example of a compartmental epidemiological model, we here use a susceptible-exposed-infected-recovered (SEIR) model whose stochastic dynamics are governed by the following equations:

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

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

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

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

where:

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

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

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

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. 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 partition exposed individuals and infected individuals into genotype classes, with basal genotype (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 genotypes are numbered chronologically according to 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, 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.

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, $N_{E\rightarrow I}$ individuals are drawn at random from the set of individuals who are currently exposed. 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, $N_{I\rightarrow R}$ individuals are drawn at random from the set of individuals who are currently infected. These individuals will transition to the recovered class during this time step. To simulate transmission, during a time step Δt, we add $N_{S\rightarrow E}$ 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 new individual enters the same

414 genotype class as its parent. If one or more mutations occur during transmission, this new
415 individual enters a new genotype class, and the sparse matrix is extended to document the new
416 genotype and its associated mutations (given as integers, without a bitstring or explicit genome
417 structure).

418 We start the simulation with one infected individual carrying a viral genotype that we consider
419 as the 'reference' genotype (genotype 0). To calculate a time series of segregating sites, we
420 define a time window length T ($T > \Delta t$) of a certain number of days and partition the simulation
421 time course into discrete, non-overlapping time windows. During simulation, we keep track of
422 the individuals that recover (transition from I to R) within a time window. For each time
423 window i, we then sample n_i of these individuals at random, where n_i is the number of
424 sequences sampled in a given time window based on the sampling scheme chosen. Because we
425 have the genotypes of the sampled individuals from the sparse matrix, we can calculate for any
426 time window i, the number of segregating sites S_i. S_i is simply the number of polymorphic sites
427 across the sampled individuals in time window i. While in our simulations, we sample
428 individuals as they recover, alternative sampling schemes can instead be assumed. For example,
429 individuals could be sampled as they transition from the exposed to the infected class, or while
430 they are in the infected class.

431 **Epidemiological inference using time series of segregating sites.** Our inference approach relies
432 on particle filtering, also known as Sequential Monte Carlo (SMC), to estimate model
433 parameters and reconstruct latent state variables. The underlying forward model we use is
434 formulated as a state-space model, with epidemiological variables (e.g., S, E, I, and R) being
435 latent (that is, unobserved) variables in the process model. The model is simulated using
436 Gillespie's τ-leap algorithm, as described in the section above. The evolutionary component of
437 the model also contributes to the process model. For the observation model, we perform k
438 'grabs' of sampled individuals, with each grab consisting of the following steps:

439 • pick (without replacement) n_i individuals from the set of individuals who recovered during
440 time window i, where n_i is the number of samples present in the empirical dataset in
441 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.

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 mass function parameterized with the mean S_i^{sim} value and evaluated at S_i. 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.

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. For inference on the mock dataset, a population size of 100,000 was used, and we set the mean duration of the exposed period to be 2 days ($1/\gamma_E$) and the mean duration of the infectious period to be 3 days ($1/\gamma_I$).

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 downsample sequences to reduce runtime.

Implementation of the transmission heterogeneity model. We implement transmission heterogeneity by splitting 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 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 (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 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 (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 of secondary infections. We set P to 0.80, to make p_h easily interpretable relative to the "20/80" rule in disease ecology (Woolhouse

483 et al. 1997). 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:

484 $\frac{R_0\gamma_I}{p_h c+(1-p_h)}$, and set $\beta_h = c\beta_l$.

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

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

499 A single-deme structured coalescent prior as defined by the following equations was
500 implemented using PhyDyn v1.3.8. All sampled sequences were assigned to the "I" class:

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

501 $$\frac{dI}{dt} = \gamma_E E - \gamma_I I$$

502 $$\frac{dR}{dt} = \gamma_I I$$

503 where: $\beta = R_0\gamma_I$. A population size of 100,000 with a single initially infected individual was
504 used. We assume infected individuals remain exposed for an average of 2 days ($1/\gamma_E$) and
505 infectious ($1/\gamma_I$) for an average of 3 days. R_0 was estimated using a uniform [1.0, 10.0] prior.

506 Sampled parameters and trees were logged every 1000 states and all MCMC chains were run
507 for at least 90M iterations. The first 10% of MCMC chains were discarded as burn-in and the ESS
508 values of all parameters were >200, as identified by Tracer v1.7.1 (10.1093/sysbio/syy032).

509 **Epidemiological model structure and parameterization used for the SARS-CoV-2 analysis.** The
510 process model we use in our application to SARS-CoV-2 sequence data from France is based on
511 a previously published epidemiological model (Salje et al. 2020). We base our process model on
512 this published model to allow for a direct comparison of inferred R_0 values between our
513 sequence-based analysis and their analysis that focuses on SARS-CoV-2 spread in France over a
514 similar time period. Their analysis was based on fitting an epidemiological model to a
515 combination of case, hospitalization, and death data. Their model structure, once implemented
516 using Gillespie's τ-leap algorithm, is given by:

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

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

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

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

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

517 where:

518 $$N_{S\to E1} \sim Pois(\beta \frac{S_t}{N} I_t \Delta t) + Pois(\beta \frac{S_t}{N} E_{2,t} \Delta t)$$

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

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

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

519 with β being the transmission rate, the average duration of time spent in the E_1 class given by
520 $1/\gamma_{E1}$ = 4 days, the average duration of time spent in the E_2 class given by $1/\gamma_{E2}$ = 1 day, and
521 the average duration of time spent in the infected class given by $1/\gamma_I$ = 3 days. Their model
522 assumes that the transmission efficiency β of exposed class 2 (E_2) and that of the infected class I

523 are the same; their model considers E_2 and I as distinct classes to interface with the case data,
524 where symptoms are assumed to not appear before an individual is infected (in class I). We
525 maintain the model structure with E_1, E_2, and I rather than reducing it to a model structure with
526 just a single E and a single I class to keep the same distribution of infection times as in their
527 model.

528 Because SARS-CoV-2 dynamics are characterized by substantial levels of transmission
529 heterogeneity (Adam et al. 2020; Miller et al. 2020; Sun et al. 2021) and we have shown in
530 Figure 1 that transmission heterogeneity impacts segregating site trajectories, we expanded the
531 compartmental epidemiological model described above to include transmission heterogeneity
532 in a manner similar to the one we used in Figure 1. Based specifically on the analysis by Paireau
533 and colleagues (Paireau et al. 2022), we set p_h to 0.10, such that 10% of infections are
534 responsible for 80% of secondary infections. Analogous to the approach we undertook for the
535 simulated data, we estimated R_0 and t_0 using the segregating site trajectory shown in Figure 6B.

536 Based on phylogenetic analyses that have indicated that early introductions of SARS-CoV-2 into
537 focal regions likely resulted from multiple introductions rather than a single one, we further
538 considered a modified version of the epidemiological model that would allow multiple
539 introductions. The modification relied on the incorporation of infections that resulted from
540 contact with an outside-of-France viral reservoir. Similar to the approach adopted by some
541 existing phylodynamic analyses, e.g., (Geidelberg et al. 2021), the viral population dynamics in
542 this reservoir are simplified to exponential growth. This infected population from outside of
543 France acts as another source of infection for susceptible individuals within France, allowing the
544 multiple introductions of SARS-CoV-2 into France.

545 As in the focal region, new genotypes are expected to emerge in the outside reservoir. As we
546 assume an infinite sites model, the genotypes that emerge in the outside reservoir and in the
547 focal region will not overlap except in the basal genotype that is first introduced to the focal
548 region. For this reason, and because the basal genotype is expected to be more prevalent than
549 any of the viral genotypes that stem from it, we consider only the repeated introduction of the
550 basal genotype into France. Starting at the time of emergence of the basal genotype in the

551 outside reservoir, we let the number of individuals infected with this basal genotype Y_t grow
552 exponentially:

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

553 where r is the intrinsic growth rate of the basal genotype and t_e is the time of emergence for
554 the basal genotype. We set the intrinsic growth rate to 0.22 day^{-1}, based on empirical estimates
555 (Dehning et al. 2020; Musa et al. 2020). To set t_e, we first identified the genotype sampled in
556 France that is genetically closest to the reference strain Wuhan/Hu-1 (MN908947.3). This basal
557 genotype differs from Wuhan/Hu-1 by 4 nucleotides: C241T, C3037T, C14408T, and A23403G.
558 Using GISAID data, we then identified sequences with collection locations outside of France
559 that carried all four of these mutations that define the basal genotype. The earliest of these
560 sequences including the four basal genotype-defining mutations was collected on January 8th,
561 2020 in the Netherlands, suggesting that the basal genotype had been circulating prior to
562 January 8^{th,} 2020. Considering the potential delay between emergence and the time of first
563 detection, we considered three distinct t_e values: December 24th, 2019, January 1st, 2020, and
564 January 8th, 2020.

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

$$N^{O}_{S\rightarrow E1} \sim Pois(\beta\eta \frac{S_t}{N} Y_t \Delta t)$$

571 As we are considering only the dynamics of the basal genotype in the outside reservoir, all of
572 these newly infected individuals will carry the basal genotype unless mutation occurs during the
573 transmission process.

574

575 **Estimation of the per genome, per transmission event mutation rate**

576 We set the per-genome, per-transmission mutation rate parameter μ to 0.33. This is based on
577 the fit of a Poisson distribution to the number of *de novo* substitutions that were observed in
578 87 transmission pairs of SARS-CoV-2 from four studies (James et al. 2020; Popa et al. 2020;
579 Braun et al. 2021; Lythgoe et al. 2021). Accession numbers for 78/87 of these transmission pairs
580 are available in Table S1. Accession numbers for the remaining pairs were provided by the
581 corresponding authors of the relevant publication (Lythgoe et al. 2021). Sequence data were
582 aligned to Wuhan/Hu-1 (MN908947.3) (Wu et al. 2020) using MAFFT v.7.464 (Katoh 2002).
583 Insertions relative to Wuhan/Hu-1 were removed and the first 55 and last 100 nucleotides of
584 the genome were masked. *De novo* substitutions for each pair were identified in Python v.3.9.4
585 (http://www.python.org) using NumPy v.1.19.4 (Harris et al. 2020). Ambiguous nucleotides
586 were considered in the identification of *de novo* substitutions (i.e. an R nucleotide was assumed
587 to match both an A and a G). The mean number of substitutions between transmission pairs is
588 the Maximum Likelihood Estimate for the λ parameter of the Poisson distribution. The 95%
589 confidence intervals were calculated using the exact method using SciPy v.1.5.4 (SciPy 1.0
590 Contributors et al. 2020) such that the lower bound was $\frac{(\mathrm{X}^2_{2Y,0.025})/2}{87}$ and the upper bound was
591 $\frac{(\mathrm{X}^2_{2(Y+1),0.975})/2}{87}$ where Y is the total number of observed substitutions.

592 The value for μ = 0.33 is consistent with population-level substitution rate estimates for SARS-
593 CoV-2, which range from 7.9 x 10^{-4} to 1.1 x 10^{-3} substitutions per site per year (Duchene et al.
594 2020; Pekar et al. 2020). With a genome length of SARS-CoV-2 of approximately 30,000
595 nucleotides and a generation interval of approximately 4.5 days (Griffin et al. 2020), these
596 population-level substitution rates would correspond to per genome, per transmission
597 mutation rates of between 0.29 and 0.41, respectively.

598 **Estimation of segregating site trajectories for the France data.**

599 We downloaded all complete and high-coverage SARS-CoV-2 sequences with complete
600 sampling dates sampled through March 17th, 2020 (https://www.france24.com/en/20200316-
601 live-france-s-macron-addresses-nation-amid-worsening-coronavirus-outbreak) in France and

602 uploaded through April 29th, 2021 from GISAID (Shu and McCauley 2017). Sequences were
603 aligned to Wuhan/Hu-1 using MAFFT v.7.464 Insertions relative to Wuhan/Hu-1 were removed.
604 Any sequences with fewer than 28000 A, C, T, or G characters were removed. Following this
605 filtering protocol our dataset included 479 sequences. We masked the first 55 and last 100
606 nucleotides in the genome as well as positions marked as "highly homoplasic" in early SARS-
607 CoV-2 sequencing data (https://github.com/W-L/ProblematicSites_SARS-
608 CoV2/blob/master/archived_vcf/_problematic_sites_sarsCov2.2020-05-27.vcf). Pairwise SNP
609 distances were calculated in a manner that accounted for IUPAC ambiguous nucleotides in
610 Python using NumPy. To subset these data to a single clade circulating within France, we
611 identified the connected components of this pairwise distance matrix with a cutoff of 1 SNP in
612 Python using SciPy and identified the shared SNPs relative to Wuhan/Hu-1 between all
613 sequences in each connected component. The largest connected component contained 308
614 sequences which shared the substitutions C241T, C3037T, C14408T, and A23403G. Our final
615 dataset included these 308 as well as 122 sequences from connected components that shared
616 these four substitutions relative to Wuhan/Hu-1. We included connected components in which
617 all sequences had an N at any of the four clade-defining sites of the largest connected
618 component. Two sequences were excluded as they differed from all other sequences in the
619 dataset by > 7 SNPs. This dataset is similar to the set of sequences analyzed in Danesh et al.
620 (2020). Sequences were binned into four-day windows, aligned such that the last window
621 ended on the latest sampling date. The number of segregating sites in each window was
622 calculated in Python using NumPy. Ambiguous nucleotides were considered in the calculation
623 of segregating sites.

624 **Phylogenetic analysis of SARS-CoV-2 sequences from France.**

625 To confirm that the subset of sequences from France obtained from finding connected
626 components formed an evolutionary lineage/clade, we first combined the 479 sequences
627 sampled from France with 100 randomly-selected complete, high-coverage, collected date
628 complete sequences sampled from outside France through March 17th, 2020 and uploaded to
629 GISAID through April 29th, 2021. These sequences were aligned to Wuhan/Hu-1 using MAFFT,
630 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 full, dense, and sparse sampling efforts. The segregating site trajectory under full sampling effort is shown with a black dashed line. Dense sampling (black lines) corresponds to 40 sequences sampled per time window. Sparse sampling (gray 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 (blue line) compared to those of the R_0 = 1.6 simulation (black 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 (blue lines) and the R_0 = 1.6 simulation (black lines). (E) Simulated dynamics of infected individuals (I) under an SEIR model with an R_0 of 1.6 and incorporating transmission heterogeneity (green, dashed line) compared to those of the original R_0 = 1.6 simulation (black line) without transmission heterogeneity. Transmission heterogeneity was included by setting the parameter p_h to 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 (green, solid line). (F) Segregating site trajectories for the shifted transmission heterogeneity simulation (green lines) and the simulation without transmission heterogeneity (black lines). (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 black has R_0 remaining at 1.6. (H) Segregating site trajectories for the three simulations shown in subplot G. Dense sampling effort (40 sequences sampled per time window) was used to generate all segregating site trajectories shown in subplots D, F, and G. 30 randomly-sampled segregating site trajectories are shown for each sampling effort in subplot B and for each epidemiological scenario in subplots D, F, and G. 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, $I_h(t_0)$ = 1, $I_l(t_0)$ = 0, and $R(t_0)$ = 0. A time step of τ = 0.1 days was used in the Gillespie τ-leap algorithm.

676

677 **Figure 2. Epidemiological inference on a simulated trajectory of segregating sites.** (A) The number of
678 sampled sequences over time, binned by 4-day time windows. Sampling was done in proportion to the
679 number of individuals recovering in a time window. In all, 500 sequences were sampled over the course
680 of the simulated epidemic. (B) Simulated segregating site trajectory from the sampled sequences, by
681 time window. (C) Estimation of R_0 using Sequential Monte Carlo (SMC). Points show log-likelihood values
682 from different SMC simulations with 2500 random drawn R_0 values, uniformly sampled from a range
683 between [1.0, 2.5). Solid black curve shows the moving average of 150 data points of the log-likelihood
684 values. The vertical red dashed line shows the maximum likelihood estimate (MLE) of R_0. The red band
685 shows the 95% confidence interval of R_0 (1.37-1.81). The vertical blue line shows the true value of R_0 =
686 1.6. MLE and 95% CI were obtained from moving average log-likelihood values. Model parameters for
687 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
688 the per genome, per transmission mutation rate μ = 0.2. Initial conditions are $S(t_0)$ = N-1, $E(t_0)$ = 0, $I(t_0)$ =
689 1, and $R(t_0)$ = 0. A time step of τ = 0.1 days was used in the Gillespie τ-leap algorithm.

690

691 **Figure 3. Joint estimation of the basic reproduction number (R_0) and the timing of the index case (t_0)**
692 **using simulated data, and comparison against PhyDyn.** (A) The likelihood surface based on the
693 segregating site trajectory shown in Figure 2B is shown over a range of R_0 and t_0 values. Blank cells
694 yielded log-likelihood values of negative infinity. The log-likelihood value shown in each cell is the mean
695 log-likelihood value calculated from 20 SMC simulations. (B) Joint density plot for R_0 and the time of the
696 most recent common ancestor (tMRCA), as estimated using PhyDyn (Volz and Siveroni 2018) on the set
697 of 500 sampled sequences shown in Figure 2A.

698

699 **Figure 4. Reconstruction of unobserved state variables.** (A) Simulated trajectory of the number of
700 segregating sites (dashed red), alongside reconstructed trajectories of the number of segregating sites
701 (gray). (B) Simulated dynamics of susceptible individuals (dashed red), alongside reconstructed dynamics
702 of susceptible individuals (gray). (C) Simulated dynamics of exposed individuals (dashed red), alongside
703 reconstructed dynamics of exposed individuals (gray). (D) Simulated dynamics of infected individuals
704 (dashed red), alongside reconstructed dynamics of infected individuals (gray). Reconstructed state
705 variables were obtained by running the particle filter using R_0 and t_0 parameter values sampled based on
706 their mean log-likelihood values shown in Figure 3A.

707

708 **Figure 5. Joint estimation of the basic reproduction number (R_0) and the timing of the index case (t_0)**
709 **using early samples from the simulation, and comparison against PhyDyn.** (A) The likelihood surface
710 based on a segregating site trajectory calculated using sequences from day 32 through day 52 of the
711 simulated pandemic. Sequences were binned into 4-day windows, with 10 individuals sampled from
712 each window. Blank cells yielded log-likelihood values of negative infinity. As in Figure 3A, the log-
713 likelihood value shown in each cell is the mean log-likelihood value calculated from 20 SMC simulations.
714 (B) Joint density plot for R_0 and the time of the most recent common ancestor (tMRCA), as estimated
715 using PhyDyn on the same set of 50 sampled sequences.

716

717 **Figure 6. Sequences and parameters used for epidemiological inference based on SARS-CoV-2**
718 **sequences from France.** (A) The number of sampled sequences over time, calculated using a 4-day time
719 window. (B) The segregating site trajectory calculated from the binned sequences shown in (A). (C)
720 Estimation of the per-genome, per-transmission mutation rate μ. The histogram shows the fraction of
721 transmission pairs with consensus sequences that differed from one another by the number of
722 mutations shown on the x-axis. The Poisson estimate of the mutation rate from these data (black) is μ =
723 0.33 (95% CI = 0.22-0.48).

724

725 **Figure 7. Joint estimation of the basic reproduction number R_0 and the time of the index case t_0 for the**
726 **France SARS-CoV-2 data.** (A) The joint log-likelihood surface based on the estimated segregating site
727 trajectory for the France data. Each cell shows the mean log-likelihood value based on 20 SMC

728 simulations. Blank cells indicate mean log-likelihood values of negative infinity. Gray cells indicate where
729 log-likelihood values were not evaluated due to extended simulation time. (B) The same joint log-
730 likelihood surface showing cells as colored only if they fell within the 95% confidence interval of the
731 parameter estimates.

732

733 **Figure 8. Joint estimation of the basic reproduction number R_0 and the transmission-reduction**
734 **parameter η for the multiple-introductions model using the France data.** The joint log-likelihood
735 surface based on the estimated segregating site trajectory for the France data is shown under three
736 different basal genotype emergence times: t_e = December 24, 2019 (A), January 1, 2020 (B), and January
737 8, 2020 (C). Each cell shows the mean log-likelihood value based on 20 SMC simulations. Blank cells
738 indicate mean log-likelihood values of negative infinity. Gray cells indicate where log-likelihood values
739 were not evaluated due to extended simulation time.

740

741 **Figure 9. Trajectories of reconstructed state variables for the France data under the multiple-**
742 **introductions model.** State variables are reconstructed for the multiple-introductions model with three
743 different values assumed for the emergence time of the basal genotype: t_e = December 24, 2019 (first
744 column), January 1, 2020 (second column), and January 8, 2020 (third column). (A-C) Segregating site
745 trajectory for the France SARS-CoV-2 data (ref), alongside segregating site trajectories from 10 sampled
746 SMC particles (gray). (D-F) Reconstructed dynamics of the total number of infected individuals ($E_1 + E_2 +$
747 I), shown in percent of France's population. (G-I) Reconstructed dynamics of the cumulative number of
748 recovered individuals over time, shown in percent of France's population. Independent estimates of the
749 fraction of the population that has been infected with SARS-CoV-2 by mid-March are shown in black.
750 Estimates are from a serological study conducted during the time window March 9-15, 2020 (Le Vu et al.
751 2021). (J-L) Reconstructed dynamics of the cumulative number of infections in France that resulted from
752 contact with infected individuals outside of France. Reconstructed state variables shown in panels (A-L)
753 are based on 10 SMC simulations using parameter combinations of R_0 and η that were sampled in
754 proportion to their likelihood values.

755

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; and (C) 6 days. Under the dense sampling scheme, sampling efforts are 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 efforts are 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 show a single representative segregating site trajectory under the dense and sparse sampling efforts, respectively. Shorter time window lengths yield segregating site trajectories that are more variable between time windows, while longer time window lengths decrease the resolution on temporal trends in the number of segregating sites. Shorter time window lengths also result in a lower number of segregating sites, reflecting the smaller number of sequences per time window.

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 (blue line; p_h = 0.8). The transmission heterogeneity simulations considered span from low levels of transmission heterogeneity (p_h = 0.3, corresponding to 30% of the most infectious individuals responsible for 80% of secondary infections), to higher levels (p_h = 0.06, corresponding to 6% of the most infectious individuals responsible for 80% of secondary infections. Intermediate levels of transmission heterogeneity are parameterized with p_h = 0.15, corresponding to 15% of the most infectious individuals responsible for 80% of secondary infections. (B) Segregating site trajectories for the simulations shown in (A). All simulations are densely sampled (40 sequences sampled per time window of 4 days).

Figure S3. Segregating site trajectories under transmission reduction scenarios occurring at a later time point of the epidemic. (A) 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

786 1000. The simulation in red has R_0 decreasing to 1.1. The simulation in yellow has R_0 decreasing to 0.75.
787 The simulation in blue has R_0 remaining at 1.6. (B) Segregating site trajectories for the three simulations
788 shown in (A).

789

790 **Figure S4. Epidemiological inference on a simulated trajectory of segregating sites, with lower**
791 **sampling effort than in Figure 2.** (A) The number of sampled sequences over time, by time window.
792 Sampling was done in proportion to the number of individuals recovering in a time window. In all, 100
793 sequences were sampled over the course of the simulated epidemic. (B) Simulated segregating site
794 trajectory from the sampled sequences. (C) Estimation of R_0 using SMC. The maximum likelihood
795 estimate for R_0 was 1.38 [95% CI = 1.17 to 1.65].

796

797 **Figure S5. Epidemiological inference on a simulated trajectory of segregating sites, with uniform**
798 **rather than proportional sampling.** (A) The number of sampled sequences, by 4-day time window.
799 Uniform sampling was performed by sampling 13 sequences per time window. Time windows with
800 fewer than 13 sequences available were not included in the analysis. As such, here, only 494 sequences
801 were used for inference. (B) Simulated segregating site trajectory from the sampled sequences. (C)
802 Estimation of R_0 using SMC. The estimate for R_0 was 1.55 [95% CI = 1.49 to 1.63]. In contrast, the
803 proportional sampling strategy shown in Figure 2 had a higher level of uncertainty in its R_0 estimate (1.
804 54, with 95% CI = 1.37 to 1.81).

805

806 **Figure S6. Epidemiological inference on the same set of sampled sequences as in Figure 2, binned at**
807 **different time window lengths.** For all time window lengths, the same set of 500 proportionally-
808 sampled sequences were used. Top row: The number of sampled sequences over time, binned by time
809 window (1, 2, 4, 6, and 10 days, respectively). Middle row: Segregating site trajectories from the set of
810 binned sequences. Bottom row: Estimation of R_0 using SMC. Panels A-C show results using a time
811 window length of 1 day. The estimate for R_0 was 1.62 [95% CI = 1.38 to 1.78]. Panels D-F show results
812 using a time window length of 2 days. The estimate for R_0 was 1.60 [95% CI = 1.40 to 1.85]. Panels G-I
813 reproduce the results shown in Figure 2, using a time window length of 4 days. The estimate for R_0 was
814 1.54 [95% CI =1.37 to 1.81]. Panels J-L show results using a time window length of 6 days. The estimate

815 for R_0 was 1.58 [95% CI = 1.37 to 1.81]. Panels M-O show results using a time window length of 10 days.
816 The estimate for R_0 was 1.54 [95% CI = 1.36 to 1.82].

817

818 **Figure S7. Inferred phylogenies for the sequences sampled from France, January 23-March 17, 2020**.
819 (A) Divergence tree, showing the number of nucleotide substitutions from Wuhan/Hu-1. Sequences
820 from France are colored in blue, with dark blue coloring indicating sequences that were included in our
821 single-lineage analysis and light blue coloring indicating sequences that were excluded from our analysis.
822 Tips colored in gray denote genetically similar sequences sampled from outside of France during this
823 time period. (B) Time-aligned maximum likelihood phylogeny, with coloring of sequences as in (A).

824

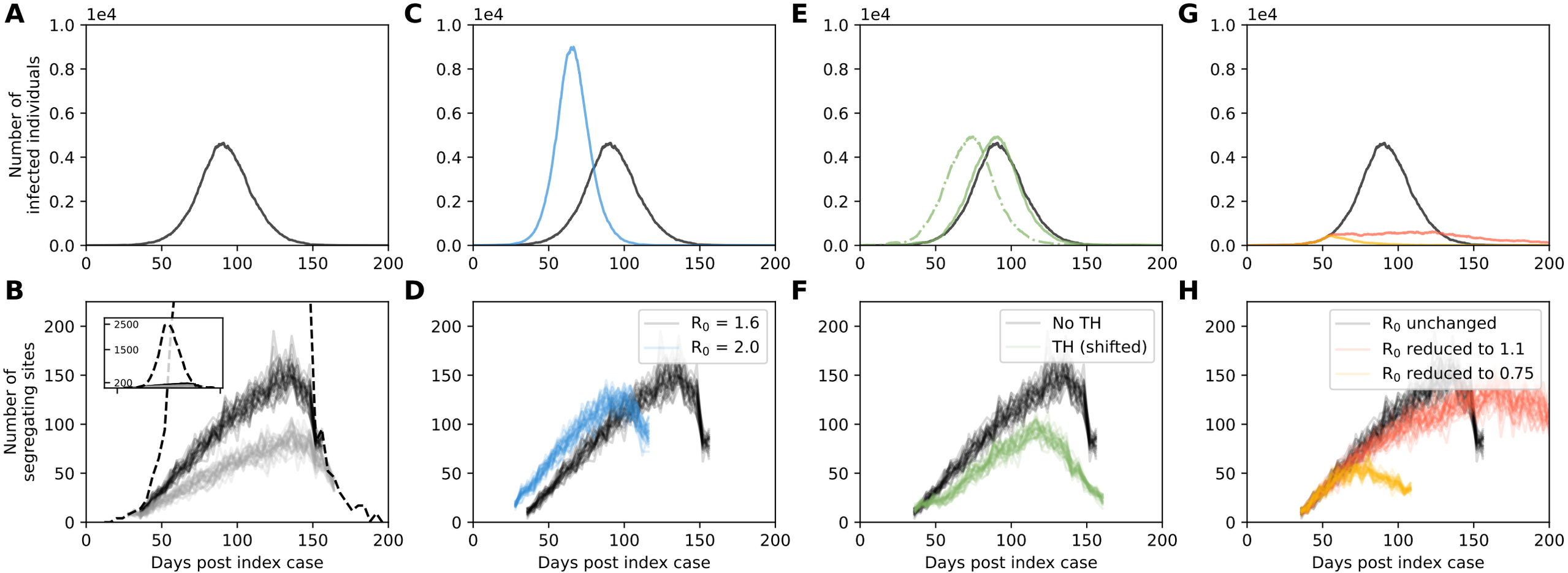
825 **Table S1.** Transmission pairs used to estimate the per-genome, per-transmission event mutation rate μ.
826 Accession numbers of the consensus sequences from the donor and the recipient of the transmission
827 pair are provided.

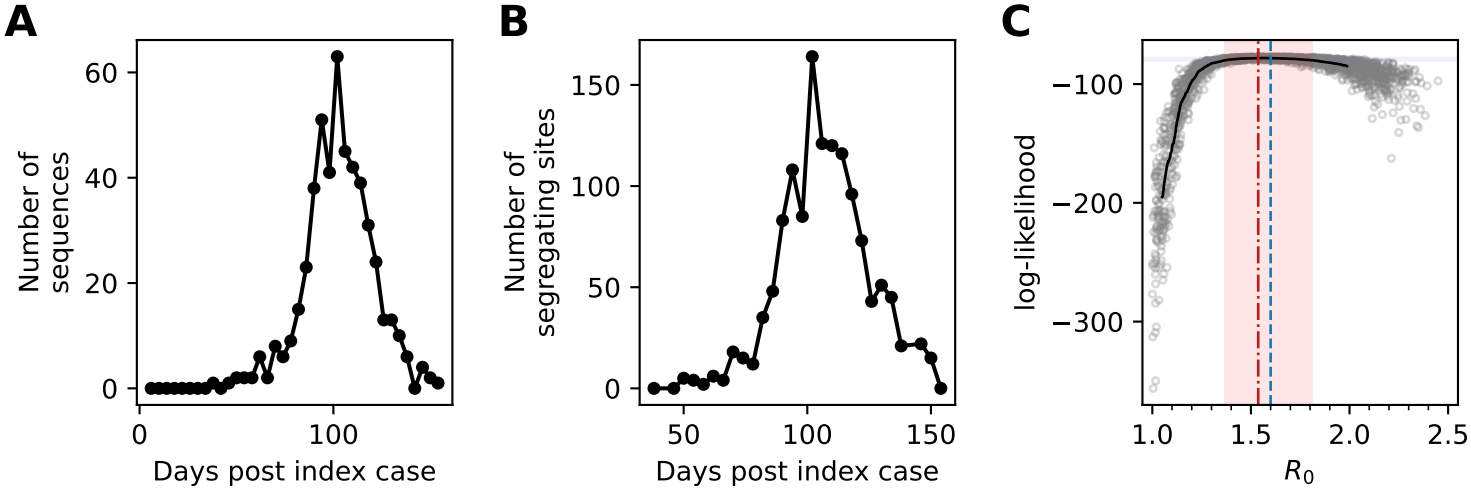
828

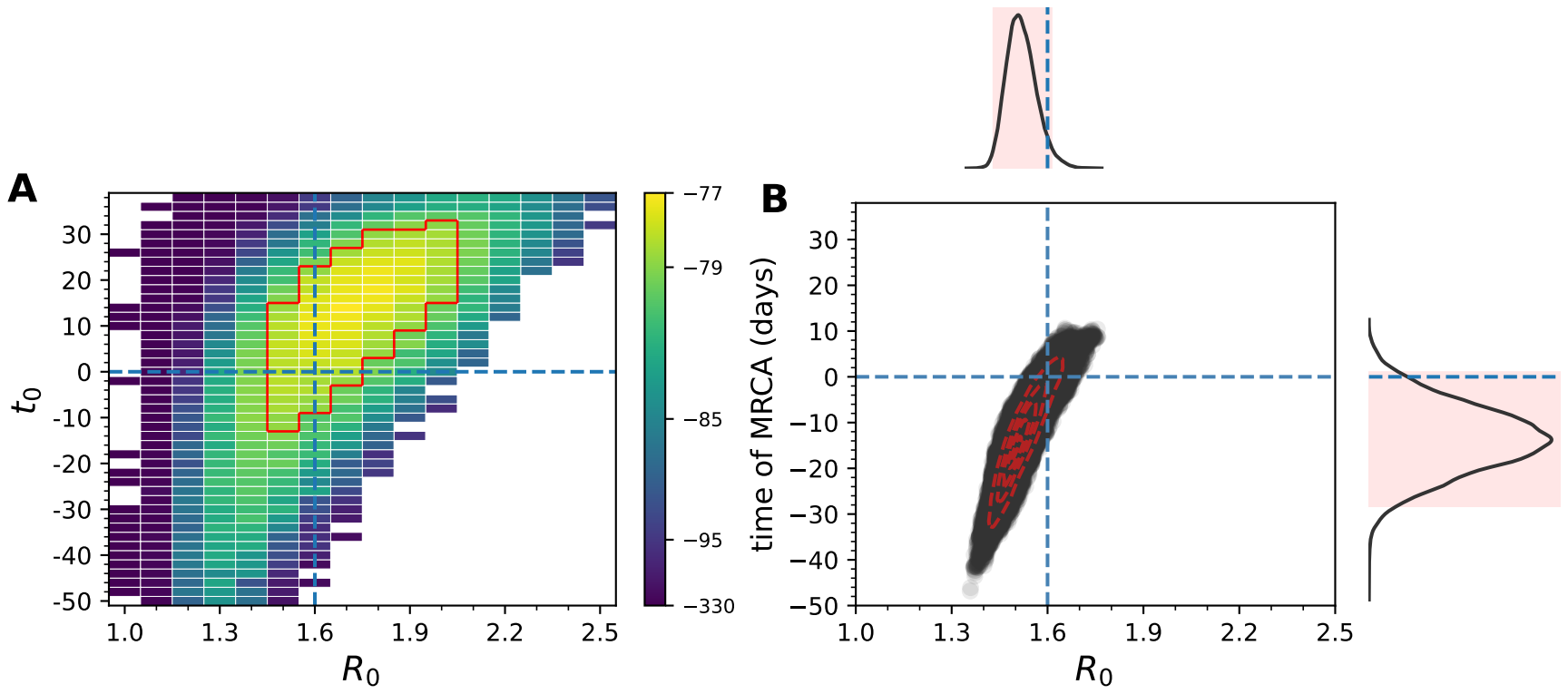
829

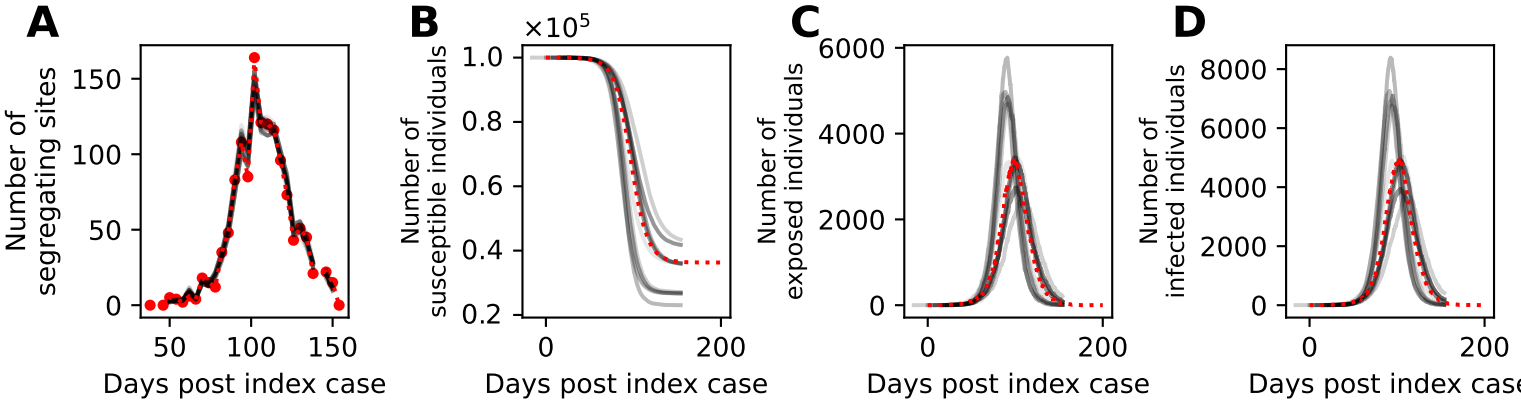
830 ## References

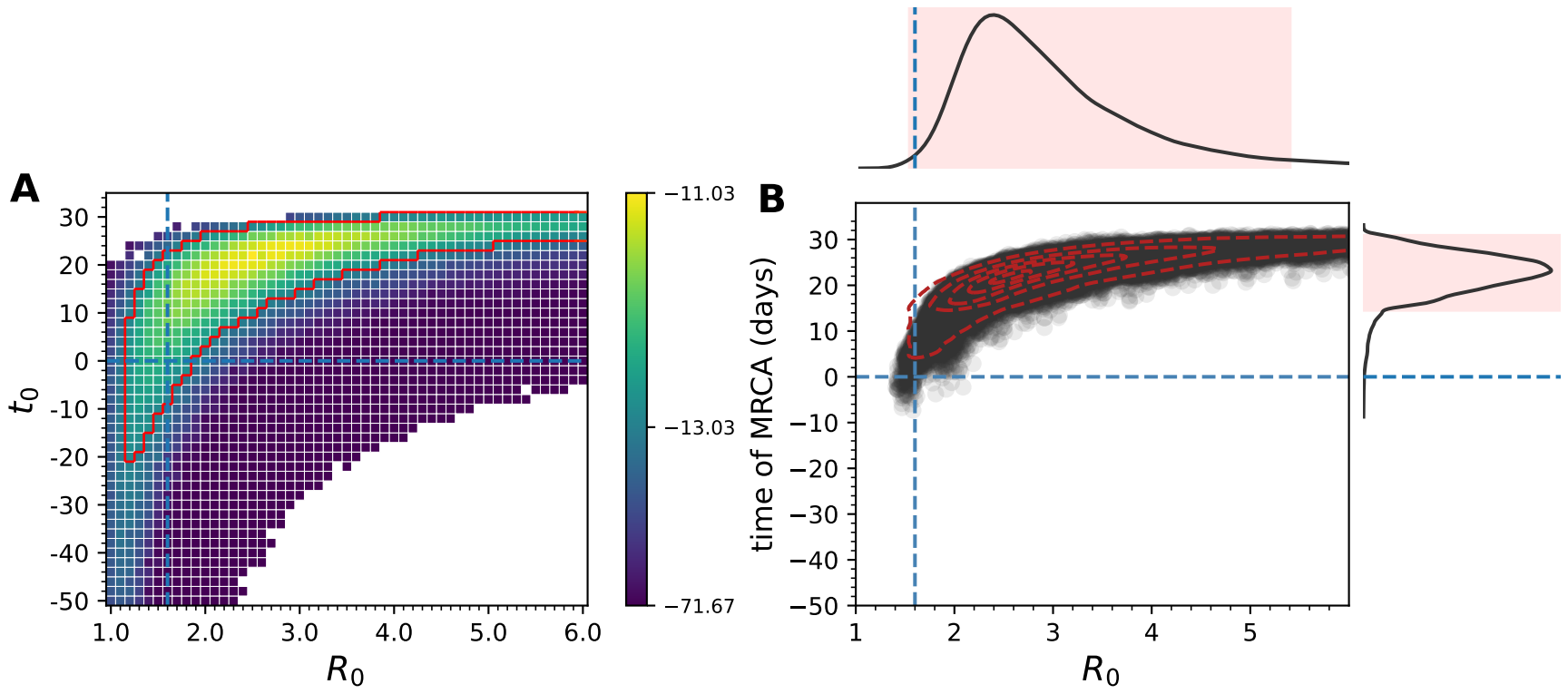
831 Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.

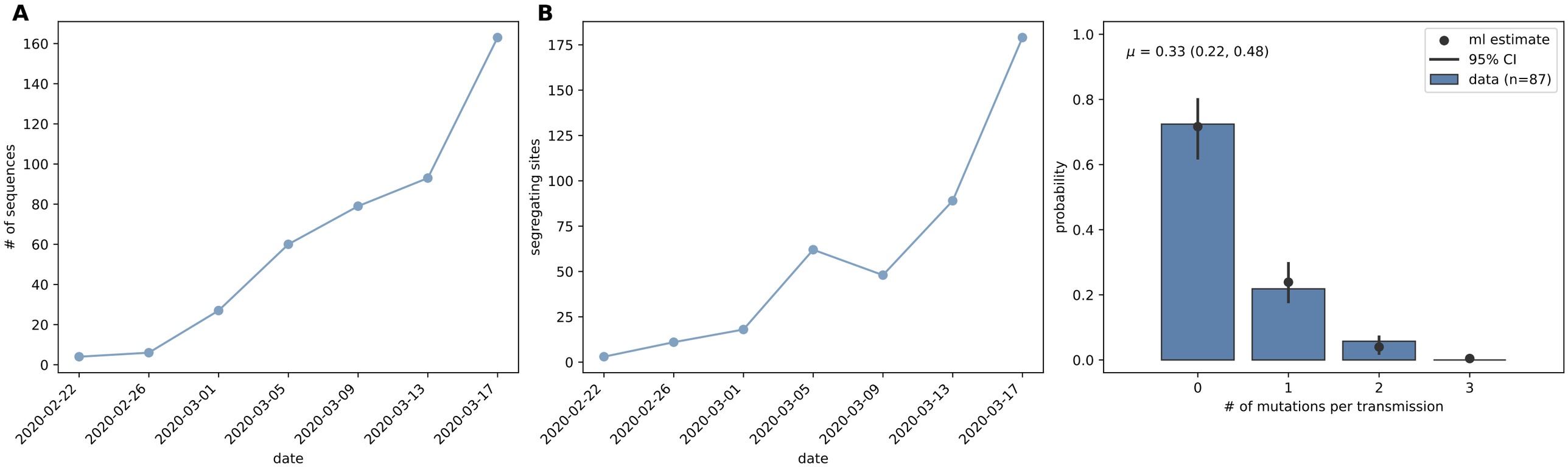

A
B
C
D
E
F
G
H
1e4
Number of infected individuals
Number of segregating sites
Days post index case
R_0 = 1.6
R_0 = 2.0
No TH
TH (shifted)
R_0 unchanged
R_0 reduced to 1.1
R_0 reduced to 0.75

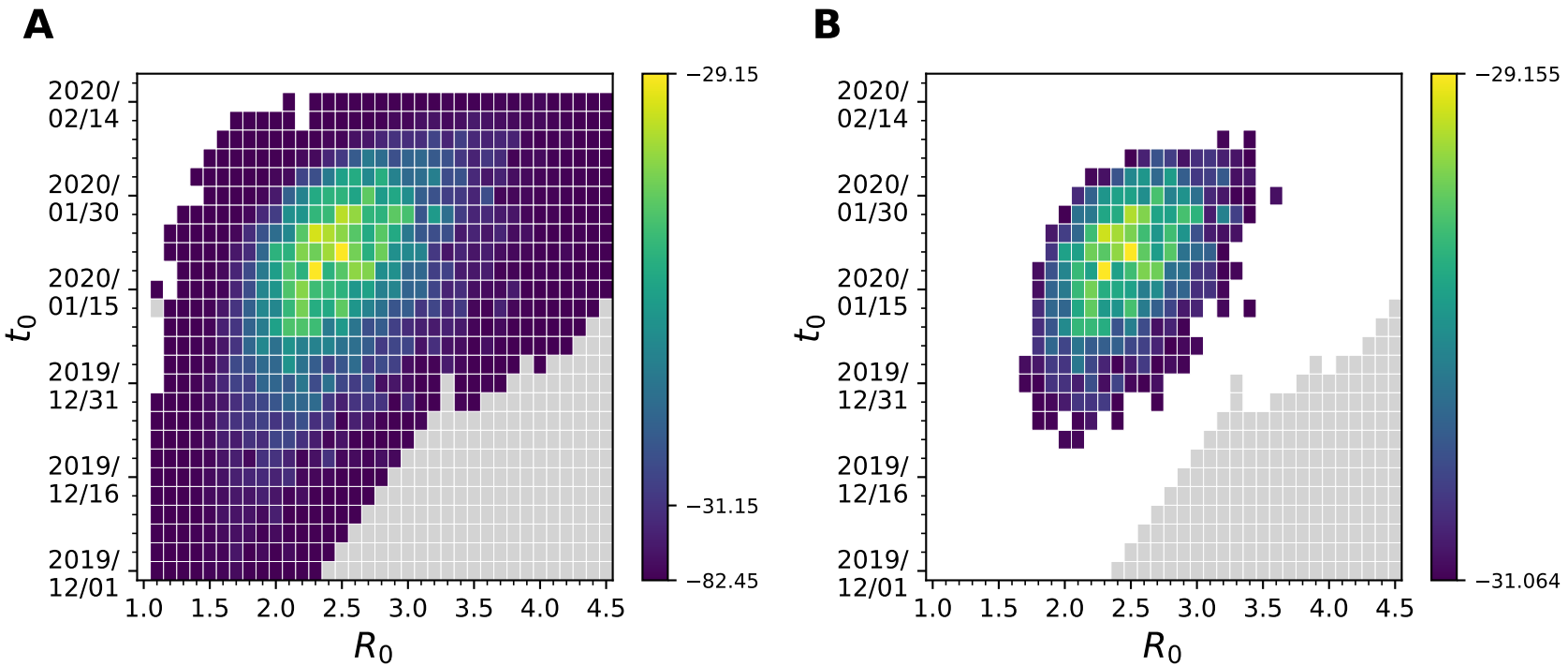

A
Number of sequences
0
20
40
60
0
100
Days post index case
B
Number of segregating sites
0
50
100
150
50
100
150
Days post index case
C
log-likelihood
−100
−200
−300
1.0
1.5
2.0
2.5
R_0

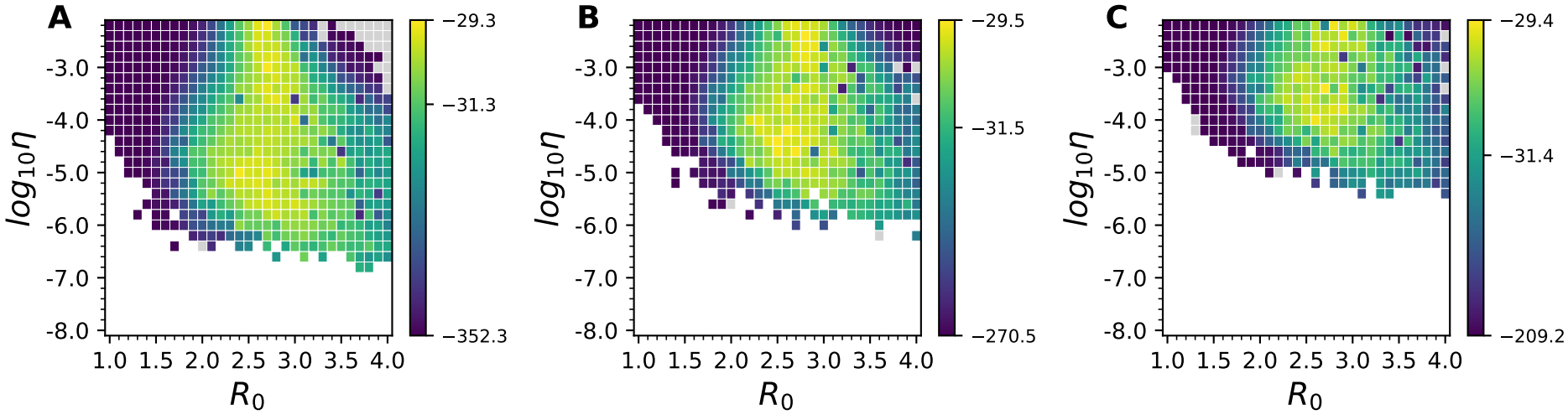

A
B
t_0
R_0
time of MRCA (days)
30
20
10
0
-10
-20
-30
-40
-50
1.0
1.3
1.6
1.9
2.2
2.5
−77
−79
−85
−95
−330

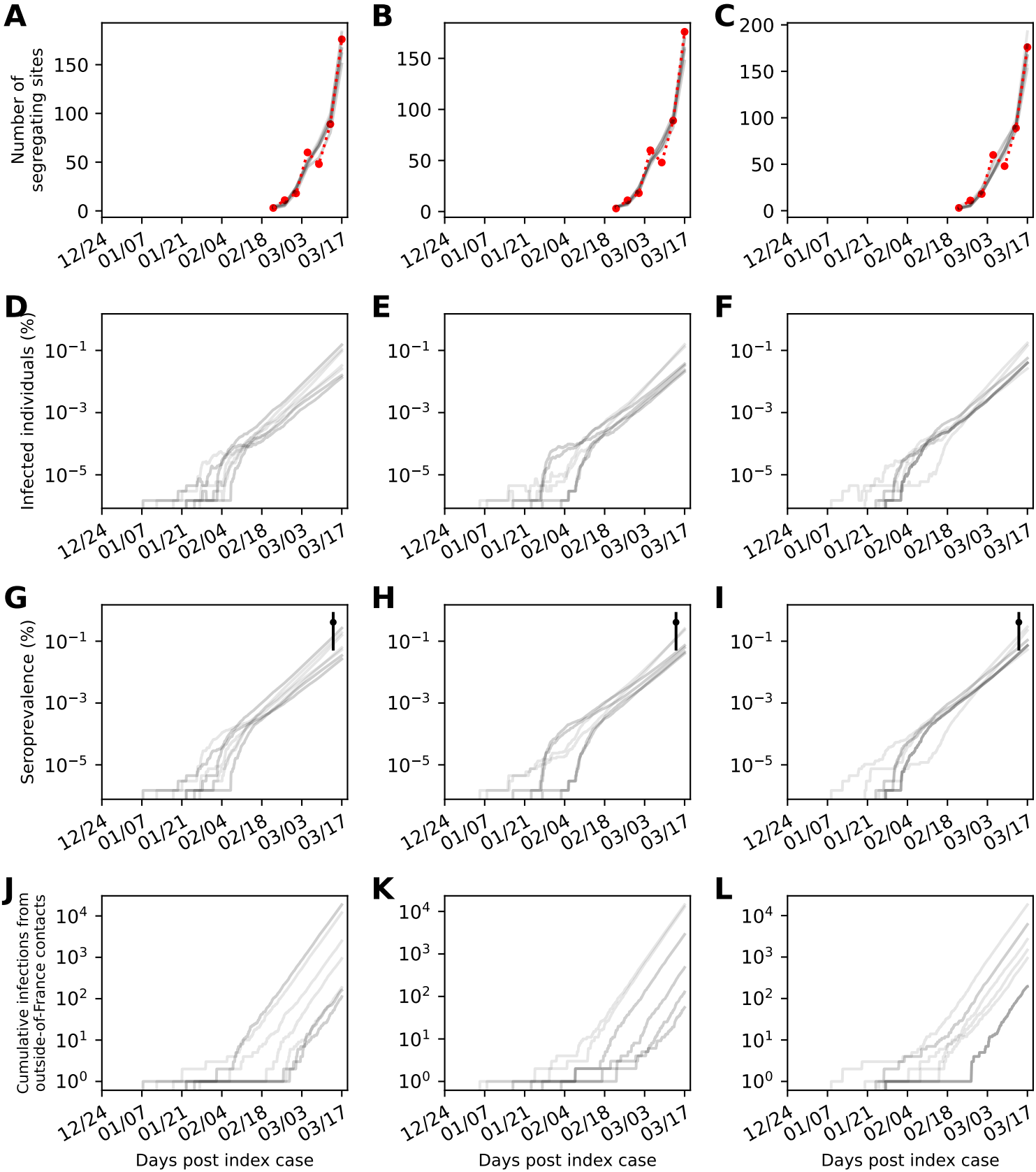

A
Number of segregating sites
150
100
50
0
50
100
150
Days post index case
B
×10^5
Number of susceptible individuals
1.0
0.8
0.6
0.4
0.2
0
200
Days post index case
C
Number of exposed individuals
6000
4000
2000
0
0
200
Days post index case
D
Number of infected individuals
8000
6000
4000
2000
0
0
200
Days post index case

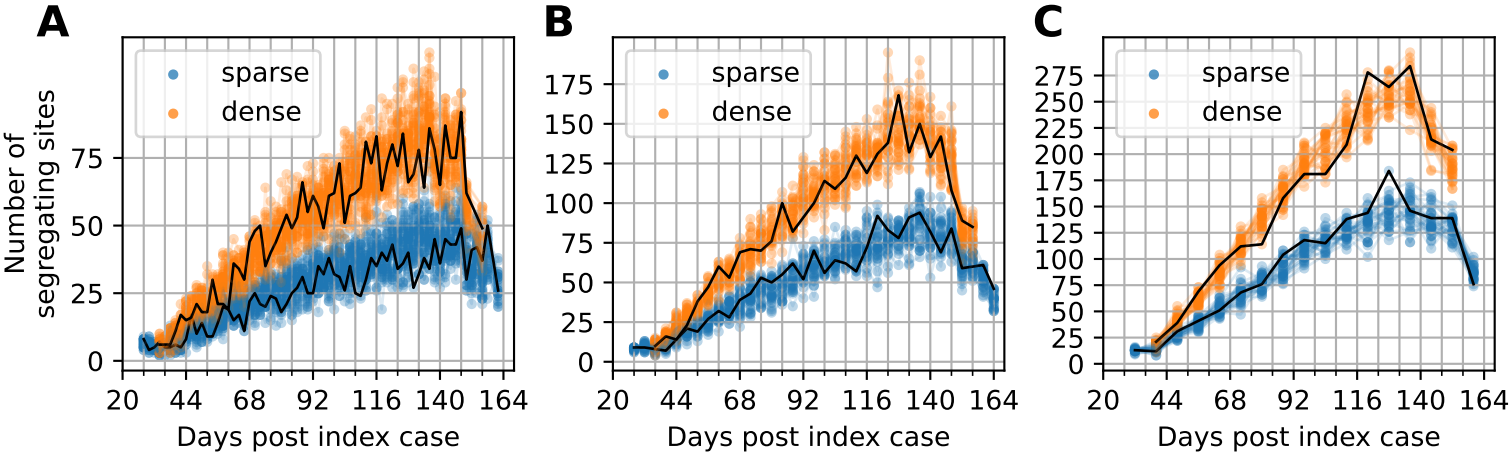

A
B
t_0
R_0
time of MRCA (days)
−11.03
−13.03
−71.67

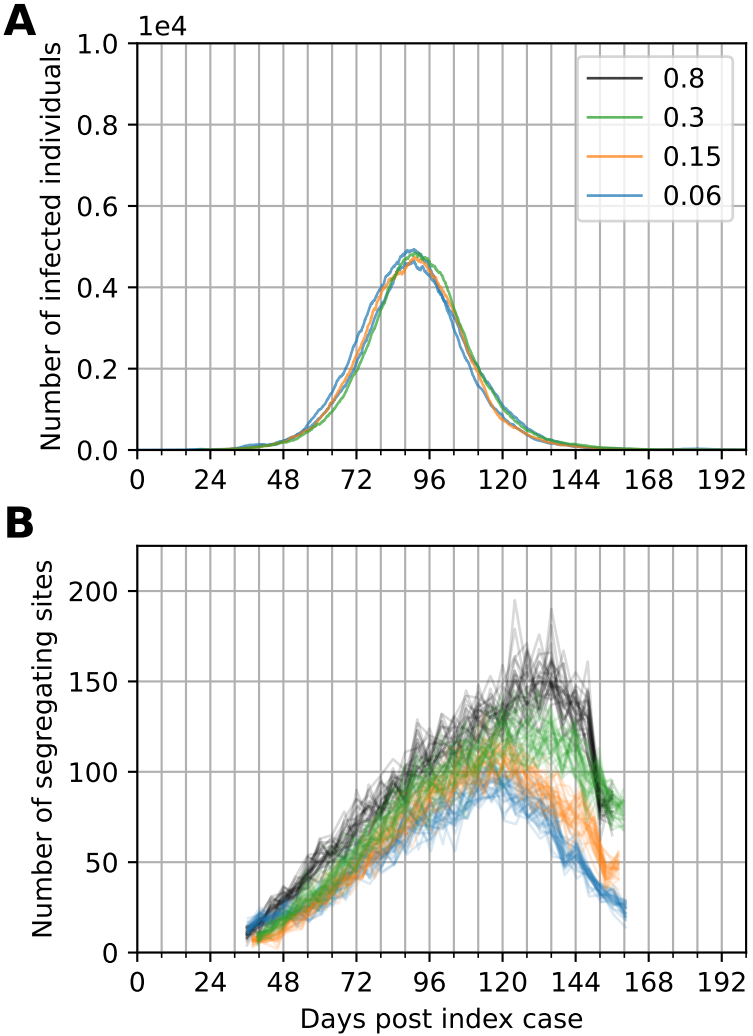

A
# of sequences
0
20
40
60
80
100
120
140
160
2020-02-22
2020-02-26
2020-03-01
2020-03-05
2020-03-09
2020-03-13
2020-03-17
date
B
segregating sites
0
25
50
75
100
125
150
175
2020-02-22
2020-02-26
2020-03-01
2020-03-05
2020-03-09
2020-03-13
2020-03-17
date
μ = 0.33 (0.22, 0.48)
ml estimate
95% CI
data (n=87)
probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
# of mutations per transmission

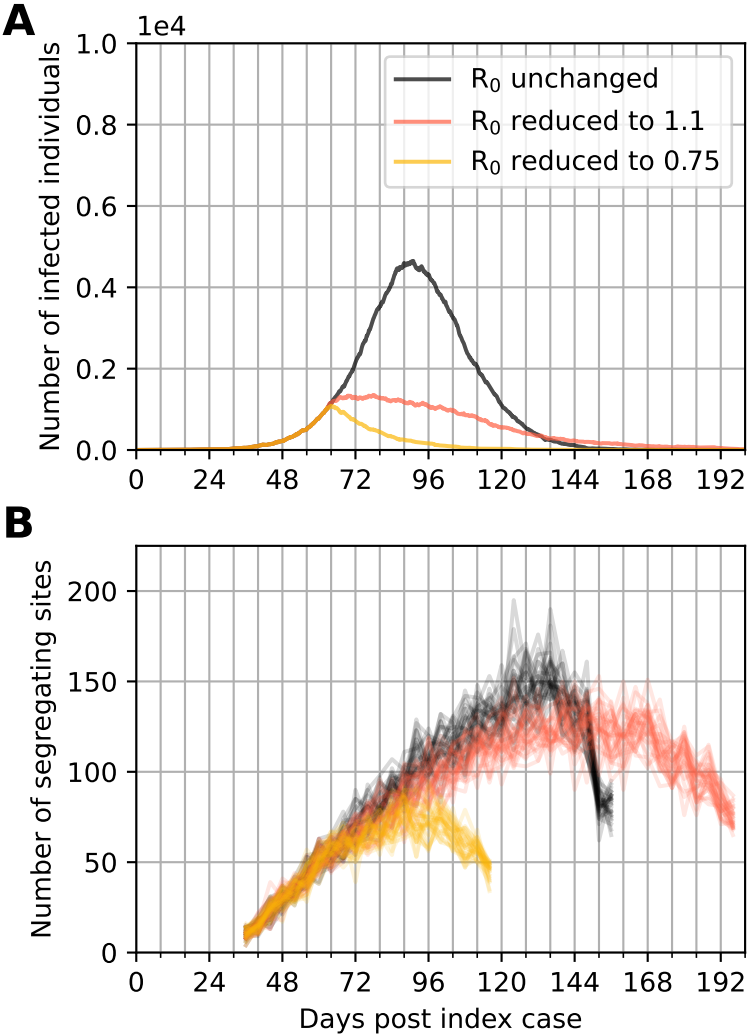

A
B
2020/ 02/14
2020/ 01/30
2020/ 01/15
2019/ 12/31
2019/ 12/16
2019/ 12/01
t_0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
R_0
−29.15
−31.15
−82.45
−29.155
−31.064

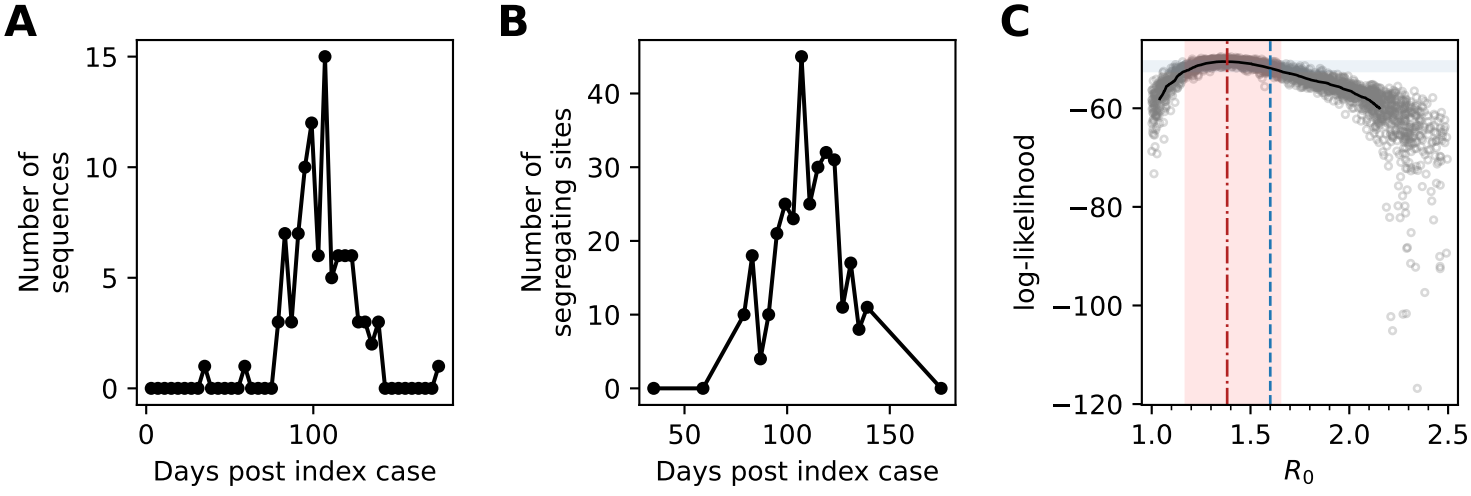

A
B
C
$log_{10}\eta$
R_0
-3.0
-4.0
-5.0
-6.0
-7.0
-8.0
1.0 1.5 2.0 2.5 3.0 3.5 4.0
−29.3
−31.3
−352.3
−29.5
−31.5
−270.5
−29.4
−31.4
−209.2

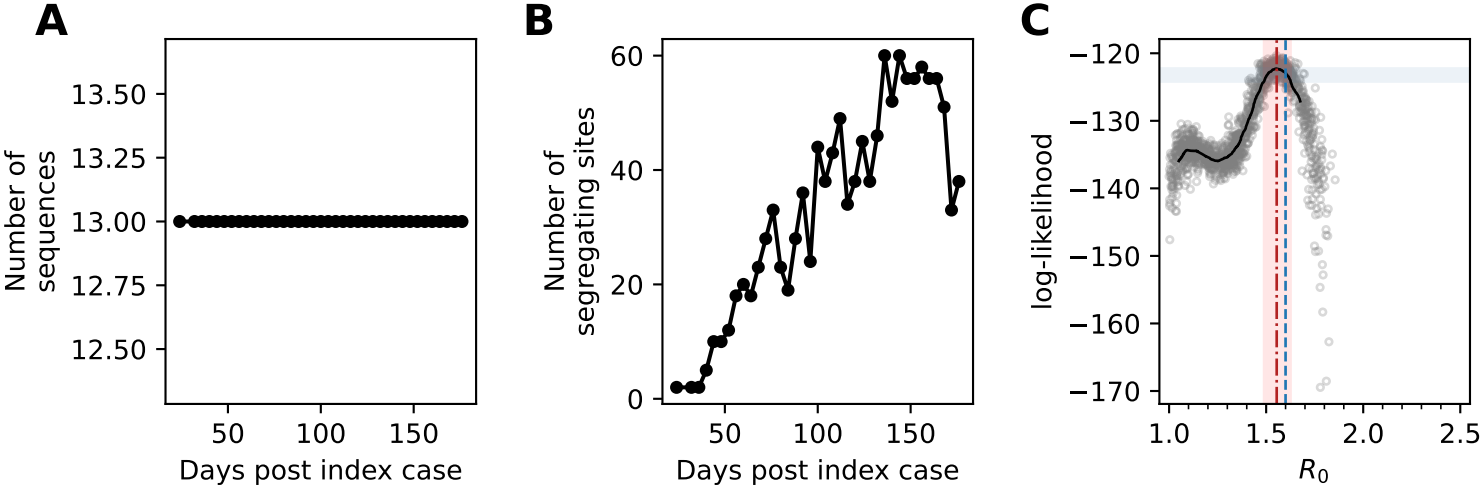

A
B
C
D
E
F
G
H
I
J
K
L
Number of segregating sites
Infected individuals (%)
Seroprevalence (%)
Cumulative infections from outside-of-France contacts
0
50
100
150
200
10^{-1}
10^{-3}
10^{-5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
12/24
01/07
01/21
02/04
02/18
03/03
03/17
Days post index case

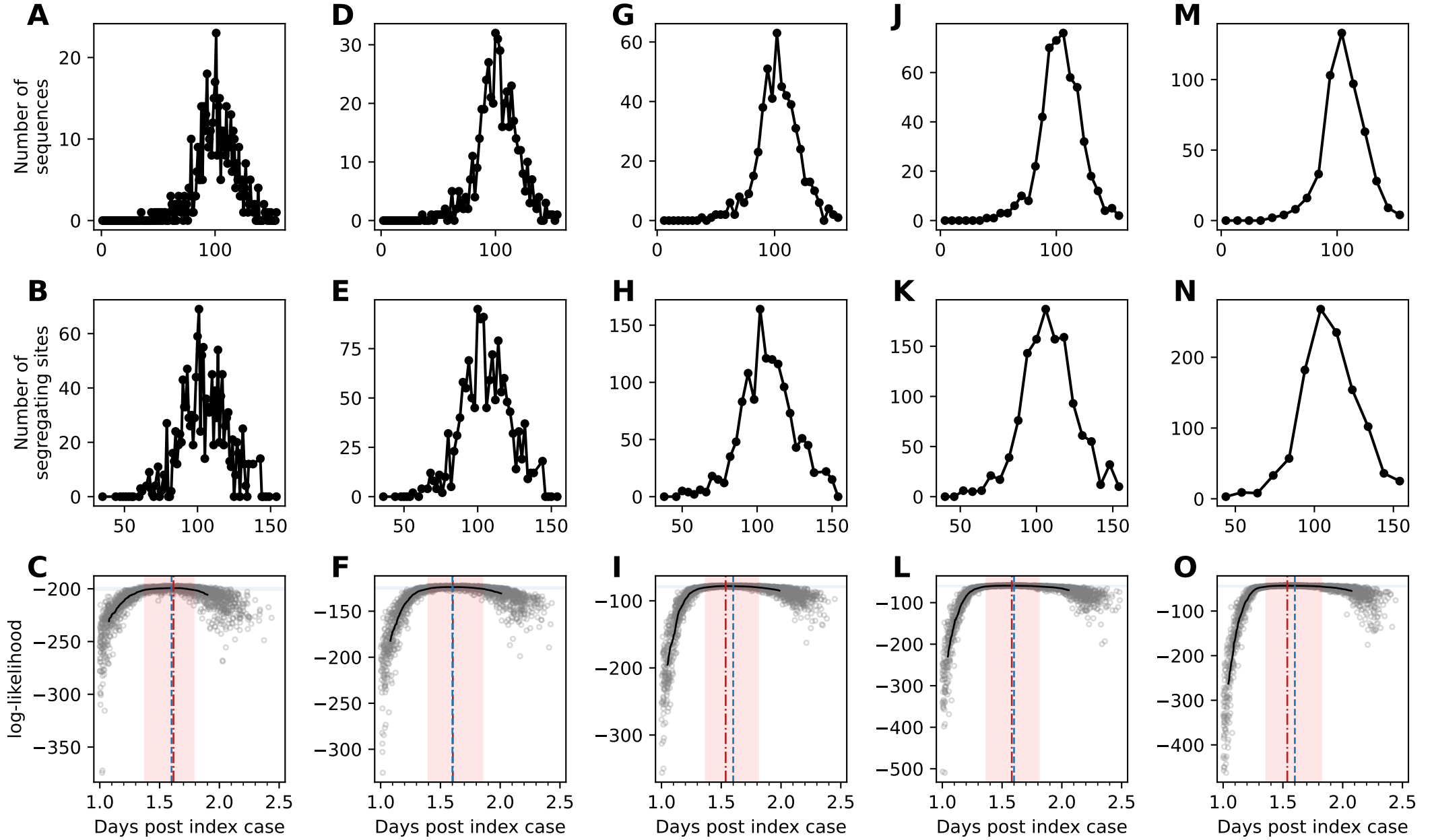

A
B
C
sparse
dense
Number of segregating sites
Days post index case
0
25
50
75
100
125
150
175
200
225
250
275
20
44
68
92
116
140
164

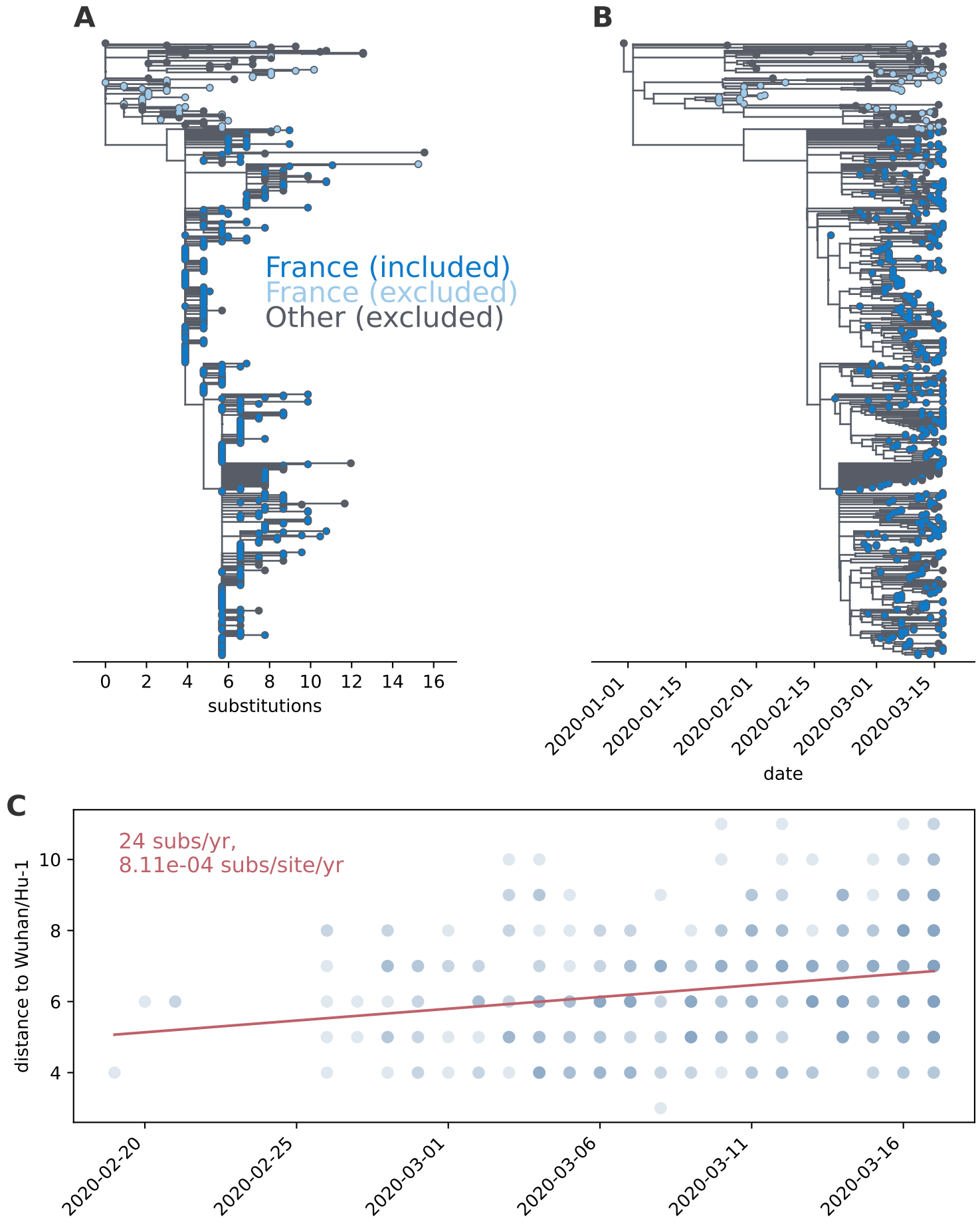

A
1e4
1.0
0.8
0.6
0.4
0.2
0.0
Number of infected individuals
0.8
0.3
0.15
0.06
0
24
48
72
96
120
144
168
192
B
200
150
100
50
0
Number of segregating sites
0
24
48
72
96
120
144
168
192
Days post index case

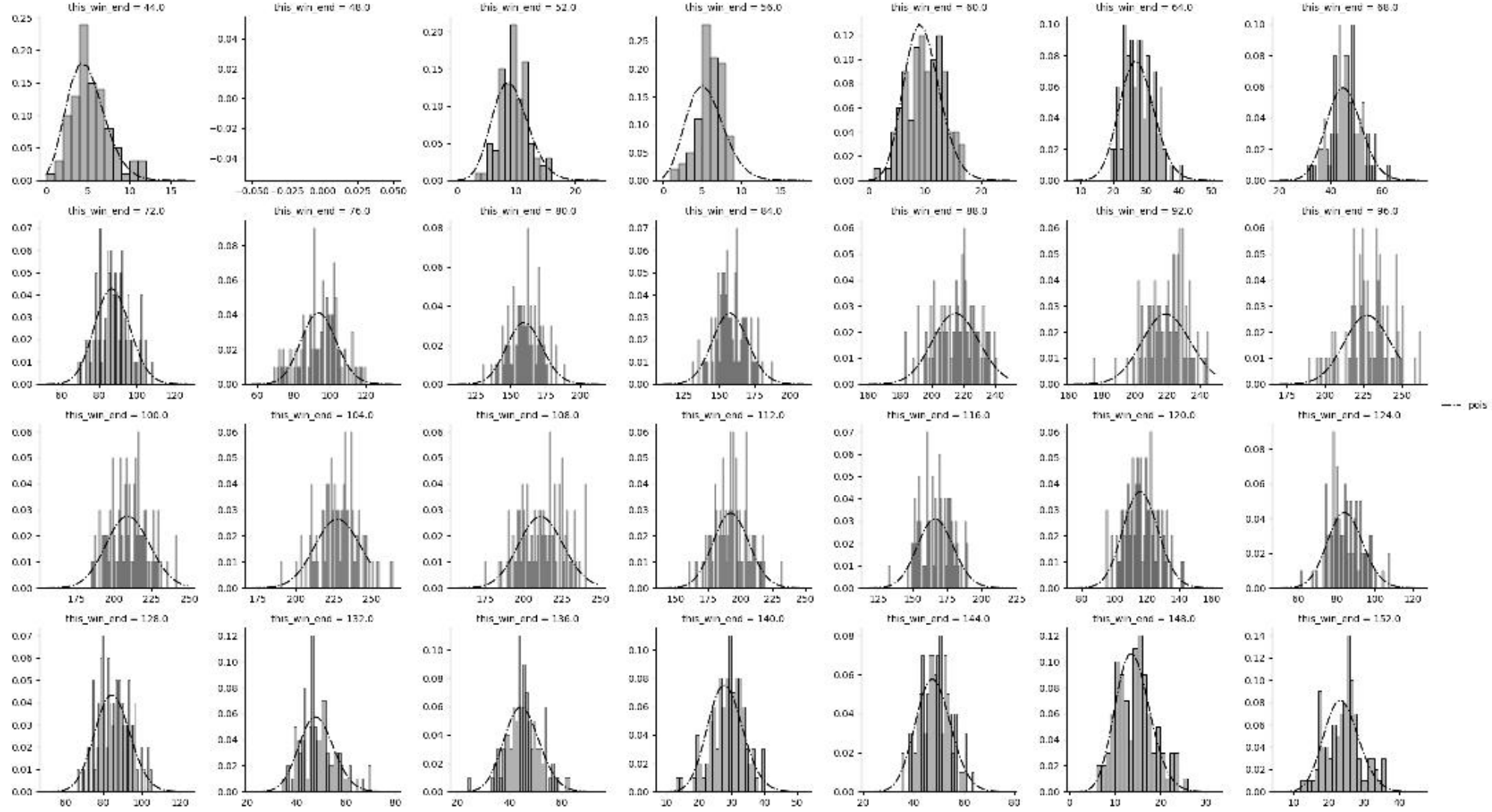

A
1e4
1.0
0.8
0.6
0.4
0.2
0.0
Number of infected individuals
R0 unchanged
R0 reduced to 1.1
R0 reduced to 0.75
0
24
48
72
96
120
144
168
192
B
200
150
100
50
0
Number of segregating sites
0
24
48
72
96
120
144
168
192
Days post index case


A
Number of sequences
Days post index case
B
Number of segregating sites
Days post index case
C
log-likelihood
R_0


A
Number of sequences
13.50
13.25
13.00
12.75
12.50
50
100
150
Days post index case
B
Number of segregating sites
60
40
20
0
50
100
150
Days post index case
C
log-likelihood
−120
−130
−140
−150
−160
−170
1.0
1.5
2.0
2.5
R_0


A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
Number of sequences
Number of segregating sites
log-likelihood
Days post index case


A
France (included)
France (excluded)
Other (excluded)
0
2
4
6
8
10
12
14
16
substitutions
B
2020-01-01
2020-01-15
2020-02-01
2020-02-15
2020-03-01
2020-03-15
date
C
24 subs/yr,
8.11e-04 subs/site/yr
distance to Wuhan/Hu-1
10
8
6
4
2020-02-20
2020-02-25
2020-03-01
2020-03-06
2020-03-11
2020-03-16


this_win_end = 44.0
this_win_end = 48.0
this_win_end = 52.0
this_win_end = 56.0
this_win_end = 60.0
this_win_end = 64.0
this_win_end = 68.0
this_win_end = 72.0
this_win_end = 76.0
this_win_end = 80.0
this_win_end = 84.0
this_win_end = 88.0
this_win_end = 92.0
this_win_end = 96.0
this_win_end = 100.0
this_win_end = 104.0
this_win_end = 108.0
this_win_end = 112.0
this_win_end = 116.0
this_win_end = 120.0
this_win_end = 124.0
this_win_end = 128.0
this_win_end = 132.0
this_win_end = 136.0
this_win_end = 140.0
this_win_end = 144.0
this_win_end = 148.0
this_win_end = 152.0
pols