Estimating the strength of selection for new SARS-CoV-2 variants

Christiaan H. van Dorp[1,*], Emma E. Goldberg[1,*], Nick Hengartner[1,2], Ruian Ke[1,2], and Ethan O. Romero-Severson[1,2,†]

[1]Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos NM, USA

[2]New Mexico Consortium, Los Alamos NM, USA

[*]These authors contributed equally to this work

[†]Correspondence: eoromero@lanl.gov

March 29, 2021

Abstract

A challenge to controlling the SARS-CoV-2 pandemic is the ability of the virus to adapt to its new human hosts, with novel and more transmissible strains of the virus being continually identified. Yet there are no generally accepted methods to consistently estimate the relative magnitude of the change in transmissibility of newly emerging variants. In this paper we consider three methods for examining and quantifying positive selection of new and emerging strains of SARS-CoV-2 over an existing wild-type strain. We consider replication at the level of countries and allow for the action of other processes that can change variants' frequencies, specifically migration and drift. We apply these methods to the D614G spike mutation and the variant designated B.1.1.7, in every country where there is sufficient sequence data. For each of D614G and B.1.1.7, we find evidence for strong selection (greater than 25% increased contagiousness) in more than half of countries analyzed. Our results also shows that the selective advantages of these strains are highly heterogeneous at the country level, suggesting the need for a truly global perspective on the molecular epidemiology of SARS-CoV-2.

Introduction

Recently, several genetic variants of SARS-CoV-2 have been identified that are either suspected or confirmed to have mutations that increase the contagiousness of the virus above the current circulating variants [1, 2, 3]. For a short while after the emergence of SARS-CoV-2, it was believed that the adaptive evolution of SARS-CoV-2 was limited, as evidence for purifying selection was found at most sites, with the clear exception of position 614 of the spike protein [4]. However, the emergence and rise of more complex variants such as B.1.1.7 in the United Kingdom (UK) has

shifted this understanding. As SARS-CoV-2 continues to adapt to transmission among humans, we can expect to see further mutations that alter the phenotype of the circulating virus [5]. Likewise, the gradual roll-out of vaccination programs globally will slowly change the immunological landscape, possibly leading to the emergence of escape strains that are partially or fully resistant to existing vaccinations [6, 7, 8].

Molecular epidemiology is a powerful framework comprised of a broad collection of methods and software designed to facilitate the analysis of pathogen genetic sequence data [1, 2, 3, 9]. These methods allow us to peer beyond what is provided by traditional epidemiological data such as case counts and death time-series, into the substructure of an epidemic by tracking the emergence and transmission of new genetic variants. Given the high extent of population mixing at both local and global levels, the time between the emergence of a new strain in one country and its global dissemination is short. Given that rapid spread, the ongoing fight against COVID-19 needs new, global tools focused on rapid modeling and assessment of the risk associated with new strains of SARS-CoV-2 to support global public health action.

Several groups have investigated the selective advantage of particular SARS-CoV-2 variants, such as D614G and B.1.1.7, either qualitatively or quantitatively. The global spread of the D614G variant was first described by Korber et al. [10]. Specifically for the UK, the selection coefficient for the D614G variant has been estimated using various phylogenetic and phylodynamic methods [3]. Estimates from these methods are highly variable, often producing inconclusive answers. The increased infectiousness of the D614G variant has also been functionally explained in terms of ACE2 receptor binding [11]. The selection coefficient of the B.1.1.7 variant has been estimated for England using a highly detailed deterministic epidemic model [12]. Phylodynamic approaches have led to similar results [2]. More worryingly, the B.1.1.7 variant is associated with increased mortality [13]. These changes to the SARS-CoV-2 phenotype embodied in D614G and B.1.1.7 likely represent only a small fraction of the phenotypic variability in the broader population.

In this paper we consider three different approaches for analyzing global sequence data to estimate the evidence for increased contagiousness of existing strains in the context of potentially high levels of between-country movement of people. Our goal is to test the robustness of results to different modeling assumptions, and to asses these different approaches as molecular surveillance tools.

Methods

We use three distinct methods to study the change in frequency over time of a SARS-CoV-2 genetic variant: isotonic regression, a population genetics model, and a stochastic epidemiological model. These models represent trade-offs in mechanistic detail and computational efficiency. The first takes a descriptive approach to the rise and fall of variant frequency based around rejecting a null hypothesis of limited or no change in frequency. The second incorporates the processes of selection and migration in the context of a deterministic theoretical population. The third additionally includes stochastic effects and more explicit epidemiological processes. By comparing results from these three models, we assess the robustness of our findings to the assumptions of particular methods. In all models, we compare a new variant with the circulating background variants. The new and background variants are labeled mt (for “mutant”) and

wt (for "wild type") respectively. All data and scripts can be downloaded from `https://github.com/eeg-lanl/sarscov2-selection`

Data

All epidemiological data was taken from the COVID-19 Data Repository curated by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [14], aggregated by week to reduce noise. The D614G and B.1.1.7 variant data were taken from the Los Alamos COVID-19 Viral Genome Analysis Pipeline [10, 15] and GISAID [16, 17], respectively.

Isotonic regression

The logic behind the isotonic regression method is that, if a variant is under selection strong enough to be worrying, then we should see a continual increase in its relative frequency. That is, for a variant under selection in a given country, we should be able to reject the hypothesis that it shows no increase with respect to its background.

Let us consider modeling the time series of pairs $(F_t^{\mathrm{wt}}, F_t^{\mathrm{mt}})$ that count the number of samples identified as the new variant (F_t^{mt}) and all other background sequences (F_t^{wt}) observed on a given day t. If we assume that the individuals whose SARS-CoV-2 virus is sequenced are randomly selected from the pool of infected individuals, then the number of observed variant sequences F_t^{mt}, conditional on the total number of sequenced individuals $F_t = F_t^{\mathrm{wt}} + F_t^{\mathrm{mt}}$, is binomial with probability p_t and size F_t. If the mutations that define the new variant (i.e., the genotype) are neutral, being neither beneficial nor deleterious, then the proportion p_t performs a random walk with constant expectation. However, if the variant has an evolutionary advantage, then the proportion p_t will have a non-decreasing expectation over time.

Here, we use that observation to devise a statistical test for the null hypothesis that a genotype is not advantageous. This approach does not, however, provide us with an estimate of how advantageous a genotype is because it does not model the competition between variants.

Let (t_i, V_i), $i = 1, \ldots, K$ denote the date and variant $V_i \in \{\mathrm{wt}, \mathrm{mt}\}$ from each of the K tested individuals. Our test is based on fitting isotonic logistic regressions to estimate a monotone non-decreasing probability p_t to that data, and use the logarithm of the likelihood ratio of the fitted isotonic model and model with constant p_t as the test statistic. Unlike in regular parametric cases, that statistic does not have an asymptotic chi-square distribution.

For that reason, we empirically evaluate the null distribution of the test statistic by refitting the isotonic regression to $(t_i, V_i^\star)$, where $V_1^\star, \ldots, V_K^\star$ is a random permutation of the original data $V_1, \ldots, V_K$. Fitting the isotonic logistic regression to M random permutations allows us to calculate empirically the country level p-value for the null hypothesis of no evolutionary advantage that are reported in Results. These were calculated in R 3.6.3 using the package `cgam` [18] to perform the isotonic logistic regression.

Population genetics model

The goal of this modeling approach is to provide a rapid means of estimating the selective advantage of a new genetic variant while also allowing for some contribution from migration. We first describe the model within each country. Then we explain how we fit it to data from all countries.

The model assumes that time is measured in discrete units of generations. Within each generation, we let selection act first and then migration. Say that p and q are the frequencies of new and background variants, respectively, at the beginning of the generation ($p+q=1$). Then say that p^* and q^* are the variant frequencies after selection, and p' and q' are the frequencies after migration and hence at the beginning of the next generation.

Selection

Define the absolute fitnesses of the two variants as $W_{\text{wt}}=\beta$ and $W_{\text{mt}}=\beta(1+s)$. So β is the geometric growth rate in number of infected people for the original genotype, and s is the selective advantage (if $s>0$) or disadvantage (if $s<0$) of the new variant.

For the moment, let us be explicit about population size. Define $N_{\text{mt}}=Np$ and $N_{\text{wt}}=Nq$ as the numbers of infected people with each variant at the beginning of this generation, where N is the total number of infected people in the population. After the selective event, which is transmission of each of the variants to new hosts, the numbers of infected people become

$$N^*_{\text{mt}}=W_{\text{mt}}N_{\text{mt}}=\beta(1+s)Np \tag{1a}$$

$$N^*_{\text{wt}}=W_{\text{wt}}N_{\text{wt}}=\beta Nq\,. \tag{1b}$$

Even if transmission (and recovery) alters the number of infected people, this change in population size does not affect the new variant frequencies, i.e.,

$$p^*=\frac{N^*_{\text{mt}}}{N^*_{\text{mt}}+N^*_{\text{wt}}}=\frac{\beta(1+s)Np}{\beta(1+s)Np+\beta N(1-p)}=\frac{(1+s)p}{1+sp} \tag{1c}$$

$$q^*=\frac{1-p}{1+sp} \tag{1d}$$

is independent of N and of β. So even with arbitrary changes in the number of infected people over time, this simple deterministic model can track only the variant frequencies. Of course, drift can have large effects when N is small, and also when a population of any size is growing rapidly. But we leave stochastic effects to our subsequent, more complex epidemic model.

Migration

Next, a fraction m of our population is replaced by immigrants. That is, some number of infected people leave our population, and an equal number of infected people arrive from elsewhere. We say that immigration is balanced by emigration because we are applying this same model to many populations (countries) simultaneously, and travel itself does not change the total number of infected people.

The change in frequency of the new variant due to migration is

$$p'=p^*(1-m)+\bar{p}m, \tag{2a}$$

where $\bar{p}$ is the frequency of the new variant among the immigrants. To be most generous to the alternative explanation that immigration is the driving force behind increases in p, we set $\bar{p} = 1$ so

$$p' = p^* + (1 - p^*)m\,. \tag{2b}$$

Note that if the number of infected people is increasing over time ($\beta > 1$ in the description of selection, above), our formulation with constant migration fraction m means that the number of infected travelers is also increasing over time.

Putting together the total effects of selection and migration for this generation, by substituting Eq. (1c) into Eq. (2b),

$$p' = \frac{(1+s)p + (1-p)m}{1+sp}\,. \tag{3}$$

At any time t,

$$p_t = \frac{[s(1+s)^t + m(1-m)^t]p_0 + m[(1+s)^t - (1-m)^t]}{s[(1+s)^t - (1-m)^t]p_0 + [m(1-s)^t + s(1-m)^t]} \tag{4}$$

(see Eq. (A-6)). We define $t = 0$ as the time at which the new variant first appears in any country. Notice that without migration ($m = 0$), Eq. (4) reduces to the logistic model derived in [19].

Fitting to data

For each country, the data we use are the numbers of observations of the background (F_t^{wt}) and the new variant (F_t^{mt}) each day (t). We fit these data with Bayesian binomial regression, using Eq. (4), with country as a random effect. This yields separate estimates of s, m, and p_0 for each country.

Because the time unit for our data is days, the estimates of s and m from the model fit must be multiplied by the generation time in order to be interpreted as per-generation processes. The mean serial interval for SARS-CoV-2 is most likely between 4 and 7.8 days [20], so we use a normal distribution with mean 5.9 and standard deviation 1.15 for the mean generation time.

When selection truly favors a variant due to its genetic composition, it should have a similar advantage in any country. There may be differences from country to country, though, due to chance effects. For example, if the early-infected people happen to be from a demographic with higher transmission or a city with looser enforcement of social distancing, selection may appear to be stronger. We therefore use a hierarchical model in which s is drawn for each country from a normal distribution, whose mean and variance we estimate in order to infer the consistency of selection.

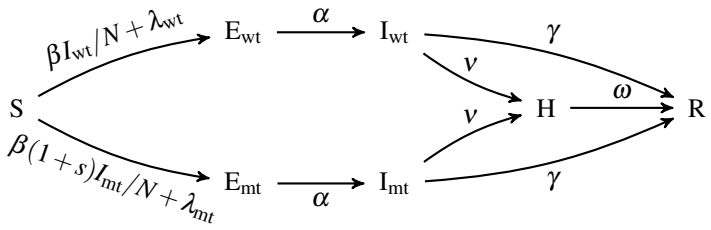
The migration rate, m, is the proportion of the country's population swapped out for the new variant each generation. This is surely quite small, especially considering travel restrictions. We therefore set an exponential prior on m with mean 0.001.

For numerical stability, we transform all frequencies to the logit scale (see Appendix A). Models were fit with Stan [21], using 4 parallel chains of length 2000, with a warm-up phase of a 1000 iterations.

Stochastic epidemic model

To take more detailed population dynamics into account, we use a stochastic compartmental Kermack-McKendrick-type model. In addition to susceptible (S), exposed (E), infectious (I), and removed (R) individuals, we keep track of individuals with severe disease (H), and stratify the exposed and infected populations into individuals infected with the background (wt) or the new variant (mt). The compartment of severe infections is used to model the observed delay between infection and death. The two strata are used to keep track of the new variant's frequency in the population, and model its selective advantage (s).

The compartmental model keeps track of the number of individuals S, E, I, H, R in the disease states S, E, I, H, R, respectively. An individual starts out susceptible, and upon infection enters the exposed compartment and then becomes infectious at rate α. An infectious individual can either become severely infected at rate ν, or recover at rate γ. Severely infected individuals either recover or die at rate ω. The total population size is denoted by N, and we write $X = (S, E_{\mathrm{wt}}, E_{\mathrm{mt}}, I_{\mathrm{wt}}, I_{\mathrm{mt}}, H, R)$. The transition rates $\eta_j(X,t)$ between the compartments are indicated by the following diagram, and the parameters are listed in Table 1.



(5)

Here the indicated rates are *per capita* and should be multiplied by the size of the source compartment (e.g., $\eta_{\mathrm{H}\to\mathrm{R}}(X,t) = H\omega$). The selective advantage of the new variant is equal to s; when $s > 0$, the mutant has a higher infection rate $\beta(1+s)$ than to the wild-type (β). The other life-history traits of the virus are assumed to be identical between wild-type and mutant. To model the effects of non-pharmaceutical interventions (NPI) such as lock-downs, the infection rate β is a smoothed, piece-wise constant function of time [22]. To account for migration, we added time-dependent terms λ_{wt} and λ_{mt} to the *per-capita* infection rate, representing the exposure of individuals in the population to SARS-CoV-2 from other regions. The precise definitions of the time-dependent parameters β, λ_{wt}, and λ_{mt} are given in Appendix B.

Observation model

The model is fit to two data streams. The first data stream consists of weekly incidence of COVID-19 deaths D. For this reason, we keep track of an auxiliary accumulator variable Θ^{HR}, which keeps track of all transitions from H to R within a week. After each time the incidence is observed, the accumulator variable Θ^{HR} is set to 0. Let δ denote the probability that a severe infection leads to death and not recovery. To account for variability in δ between demographic groups or reporting errors, we use an over-dispersed negative binomial instead of a binomial or Poisson

symbol	unit	value				description	source
		D614G		B.1.1.7		*variant*	
		UK	NL	UK	NL	*region*	
s	-	0.28	0.27	0.55	0.28	selective advantage novel variant	est.[a]
β_0	d^{-1}	0.81	0.80	0.40	0.39	infection rate before lockdown	est.
β_1	d^{-1}	0.15	0.12	0.25	0.22	infection rate during lockdown	est.
β_2	d^{-1}	0.21	0.33	0.34	0.37	infection rate after lockdown relaxation	est.
β_3	d^{-1}	-	-	0.18	0.20	infection rate after second lockdown	est.
t_0	d	54	54	243	243	initial time	-
t_1	d	86.6	82.5	305	298	time of lockdown	est.
t_2	d	188	181	337	336	time of lockdown relaxation	est.
t_3	d	-	-	369	365	time of second lockdown	est.
p_0	-	0.28	0.37	$3.0 \cdot 10^{-4}$	-	initial mutant frequency	est.[b]
λ_0	d^{-1}	-	-	-	$5.7 \cdot 10^{-3}$	*per capita* infection rate due to travel	est.[b]
r_D	-	105	115	90.2	93.0	overdispersion parameter death incidence	est.
r_F	-	72.8	68.3	70.3	71.6	overdispersion parameter sequence data	est.
τ	d^{-1}	0.011	0.018	0.021	0.013	overdispersion of the process noise	est.
ζ	-	$8.2 \cdot 10^{-6}$	$1.7 \cdot 10^{-5}$	$3.3 \cdot 10^{-4}$	$3.1 \cdot 10^{-4}$	initital fraction infected	est.
ξ	-	-	-	0.06	0.05	fraction of the population immune at time t_0	[23, 24][c]
N	-	66.5	17.4	66.5	17.4	population size (million)	-
$1/\alpha$	d	3	.	.	.	mean duration of incubation period	[25, 26]
γ	d^{-1}	$1/4$	.	.	.	recovery rate from infectious stage	[27, 28][d]
ν	d^{-1}	$\gamma/50$	.	.	.	rate of developing severe infection	[29, 30][e]
$1/\omega$	d	12.5	.	.	.	duration of severe infection	[25, 26][f]
δ	-	0.3	.	.	.	probability of dying from severe infection	[29, 30, 31][e]

Table 1: Parameters for the epidemic model for the United Kingdom (UK) and the Netherlands (NL) and the D614G and B.1.1.7 variants. Notes: [a]Estimated. [b]For the Dutch B.1.1.7 model, the initial frequency is fixed to 0. Instead, the variant is introduced due to travel from the UK. [c]More recent estimates for sero-prevalence in the Netherlands are taken from https://www.rivm.nl/pienter-corona-studie/resultaten. [d]The generation interval in the SEIHR model with exponentially distributed transition times is equal to $1/\alpha + (\gamma+\nu)^{-1} \approx 1/\alpha + 1/\gamma$. Hence, with an average incubation period of 3 days, we need an average infectious period of 4 days to get an average generation time of 7 days. [e]By taking the probability of developing severe infection equal to 0.02 and the probability of dying from severe infection equal to 0.3, we arrive at a case fatality rate of 0.6%. Our choice is also comparable to mortality rates for ICU patients [31]. [f]The average time between symptom onset and death is 16.5 days. After subtracting the duration of the infections period $(\gamma+\nu)^{-1} \approx 1/\gamma$, we get an average duration of severe infection of 12.5 days.

likelihood function for the observed death counts. At the time of the n-th observation t_n, we then get

$$D_n \sim \text{NegBinom}\left(\delta \cdot \Theta^{\text{HR}}(t_n), r_D\right) \tag{6}$$

where the parameterization of the NegBinom(ℓ, r) distribution is such that it has mean ℓ and variance $\ell + \ell^2/r$.

The second data stream consists of the number viral samples F that were sequenced each week, and the number sequences F^{mt} identified as the new variant. We assume that these sequences are collected from individuals that transition from the exposed to the infectious compartment, and hence we again define accumulator variables $\Theta^{\text{EI}}_{\text{wt}}$ and $\Theta^{\text{EI}}_{\text{mt}}$ to keep track of such transitions (for wild-type and mutant infections, respectively) during the week between subsequent observation times. We define $f_{\text{mt}} = \Theta^{\text{EI}}_{\text{mt}}/(\Theta^{\text{EI}}_{\text{wt}} + \Theta^{\text{EI}}_{\text{mt}})$ for the fraction of individuals that were infected with the new variant. To allow for over-dispersion of sampling, we use a beta-binomial likelihood function:

$$F^{\text{mt}}_n \sim \text{BetaBinom}\left(F_n, f_{\text{mt}}(t_n) r_F, (1 - f_{\text{mt}}(t_n)) r_F\right) \tag{7}$$

where the parameter r_F determines the level of over-dispersion of the sampling process.

We fit the model to the two data streams using sequential Monte-Carlo (SMC), where parameters are estimated with iterated particle filtering as described in [32]. The details of the procedure are given in Appendix B.

Diffusion approximation of the epidemic model

Exact simulation of the Markov jump process (MJP) that defines our stochastic epidemic model is very computationally intensive. We therefore switch to a diffusion approximation of the MJP when population sizes become large in order to do inference more efficiently. This formalism allows us to incorporate two sources of noise. The first being the process noise inherent to the MJP, which becomes negligible when the sizes of the compartments are large. We therefore introduce a second noise term that captures other origins of stochasticity that the MJP can not account for and acts on predominantly large population sizes.

As above, we denote the state of the n-dimensional model (where $n = 7$) by $X^i(t)$ with $i = 1, \ldots, n$. The discrete, stochastic model is defined by $k = 9$ state transitions

$$X \xrightarrow{\eta_j(X,t)} X + \varepsilon_j\,, \quad j = 1, \ldots, k \tag{8}$$

where $\varepsilon_j \in \mathbb{Z}^n$ is the increment of the j-th transition. For instance, the transition H $\to$ R corresponds to the increment $(0, \ldots, 0, -1, 1)$. Using the Kramers-Moyal expansion of the master equation, the MJP is mapped to a system of stochastic differential equations (SDE) that can be derived from the transitions η_j and increments ε_j as follows [33]

$$\mathrm{d}X^i = \sum_{j=1}^{k} \varepsilon^i_j \eta_j(X,t)\mathrm{d}t + \sum_{j=1}^{k} \varepsilon^i_j \sqrt{\eta_j(X,t)}\mathrm{d}B^j_t\,, \quad i = 1, \ldots, n \tag{9}$$

where B_t is a 9-dimensional Brownian motion, corresponding to the 9 transitions of the MJP in Eq. (5). The SDE in Eq. (9) is of the form $\mathrm{d}X = \mu(X,t)\mathrm{d}t + \sigma(X,t)\mathrm{d}B_t$, where μ and σ describe the drift and volatility, respectively. The

volatility matrix $\sigma(X,t)$ encodes the intrinsic noise of the MJP, which is negligible compared to X when X is large. We therefore add a small second noise term to the system of SDEs that is proportional to X. After this adjustment, the SDE becomes

$$\mathrm{d}X^i = \mu^i(X,t)\mathrm{d}t + \sigma^i(X,t)\mathrm{d}B_t + \tau X^i \mathrm{d}\widetilde{B}_t^i, \quad i = 1,\ldots,n \tag{10}$$

where $\widetilde{B}_t$ is a n-dimensional Brownian motion, independent of B_t. The parameter $\tau \ll 1$ determines the magnitude of the additional noise term.

In Appendix B, we further describe in detail the algorithm used to switch from a discrete (MJP) to a continuous (SDE) model, and the way the initial condition of the system is determined.

Results

Isotonic regression results

Our first approach, isotonic regression, provides a simple statistical test for relative increase of a new variant. The null hypothesis is that the daily proportion of SARS-CoV-2 cases that are the mutant variant does not increase over time. The isotonic regression method rejected this null hypothesis in 33 of 38 countries with sufficient data for D614G and all 24 of the countries with sufficient data for B.1.1.7 at the 5% significance level. The results are shown as a map in Fig. 1CD. In general the isotonic regression method identifies the same set of countries as the population genetic method (see below). However the computed p-value is not directly translated into an estimate for the strength of selection that is estimated from the other methods.

Population genetic model results

Our hierarchical population genetics model estimates a normal distribution from which is drawn the selection coefficient for each country. The mean of that global distribution for s is 0.25 (90% credible interval (CrI): $[0.14, 0.35]$) for D614G and 0.26 (90% CrI: $[0.14, 0.36]$) for B.1.1.7, both confidently greater than zero (Fig. 2AD). After allowing for variance around that mean, however (Fig. 2BE), the overall global distribution of s is substantially wider, both for D614G (90% CrI: $[-0.16, 0.68]$, Fig. 2C) and for B.1.1.7 (90% CrI: $[-0.07, 0.60]$, Fig. 2F). Uncertainty in our estimates of global s come from noise in the data, uncertainty in the mean generation time, and heterogeneity among countries.

The overall selective advantages of these two variants are estimated to be very similar, but that does not mean that they are equally transmissible. The strength of selection for a variant is measured relative to all the other genotypes present over that time frame. Because B.1.1.7 emerged after D614G became globally common, the absolute fitness of B.1.1.7 is likely greater.

Estimates of the selection strength for each variant in each country are shown in Fig. 3. Each variant shows much heterogeneity among countries. Our model allows for a random component in country-to-country variation in s, but the differences in estimated s among the countries likely overstates the differences in actual transmission advantage. Each country surely experiences many processes that are not included in our simple model—such as superspreader events, nonrandom sampling, or waves of travellers arriving from places with different variant frequencies—and all of

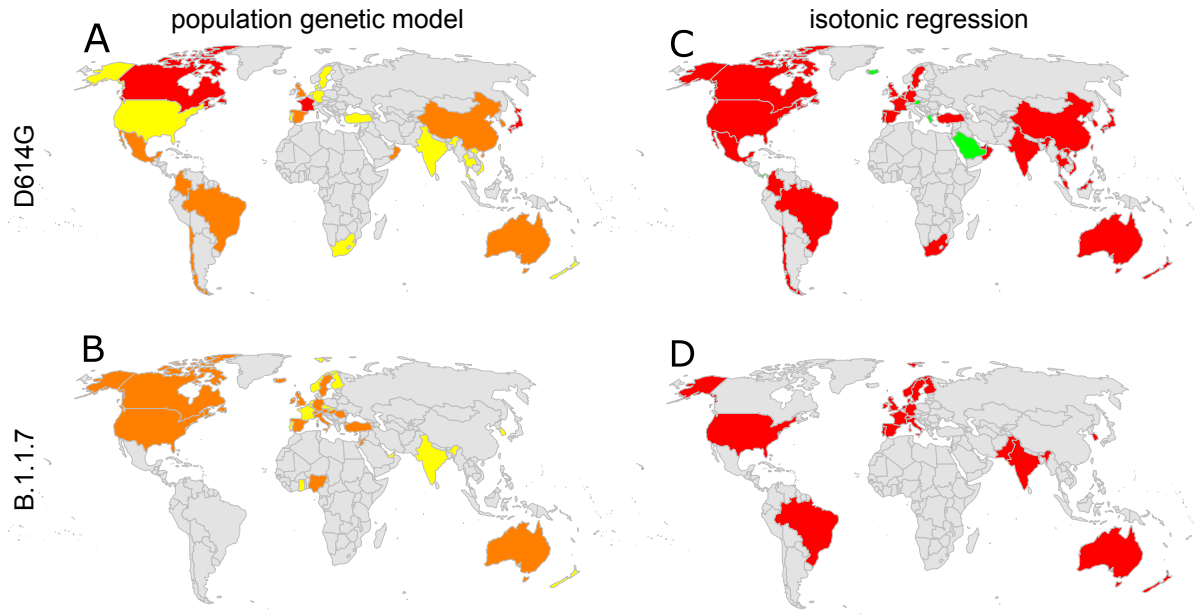


Figure 1: Map of the results of the population genetic model and isotonic regression method. Panels A and B show the estimated selection strength for D614G and B.1.1.7 respectively. Yellow, orange, and red indicated selection coefficient less than 0.3, between 0.3 and 0.6, and above 0.6 respectively. Panels C and D show the p-values from the isotonic regression method for D614G and B.1.1.7 respectively. Red and green show p-values below and above 0.05 respectively.

this heterogeneity is bundled by the model into differences in s (and m, which is constrained to be small). Furthermore, strong selection for one variant in one country does not necessarily correspond to strong selection for the other variant in that country (Fig. S2), suggesting that factors beyond country-level covariates underlie the overall heterogeneity.

In estimating the selective advantage of each variant, our model allows for a contribution of migration in elevating the variant frequencies (see Methods). Because selection and migration to some extent provide alternative explanations for change in variant frequency, we find some negative correlation between these two processes (Figs. S3 and S4). The estimates of migration are not particularly distinguishable among countries (Fig. S1), but estimates of selection nevertheless show clear differences among countries (Fig. 3). We thus conclude that the selective effect of a variant can be estimated even allowing for a reasonable amount migration.

Fits of the population genetic model to each country are shown in Figs. S5 and S6 for D614G and B.1.1.7, respectively. The countries differ dramatically in sampling effort and data availability over time. The robustness of the population genetic model is tested with a more detailed epidemic model below. For this analysis, we focused on two countries—the United Kingdom and the Netherlands—for which a lot of data is available. Our focal countries show different stages of the variant trajectories, but in both the model fits show relatively narrow credible intervals.

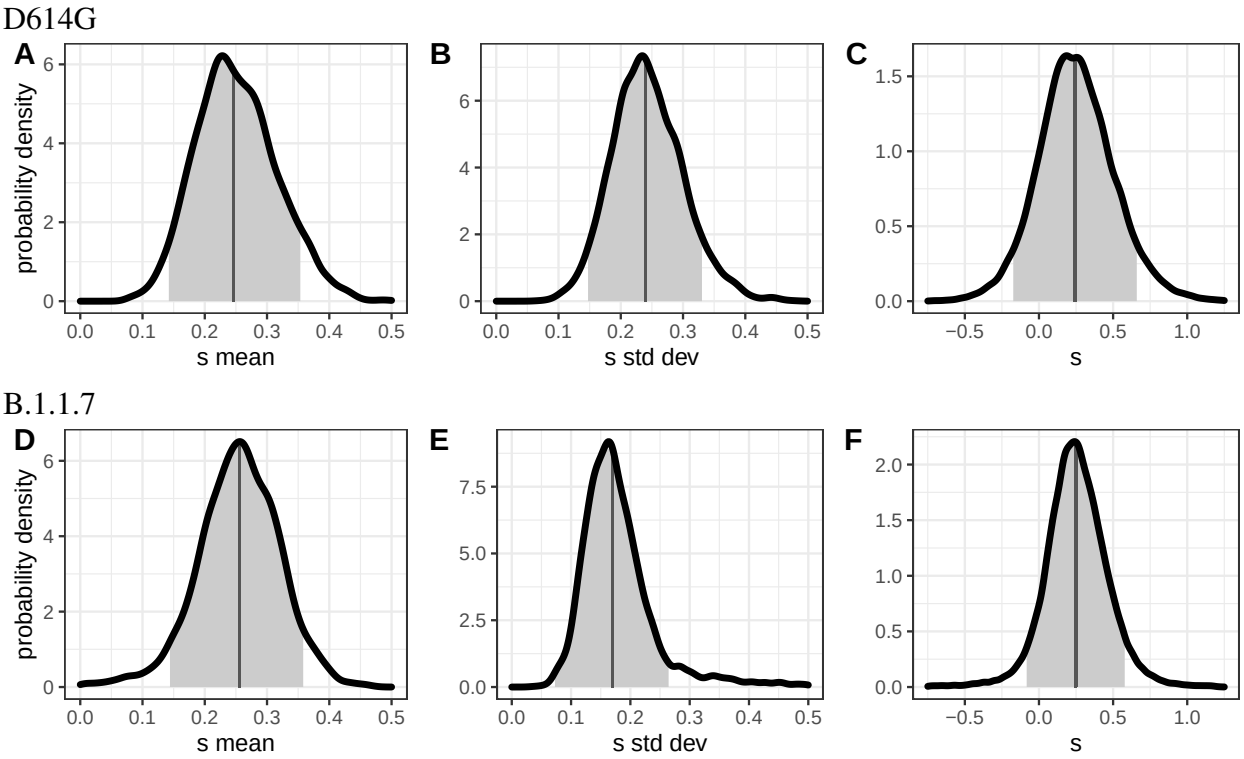


Figure 2: Estimated global distribution of selection coefficients for the D614G and B.1.1.7 variants from the population genetics model. The hierarchical model estimates the mean (left column) and standard deviation (middle column) of a normal distribution from which the selection coefficient, s, of each country is drawn. To compose the overall (posterior predictive) distribution of s (right column), we sampled many values of the mean and standard deviation from the posterior distribution of the model, and for each drew many samples from the normal distribution it defines. Thus, panels A and D provide the estimated mean selection coefficient across countries, while panels C and F provide the overall distribution for each country's selection coefficient. In each panel, dark vertical lines mark the median, and the 90% CrIs are shaded.

Stochastic model results

The overall stochastic model fits to D614G and B.1.1.7 for the UK and Netherlands are shown in Fig. 4 and 5 respectively. The models provide very good fits to the data in both cases matching both the death time series and the change in proportions of the mutant variant in both countries. D614G shows a very similar pattern in the UK and Netherlands where the mutant was spreading in a way that is nearly indistinguishable from the wild-type strains for a period of several weeks in the early epidemic period. However, in both countries the model predicts that the mutant strain very quickly outpaces the wild-type strains and continues to become more relatively prevalent even when the overall prevalence is declining by orders of magnitude.

The dynamics of B.1.1.7 in the UK and Netherlands are substantially different from both D614G and each other. In both cases the mutant strain is much slower to rise, occurring over a period of months rather than weeks

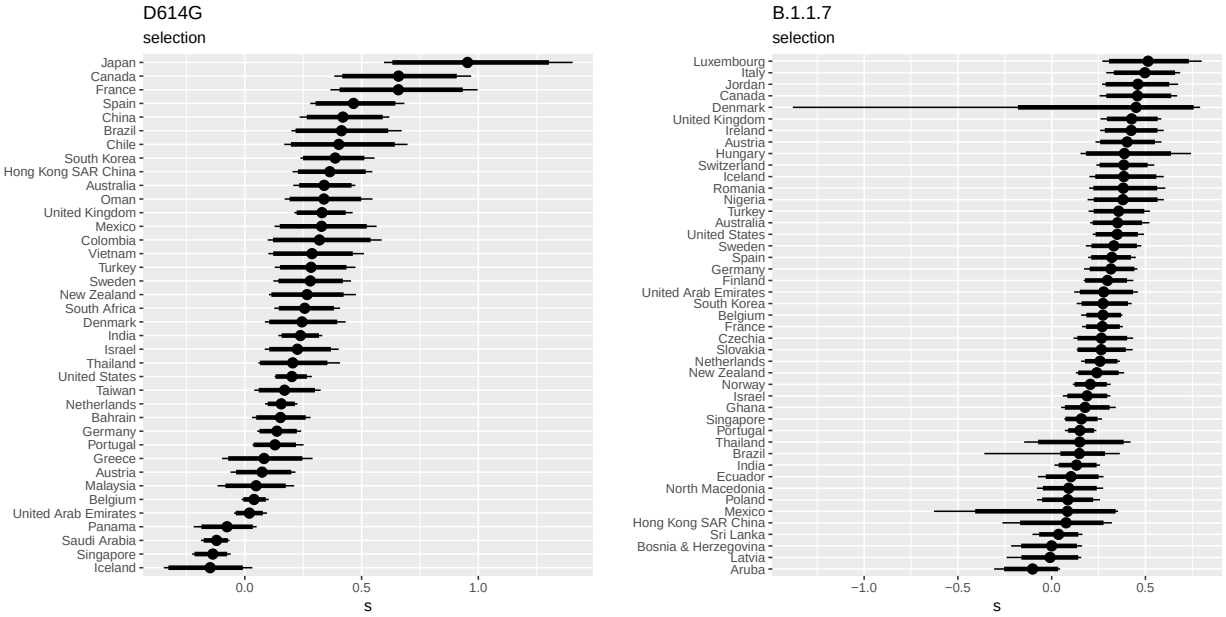


Figure 3: Selection coefficients for each country, from the population genetic model. Points mark the median, and thick and thin lines are 90% and 95% CrIs, respectively. Corresponding estimates of migration are in Fig. S1.

as was the case with D614G. In both countries, the mutant strain rises exponentially despite the large changes in the overall prevalence of COVID-19 due to changing policies and behaviors during this time. Specifically in the UK, the rise in death incidence after December 14 2020 is preceded by the rapid increase of the B.1.1.7 strain, both in frequency and absolute numbers (Fig 5). This suggests that the interventions ongoing in the UK were sufficient to bring the background strains below threshold but not B.1.1.7.

To further substantiate this, we calculated the instantaneous effective reproduction number (R_e) using the inferred trajectories of the stochastic model (Fig S7). The effective reproduction number of the wild-type fluctuates abound the threshold value 1 between November and December, following the increased NPI (non-pharmaceutical intervention) initiated end October [34]. As the B.1.1.7 variant has a $\sim 50\%$ higher reproduction number, these NPI were not sufficient for controlling the growth of the variant, leading to a doubling of the death incidence in January 2021 and the necessity of further stringent restrictions. This suggests that new variants with an increased fitness are particularly dangerous when in-place NPI are only resulting in a marginal control of the epidemic.

As the B.1.1.7 variant was most likely introduced to the Netherlands from the UK, we incorporated external forces of infection in the Netherlands (λ_{wt} and λ_{mt} in Eq. (5)) to account for this fact (see Appendix B). This process allows a source of infection in the Netherlands, governed by rate λ_0 (Table 1), that is proportional to the prevalence of B.1.1.7 in the UK. We forced the migration process to zero after December 21, 2020 to account for travel restrictions from the UK to the Netherlands. Based on a sample of 100 reconstructed trajectories of the stochastic model (Fig. 5B, green curves), we estimate that at that time of the travel restriction being in place, 6694–10287 individuals were infected with the B.1.1.7 variant in the Netherlands. That is, the model suggests that the establishment of B.1.1.7 in

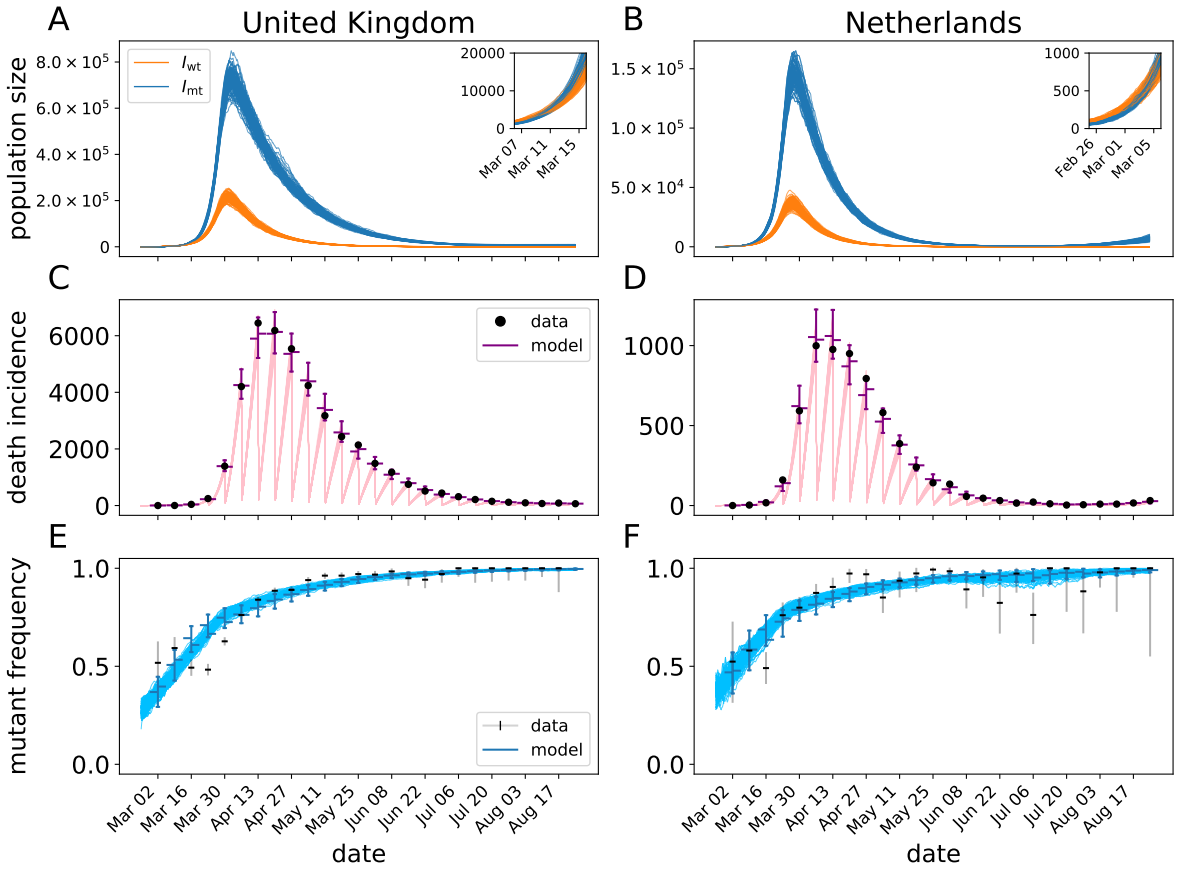


Figure 4: Mechanistic model fit to D614G data in the Netherlands and UK. Panels A and B show many realizations of the prevalence of the background, I_{wt}, and D614G variant, I_{mt} for the maximum likelihood model fit. Panels C and D show the number of deaths accumulated up to the week scale (black dots) and the model fits to those data. The bars around the points indicate the 95% predictive interval for the data according to the model. Panels E and F show the proportion of sequenced genomes with a glycine on position 614 of the spike protein in a given week. The blue lines are realizations of the model fit to the data. Vertical bars on the data indicate the 95% confidence intervals (CI) for the proportion based on the number of sampled genomes.

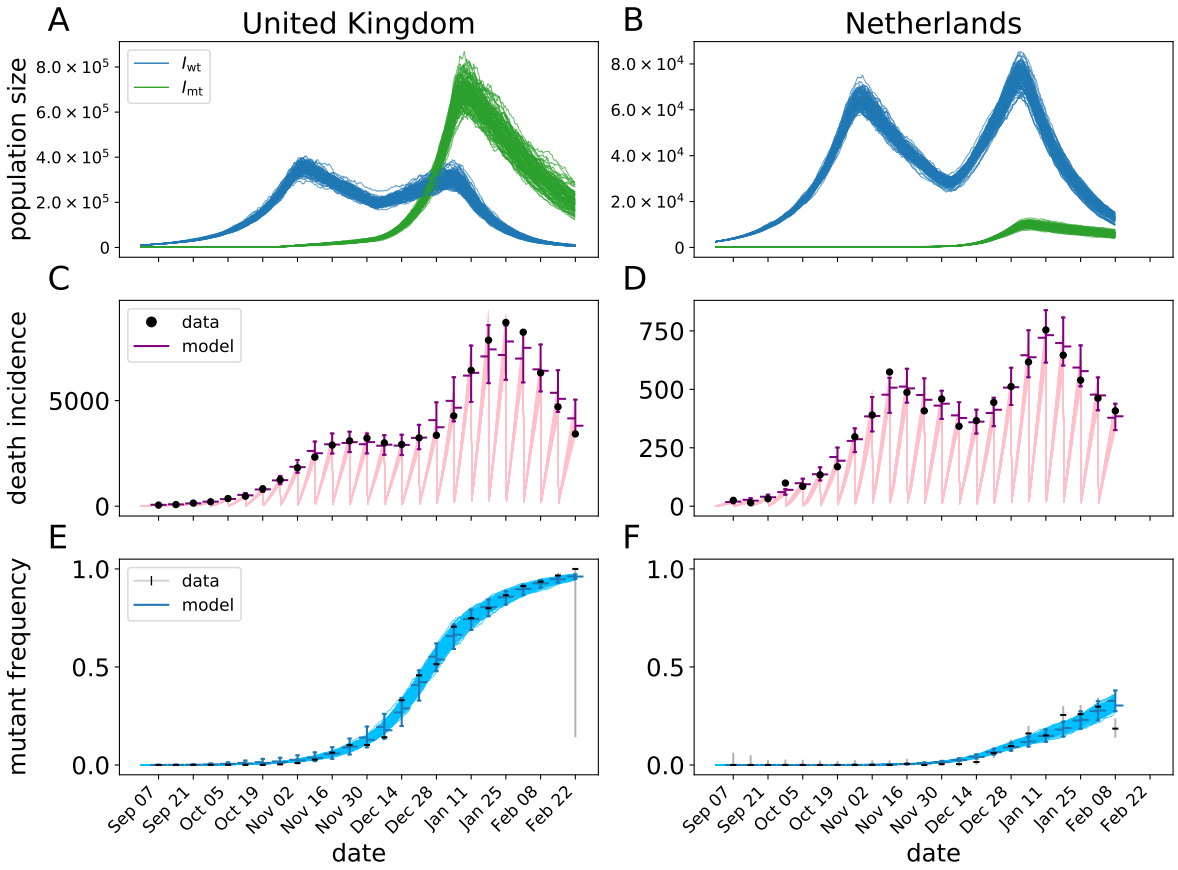


Figure 5: Mechanistic model fit to B.1.1.7 data in the Netherlands and UK. Panels A and B show many realizations of the prevalence of the background, I_{wt}, and B.1.1.7 variant, I_{mt} for the maximum likelihood model fit. Panels C and D show the number of deaths accumulated up to the week scale (black dots) and the model fits to those data. The bars around the points indicated the 95% predictive internal for the data according to the model. Panels E and F show the proportion of sequenced genomes classified as the B.1.1.7 variant in a given week. The blue lines are realizations of the model fit to the data. Vertical bars on the data indicate the 95% CI for the proportion based on the number of sampled genomes.

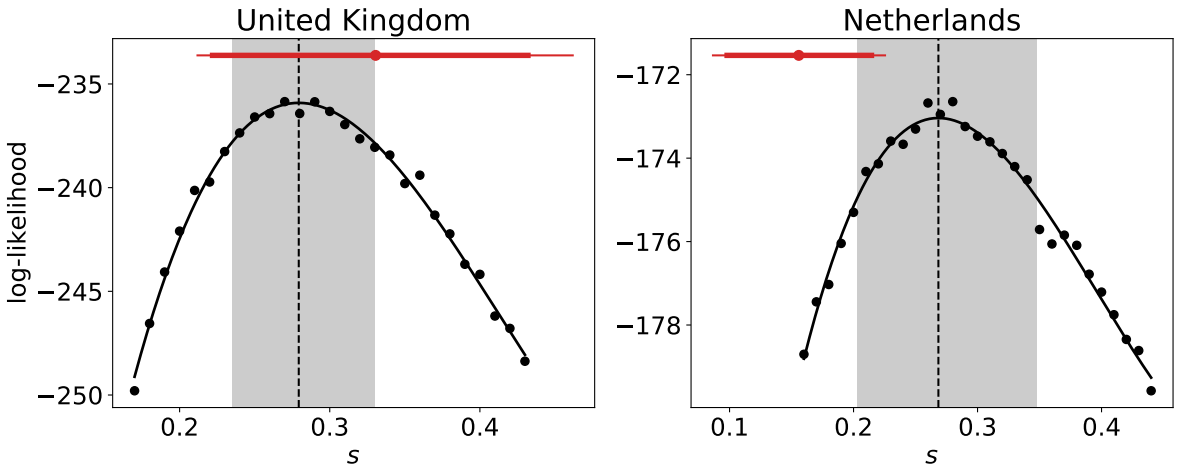


Figure 6: Profile likelihood for the selection effect of D614G in the UK and Netherlands. The maximum likelihood estimate (MLE) is indicated by a dashed line and the 95%CI is indicated by a grey box. For the UK, we estimated a selection coefficient $s = 0.28$ (95% CI: $[0.23, 0.33]$), and for the Netherlands we estimated $s = 0.27$ (95% CI: $[0.20, 0.35]$). The red box plots show the estimates from the population genetics model (Fig. 3).

the Netherlands was facilitated by migration from the UK, however, the major factor in the spread of B.1.1.7 in the Netherlands is its selective advantage s.

The profile likelihoods of s, the increase in contagiousness for the mutant variant, for the D614G and B.1.1.7 are shown in Fig. 6 and 7. The profile likelihoods show well constrained estimates of s. In general the confidence intervals in the mechanistic model are comparable with or more narrow than the population genetic model, which is consistent with the model being both more constrained and using more data. In addition, the population genetics model takes the uncertainty in the average length of the generation interval into account. The point estimates of s from both models are close to one another with the largest difference being about 0.12 for B.1.1.7 in the UK. Point estimates of the other model parameters are listed in Table 1. The fact that the population genetics model results in smaller estimates of s in some cases can be understand by looking at the population dynamics. The emergence of the variant is in all cases followed by increased NPI, leading to a decrease in the number of infected individuals, and also the relative growth rate of the variant. This is interpreted by the population genetics model as a smaller selection coefficient.

Discussion

We have illustrated three different approaches to measuring selection effects from the global SARS-CoV-2 genetic sequence data. Our analyses all point to very strong but heterogeneous selective advantages for the D614G and B.1.1.7 variants at the country level, even allowing for both migration and drift. It is important to note that our methods look for an advantage of a given variant over whatever variants are circulating at the time, rather than against a fixed reference strain. This means that the estimates of a fitness advantage of B.1.1.7 in many countries is relative to a background that consists mostly of D614G. Thus, the fitness of B.1.1.7 exceeds that of D614G, which itself exceeds

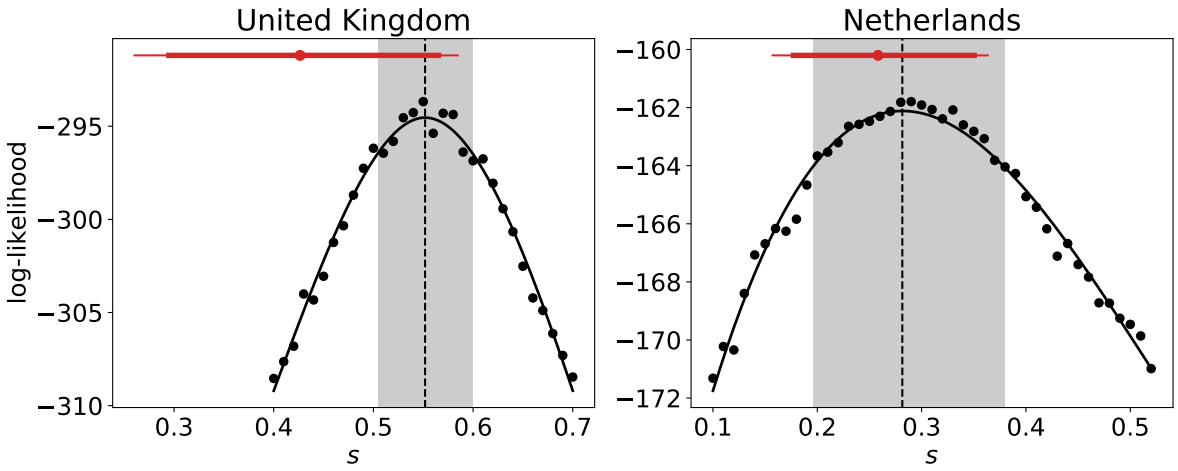


Figure 7: Profile likelihood for the selection effect of B.1.1.7 in the UK and Netherlands. The MLE is indicated by a dashed line and the 95%CI is indicated by a grey box. For the UK, we estimated a selection coefficient $s = 0.55$ (95% CI: $[0.50, 0.60]$), and for the Netherlands we estimated $s = 0.28$ (95% CI: $[0.20, 0.38]$). The red box plots show the estimates from the population genetics model (Fig. 3).

the original genotype. Our stochastic epidemiologic model suggests that the level of transmissibility of currently circulating variants is sufficient to reduce the qualitative effect of non-pharmaceutical interventions (NPIs) in Europe.

SARS-CoV-2 is rapidly adapting to its new human hosts, and strains with elevated contagiousness will likely continue to emerge as the virus continues to adapt. This situation is challenging as elevated contagiousness narrows the range under which vaccination programs can eliminate the virus, and it also opens up the possibility of escape mutations allowing infection among vaccinated persons. Integrating molecular epidemiology surveillance into SARS-CoV-2 pipelines is essential for not only monitoring the emergence of new strains, but for establishing an early warning system to monitor for escape mutations in the era of vaccine roll-out. Each approach that we examined has its own strengths and weaknesses for how it can fit into an expanded molecular epidemiology surveillance system.

The isotonic regression method is easy to compute and based on the very straightforward premise that a consistent selective advantage should produce a continually increasing frequency of the new variant in all countries where it has been observed. However, because the method is based on a hypothesis-testing framework, there is no way to determine the strength of selection relative to the background strains. Likewise, the method could be misleading in a context where the genetic background is rapidly changing (e.g., other strains under positive selection are introduced into the population during the study period). We believe that a regression-based approach is, nevertheless, very useful for rapidly evaluating evidence of selection potentially in large-scale molecular surveillance pipelines.

The population genetic model is more mechanistically explicit than the regression approach and, therefore, gives a direct estimate of the selection effect. The model also allowed us to integrate a simple migration process and to jointly estimate the parameters of selection and migration. The population genetic model is also simple enough that it was coded in a popular statistical language and fit to the global data in a matter of hours on a standard laptop

computer. Its framework to estimate country-level selection effects shaped by an overall global distribution makes it potentially quite useful for general molecular surveillance purposes. However the primary weakness of the population genetic model is that it does not account for random fluctuations in the underlying populations, which, given that the point of estimating selection effects in near real-time is to give warning before the new variant becomes widespread, is potentially a problem.

The stochastic mechanistic model solves this problem by explicitly modeling stochastic effects that could produce changes in variant frequency by chance alone, in addition to selection and migration. Our approach takes this idea one step further by allowing noise above what would be expected in a typical homogeneous stochastic model; our logic behind this choice is that the epidemiological model assumes homogeneous mixing at the state level, which we know to be unrealistic. In reality, transmission is occurring at much smaller local scales that can lead to sudden jumps in both the number of cases and number of observed variants; allowing for extra noise in the stochastic model makes the method less sensitive to mis-specifications such as an over-simplified population structure. Despite being more complex and using additional data, we found that the mechanistic model was in agreement with the population genetic model, suggesting that the population genetic model is a reasonable balance between computability and accuracy.

All of our models (and most of the other published models) make the assumption that genomes are selected at random from the set of all possible cases. If, for example, samples were sequenced specifically because they were in contact with someone that was known to be infected by the variant under study, the data may be biased toward over-estimating the spread and hence selective advantage of the new variant. There is almost certainly some bias from the non-random processes by which samples are obtained and sequenced; however, we believe that our results are still overall valid for three reasons. First, it is unlikely that the same level of bias from non-random sampling would occur in each country to produce a similar pattern in each country; that is, countries represent semi-independent systems. Second, the evidence for selection effects includes parts of the time series before people were concerned about the spread of new variants, and, therefore, were unlikely to preferentially sequence the new variants. Third, the UK has put effort into developing a representative sample of SARS-CoV-2 genomes in their country and the estimates for the selection effects in the UK for D614G and B.1.1.7 are very close and slightly above the population average for these lineages.

Several studies using other methods have similarly found D614G and B.1.1.7 to each have a selective advantage over other variants circulating contemporaneously [10, 19, 3, 12, 2]. In contrast, however, van Dorp et al. [35] found no support for a selective advantage of any of the variants they tested, including D614G. We suspect this is because their statistical test required the repeated emergence of a variant in order to draw any power. Although phylogenetic replication is an appropriate requirement in many situations, it is too conservative for identifying variants of concern on the timescale at which they emerge. Instead, to test for a selective advantage of variants that have arisen only once, power can be obtained from fitting explicitly epidemiological models within one location (e.g., our stochastic model, and others [19, 3, 12, 2]) or looking for consistent effects in multiple locations with largely-distinct conditions (e.g., our isotonic regression and population genetic models, and Korber et al. [10]).

A central question in any modelling endeavor is how much detail is required to accurately address the problem

in question. In the last year, a large number of models have been developed to study various aspects of the SARS-CoV-2 pandemic, ranging from very simple [25] to extremely detailed [12]. For the purpose of estimating selective advantages of variants, we argued that (relatively) simple models are sufficient. With our most complex model, we took a pragmatic approach and incorporated an additional noise term that can account for some of the unavoidable model misspecification. Such a noise term could potentially also be included in methods that allow for more efficient Bayesian inference with stochastic models [36].

The emergence of new variants with increased contagiousness or resistance mutations has the potential for significant implications for control of COVID-19 especially given that very few countries have been able to use NPIs alone to bring the viral growth rate sub-critical for extended periods of time. Integrating modeling into surveillance systems will help facilitate early-warning systems and improve our ability to design both pharmaceutical and non-pharmaceutical interventions that can stop the spread of COVID-19.

Acknowledgements

Portions of this work were done under the auspices of the U.S. Department of Energy under contract 89233218CNA000001 and supported by National Institutes of Health (www.nih.gov) grants P01-AI131365, R01-OD011095, and R01-AI028433 (CHvD). RK, NH, and ERS were funded by the US National Science Foundation RAPID grant PHY-2031756.

Literature Cited

[1] Tegally, H. et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* (2020).

[2] Volz, E. et al. Transmission of SARS-CoV-2 lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv* (2021).

[3] Volz, E. et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64–75 (2021).

[4] Dearlove, B. et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proceedings of the National Academy of Sciences* **117**, 23652–23662 (2020).

[5] Fontanet, A. et al. SARS-CoV-2 variants and ending the COVID-19 pandemic. *Lancet* (2021).

[6] McCarthy, K. R. et al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* (2021).

[7] Wibmer, C. K. et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *bioRxiv* (2021).

[8] Weisblum, Y. et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9**, e61312 (2020).

[9] Grubaugh, N. D. et al. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* **4**, 10–19 (2019).

[10] Korber, B. et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the

COVID-19 Virus. *Cell* **182**, 812–827 (2020).

[11] Yurkovetskiy, L. et al. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183**, 739–751 (2020).

[12] Davies, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* (2021).

[13] Davies, N. G. et al. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature* (2021).

[14] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**, 533–534 (2020).

[15] COVID-19 Viral Genome Analysis Pipeline. https://cov.lanl.gov (2020).

[16] Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33–46 (2017).

[17] Global Initiative on Sharing All Influenza Data. http://www.gisaid.org/ (2008).

[18] Liao, X. & Meyer, M. C. cgam: An r package for the constrained generalized additive model. *Journal of Statistical Software, Articles* **89**, 1–24 (2019).

[19] Chen, C. et al. Quantification of the spread of SARS-CoV-2 variant B.1.1.7 in Switzerland. *medRxiv* (2021).

[20] Ali, S. T. et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science* **369**, 1106–1109 (2020).

[21] Carpenter, B. et al. Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).

[22] Rozhnova, G. et al. Model-based evaluation of school- and non-school-related measures to control the COVID-19 pandemic. *Nature Communications* **12**, 1614 (2021).

[23] Ward, H. et al. Antibody prevalence for SARS-CoV-2 following the peak of the pandemic in england: REACT2 study in 100,000 adults. *medRxiv* (2020).

[24] Vos, E. R. A. et al. Nationwide seroprevalence of SARS-CoV-2 and identification of risk factors in the general population of the Netherlands during the first epidemic wave. *J Epidemiol Community Health* (2020).

[25] Sanche, S. et al. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerg Infect Dis* **26**, 1470–1477 (2020).

[26] Zhou, F. et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).

[27] Ali, S. T. et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science* **369**, 1106–1109 (2020).

[28] Lavezzo, E. et al. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. *Nature* **584**, 425–429 (2020).

[29] Wu, J. T. et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat Med* **26**, 506–510 (2020).

[30] Verity, R. et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* **20**, 669–677 (2020).

[31] Grasselli, G. et al. Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA* **323**, 1574–1581 (2020).

[32] Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S. & King, A. A. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences* **112**, 719–724 (2015).

[33] van Kampen, N. G. *Stochastic processes in physics and chemistry*. Elsevier, Amsterdam, 3rd edition (2007).

[34] Hale, T., Webster, S., Petherick, A., Phillips, T. & Kira, B. Oxford COVID-19 Government Response Tracker (2020).

[35] van Dorp, L. et al. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nature Communications* **11**, 5986 (2020).

[36] Fintzi, J. et al. Using multiple data streams to estimate and forecast SARS-CoV-2 transmission dynamics, with application to the virus spread in Orange County, California (2020).

[37] Mumford, D., Series, C. & Wright, D. *Indra's Pearls: The Vision of Felix Klein*. Cambridge University Press (2002).

[38] Douc, R. & Cappe, O. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, 64–69 (2005).

Appendix A: Population genetics model

Variant frequencies over time

Here we derive the equation for the mutant frequency p_n as a function of the generation n under the population genetics model with both selection and migration (Eq 4). In the main text, we derived that the new variant's frequency p' in the next generation is

$$p' = \frac{(1+s)p + (1-p)m}{1+sp}. \tag{A-1}$$

We can write this slightly differently as a Möbius transformation of p:

$$p' = \frac{(1+s-m)p+m}{sp+1} \equiv \begin{pmatrix} 1+s-m & m \\ s & 1 \end{pmatrix} \cdot p. \tag{A-2}$$

In general a Möbius transformation has the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot x \equiv \frac{ax+b}{cx+d}, \tag{A-3}$$

and it has the nice property that $A \cdot (B \cdot x) = (AB) \cdot x$ for two matrices A and B (see e.g., [37]). Hence, in order to find the variant frequency in the n-th generation, we need the n-th matrix power of

$$M = \begin{pmatrix} 1+s-m & m \\ s & 1 \end{pmatrix}. \tag{A-4}$$

For this we have to diagonalize M. The eigenvalues of M are equal to $1+s$ and $1-m$. Then eigenvectors are given by $(1,1)^T$ and $(m,-s)^T$. Hence, we can write $M = U\Lambda U^{-1}$ where

$$U = \begin{pmatrix} 1 & m \\ 1 & -s \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} 1+s & 0 \\ 0 & 1-m \end{pmatrix}. \tag{A-5}$$

Then $M^n \cdot p_0 = U\Lambda^n U^{-1} \cdot p_0$, which is equal to

$$p_n = \frac{[s(1+s)^n + m(1-m)^n]p_0 + m[(1+s)^n - (1-m)^n]}{s[(1+s)^n - (1-m)^n]p_0 + [m(1+s)^n + s(1-m)^n]}. \tag{A-6}$$

Transformation to the logit scale

To avoid numerical issues during inference, we transform the mutant frequency to the logit-scale, and re-parameterize the model. Let $R = \left(\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right)$. Then

$$\text{logit}(x) = \log(R \cdot x). \tag{A-7}$$

Let $r_0 = R \cdot p_0$ and $r_n = R \cdot p_n$. Then

$$r_n = (RU\Lambda^n U^{-1} R^{-1}) \cdot r_0, \tag{A-8}$$

and we have $RU = \begin{pmatrix} 1 & m \\ 0 & -(m+s) \end{pmatrix}$. Hence,

$$\left(RU\Lambda^n (RU)^{-1}\right) = \begin{pmatrix} (1+s)^n & \frac{m}{m+s}\left[(1+s)^n - (1-m)^n\right] \\ 0 & (1-m)^n \end{pmatrix} \tag{A-9}$$

and therefore

$$r_n = \left(\frac{1+s}{1-m}\right)^n r_0 + \frac{m}{m+s}\left[\left(\frac{1+s}{1-m}\right)^n - 1\right]. \tag{A-10}$$

This equation has a removable singularity at $s = -m$. To see this, we write

$$\frac{m}{m+s}\left[\left(\frac{1+s}{1-m}\right)^n - 1\right] = \frac{m}{1-m}\frac{\left(\frac{1+s}{1-m}\right)^n - 1}{\frac{1+s}{1-m} - 1} = \frac{m}{1-m}\sum_{k=0}^{n-1}\left(\frac{1+s}{1-m}\right)^k. \tag{A-11}$$

Now we define $\sigma = \frac{1+s}{1-m}$ and $\mu = m/(1-m)$. Then we get the following re-parameterized model:

$$r_n = \sigma^n r_0 + \mu\frac{1-\sigma^n}{1-\sigma}. \tag{A-12}$$

Appendix B: Stochastic epidemiological model

Time-dependent infection rate

To model the effects of government restrictions such as lock-downs, the infection rate β is a (smoothed) piece-wise constant function of time. We allow for $n \in \{3,4\}$ epidemic stages to model e.g. unrestricted spread, lockdown, and relaxation of the lockdown [22]. We let β vary smoothly between the epidemic stages, thereby allowing for an uptake period for (release of) restrictions. More precisely, $\beta = \beta(t)$ is defined as

$$\begin{aligned} \beta(t) &= \beta_0\left(1 - H_\upsilon(t - t_1)\right) \\ &\quad + \sum_{i=1}^{n-2} \beta_i \cdot H_\upsilon(t - t_i)\left(1 - H_\upsilon(t - t_{i+1})\right) \\ &\quad + \beta_{n-1} \cdot H_\upsilon(t - t_{n-1}) \end{aligned} \tag{B-1}$$

where H_υ is a smoothed Heaviside function defined by $H_\upsilon(t) = (1 + \exp(-t/\upsilon))^{-1}$, with parameter υ determining the duration of the transition period between different epidemic stages. The time-varying infection rate β is illustrated in Fig. S7A. To determine the parameter υ, we require that $H_\upsilon(\Delta t/2) = 95\%$, where Δt is the length of the uptake period. Solving for υ, we find $\upsilon = \Delta t/(2 \cdot \mathrm{logit}(0.95))$. We set $\Delta t = 7$ days, hence the uptake period takes a week.

In the case of variant B.1.1.7 in the Netherlands, we added an external force of infection, λ_{wt} and λ_{mt} for wild-type and mutant respectively, to account for contacts with infectious individuals from the UK. Because the number of infected individuals the UK varied over time, the external infection forces of infection λ_{wt} and λ_{mt} depend on time as

follows

$$\lambda_{\mathrm{wt}}(t) = \lambda_0 \mathbb{E}[I_{\mathrm{wt}}(t)/N(t)], \quad \lambda_{\mathrm{mt}}(t) = \lambda_0 \mathbb{E}[I_{\mathrm{mt}}(t)/N(t)]. \tag{B-2}$$

Here, the trajectories $I_{\mathrm{wt}}(t)/N(t)$ and $I_{\mathrm{mt}}(t)/N(t)$ correspond to the fraction of individuals in the UK that are infected with wild-type and mutant respectively. The expectation is taken over the filtered trajectories that were reconstructed with the SMC algorithm. Finally, the scaling parameter λ_0 represents the product of the contact rate between individuals in the two countries, and the probability of infection per contact. When we fit the model to the Dutch data, the parameter λ_0 is estimated, but the fraction of infectious UK citizens is assumed known.

Initial condition

To complete the description of the dynamic model, we have to specify the initial conditions. Let ζ and ξ denote the fractions of infected and removed individuals at time t_0, respectively. To determine the correct balance between exposed, infectious, and severely infected individuals, we compute the eigenvalues of the Jacobian matrix of the infinite population limit of the MJP model. The linearized system without the S and R compartments around the disease-free steady state with $S = (1-\xi)N$ and $R = \xi N$ equals

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} E \\ I \\ H \end{pmatrix} = \begin{pmatrix} -\alpha & \beta(1-\xi) & 0 \\ \alpha & -\gamma-\nu & 0 \\ 0 & \nu & -\omega \end{pmatrix}\begin{pmatrix} E \\ I \\ H \end{pmatrix} \tag{B-3}$$

We use the eigenvector $X_0 = (E_0, I_0, H_0)^T$ with $\sum_{i=1}^{3} X_0^i = \zeta N$ corresponding to the dominant eigenvalue of the Jacobian matrix in Eq. (B-3) to define the initial condition of the model. The parameter p_0 determines the initial fraction of infections with the mutant virus, and hence we have $E_{\mathrm{mt},0} = p_0 E_0$, $I_{\mathrm{mt},0} = p_0 I_0$, $E_{\mathrm{wt},0} = (1-p_0)E_0$, and $I_{\mathrm{wt},0} = (1-p_0)I_0$. We then set $S_0 = (1-\xi-\zeta)N$ and $R_0 = \xi N$. Finally, the initial state of the stochastic model is randomized by sampling from a Poisson distribution with mean equal to the deterministic initial value.

Hybrid model simulation

Because for small population sizes the diffusion approximation defined by the system of SDEs in Eq. (9) breaks down, as it e.g., does not allow for fixation of the mutant, we model small population sizes discretely using an adaptive tau-leap approximation of the Markov jump process (Eq. (5)). Hence, we implemented a hybrid algorithm in which variables can switch between a discrete and continuous type.

At any particular time t, the system consists of continuous components X^i for $i \in \mathscr{C}(t)$ and discrete elements X^i for $i \in \mathscr{D}(t) = \{1,\ldots,n\} \setminus \mathscr{C}(t)$. Since for $i \in \mathscr{C}(t)$, the continuous element $X^i(t)$ is generally non-constant, the transition rates $\eta_j(X,t)$ will in general be time dependent. This means that we have to integrate the following system

of SDEs and ODEs

$$\begin{aligned} \mathrm{d}X^i &= \sum_{j=1}^{k} \varepsilon_j^i \eta_j(X,t)\mathrm{d}t + \sum_{j=1}^{k} \varepsilon_j^i \sqrt{\eta_j(X,t)}\mathrm{d}B_t^j + \tau X^i \mathrm{d}\widetilde{B}_t^i, \quad i \in \mathscr{C}(t) \\ \frac{\mathrm{d}H^j}{\mathrm{d}t} &= \eta_j(X,t), \quad j \in \mathscr{E}(t) \equiv \{j \in \{1,\ldots,k\} : \exists i \in \mathscr{D}(t) : \varepsilon_j^i \neq 0\} \end{aligned} \tag{B-4}$$

Hence, we have to keep track of those transition rates η_j for which the increment ε_j^i is non-zero for a discrete component X^i. The initial conditions for the hybrid system Eq B-4 are given by $X^i(t_m) = x_m^i$ and $H_j(t_m) = 0$. We then integrate the system until time $t_{m+1} = t_m + h_m$. At this point, we sample the number of stochastic events Y_m that occurred in the time interval $(t_m, t_{m+1}]$, from the Poisson distribution

$$Y_m \sim \text{Poisson}\left(\textstyle\sum_{j\in\mathscr{E}(t_m)} H_j(t_{m+1})\right) \tag{B-5}$$

Thereafter, Y_m events with index j are sampled from the categorical distribution, with probability proportional to the cumulative transition rate $H_j(t_{m+1})$. The increments ε_j are then added to the discrete part of the state X

$$X^j(t_{m+1}) \mapsto X^j(t_{m+1}) + \varepsilon_i^j, \quad i \in \mathscr{D}(t_m) \tag{B-6}$$

After applying these Y_m discrete transitions, we have to re-evaluate which components of the state are discrete and which are continuous. We choose a fixed threshold $T = 50$ below and above the populations are discrete and continuous, respectively. Hence, at time t_{m+1} we update the partition of $\{1,\ldots,n\}$ as follows

$$\mathscr{D}(t_{m+1}) = \{i : X^i < T\}, \quad \mathscr{C}(t_{m+1}) = \{i : X^i \geq T\} \tag{B-7}$$

Finally, we set the next initial condition $x_{m+1} = X(t_{m+1})$ and $H(t_{m+1}) = 0$ and repeat the process.

The tau-leap step size h_m is chosen adaptively such that the expected number of events $\mathbb{E}[Y_m]$ within each τ-leap interval $(t_m, t_{m+1}]$ is approximately equal to 1. To accomplish this, we choose

$$h_m = \min\left\{h_{\max}, \left(\textstyle\sum_{j\in\mathscr{E}(t_m)} \eta_j(X(t_m), t_m)\right)^{-1}\right\} \tag{B-8}$$

where $h_{\max} = 1\text{d}$. Between jumps, the hybrid system (Eq B-4) is integrated using the Euler-Mayurama method with a step size of $\min\{0.01, h_m\}$.

Sequential Monte-Carlo

The method used for inference is described in full detail and generality elsewhere [32]. Here we give a brief description highlighting some of the choices made for this particular model and data set.

In order to reconstruct the latent epidemic trajectories X, given the observed data O, consisting of death incidence data D and genetic data F^{mt} and F, we use sequential Monte-Carlo (SMC). We simulate $J = 10^4$ replicates of the model (particles) forward in time from one observation time (t_{i-1}) to the next (t_i), each with different initial

conditions $X_j(t_{i-1})$. Given each of the J predicted states $X_j(t_i)$ of the model at time t_i, we calculate the likelihood $w_j = L(O_i|X_j(t_i),\theta)$. We then sample with replacement J particles with probability proportional to the weight w_j using a systematic resampling method [38]. The re-sampled particles are used as initial condition at time t_i, and we repeat the above steps until we reach the final observation.

The Monte-Carlo estimate of the conditional likelihood of observation $O_i|O_{i-1}$, is given by the average of the weights

$$L(O_i|O_{i-1},\theta) = \frac{1}{J}\sum_{j=1}^{J} L(O_i|X_j(t_i),\theta) \tag{B-9}$$

where we write $L(O_1|O_0,\theta) \equiv L(O_1|\theta)$. The total likelihood of the time series given θ is equal to

$$L(O_1,\ldots,O_N|\theta) = \prod_{i=1}^{N} L(O_i|O_{i-1},\theta) \tag{B-10}$$

All likelihood computations are done on the log-scale to minimize floating-point errors.

To estimate parameters, we extended the state X with the parameter vector θ, allowing the parameters to be perturbed after each observation time. For the j-th particle, we now have a state (X_j,θ_j), and weight $w_j = L(O_i|X_j(t_i),\theta_j)$. The extended-SMC algorithm is then iterated $M = 200$ times, and after each iteration m the magnitude of the parameter perturbations is reduced. The perturbations are Gaussian $\theta_j \mapsto \theta_j + a^m\varepsilon$, with $\varepsilon \sim \mathcal{N}(0,\Sigma)$. Here, Σ is a diagonal matrix with diagonal elements given in Table S1. For bounded parameters, the perturbations are reflected in the boundary of the domain. The ratio $a \in (0,1)$ reduces the magnitude of the parameter perturbations. We choose $a = \sqrt[M]{10^{-2}}$ such that after M iterations the magnitude of the perturbations is reduced by 99%. After each iteration of the extended-SMC algorithm, for each particle we reset the state X_j to the a randomly sampled initial state of the epidemic model, while θ_j is inherited from the previous extended-SMC iteration. To speed-up the computations, we implemented the model and SMC algorithm in C++ and used multi-threading to update particles between observations in parallel.