SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers

Marlin D. Figgins[1,2,*] and Trevor Bedford[1,3]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[2]Department of Applied Mathematics, University of Washington, Seattle, WA, USA
[3]Howard Hughes Medical Institute, Seattle, WA, USA
[*]Corresponding author: mfiggins@uw.edu

December 9, 2021

Abstract

Accurately estimating relative transmission rates of SARS-CoV-2 Variant of Concern and Variant of Interest viruses remains a scientific and public health priority. Recent studies have used the sample proportions of different variants from sequence data to describe variant frequency dynamics and relative transmission rates, but frequencies alone cannot capture the rich epidemiological behavior of SARS-CoV-2. Here, we extend methods for inferring the effective reproduction number of an epidemic using confirmed case data to jointly estimate variant-specific effective reproduction numbers and frequencies of co-circulating variants using case data and genetic sequences across states in the US from January to October 2021. Our method can be used to infer structured relationships between effective reproduction numbers across time series allowing us to estimate fixed variant-specific growth advantages. We use this model to estimate the effective reproduction number of SARS-CoV-2 Variants of Concern and Variants of Interest in the United States and estimate consistent growth advantages of particular variants across different locations.

Introduction

As SARS-CoV-2 evolves, variants may emerge that increase in their ability to transmit and escape acquired immunity [1]. Quantifying the observed growth advantages of SARS-CoV-2 variants allows us to understand which variants are able to thrive in different locations [2,3]. Relating genomic data of SARS-CoV-2 lineages to epidemic surveillance data is difficult. Although it is typical to use phylodynamic methods to analyze genetic sequence data from epidemics, the sheer amount of data as well as challenges to describing fitness effects in phylodynamic models make these methods hard to apply to potential differences in transmission rate among circulating variants. In order to deal with the limitations of phylodynamic inference, previous studies have estimated the growth of lineages using observed frequencies in sequenced SARS-CoV-2 samples [4–7]. Such methods often model the

frequency of lineages using multinomial logistic regression [6, 7], which generally assumes that genetic variants have a fitness advantage over one another which is fixed in time and acts as a estimate for the selective advantage of different variants at the level of frequencies. Although a consistent increase in frequency of one variant over another is expected to reflect differences in transmission rate, these models do not directly account for the complicated infection and transmission dynamics which influence which variants lead to local and regional epidemics. When dealing with competition between variants, variants which are declining in frequency can still lead to an increasing number of infections. Similarly, growth in frequency does not necessarily entail an increase in absolute infections.

To better capture epidemiological dynamics, there are methods which describe the growth in number of infections using confirmed case, hospitalization, or death data to estimate changes in the effective reproduction number R_t, the average number of infections a single infectious individual generates, during a given outbreak. Although these methods are excellent for describing overall epidemic growth rates, they cannot capture the evolutionary dynamics and fitness changes between different variants since they often assume the population dynamics are described by a singular R_t trajectory [8,9], which internally is unrelated to the genetic and phenotypic composition of the population. This is of particular importance in the analysis of an epidemic in which a dominant lineage may be declining overall, but some sublineage is rapidly increasing in frequency and absolute prevalence, creating the potential for a secondary wave of infections that may go unnoticed at first glance. To overcome this we require models that partition case counts into contributions from different variants to estimate variant-specific effective reproduction numbers.

The current COVID-19 pandemic serves as an important example of this phenomenon. After initial emergence in late 2020, over the course of 2021, Variant of Concern (VOC) and Variant of Interest (VOI) viruses spread throughout the world and replaced existing viral diversity. Multiple WHO designated [10] VOC and VOI viruses circulated in spring and early summer 2021, but this diversity was largely replaced by Delta variant viruses which became globally dominant in late summer 2021. Although it's now clear that Delta had greater transmissibility than other variants, rigorous estimates of the relative fitness of circulating VOC and VOI viruses are of interest. Here, we develop a joint epidemiological and population genetic model of SARS-CoV-2 to assess the growth of different variants over time and infer differences in the effective reproduction numbers of SARS-CoV-2 variants as well as underlying frequency of variants under noisy sampling. We apply this model to sequence data and case count data from United States between January and October 2021 to estimate differences in transmissibility between circulating VOC and VOI viruses.

Results

Model Overview We implement two models of variant-specific effective reproduction number based on a renewal equation framework of epidemic spread (see Methods), a free R_t model and a fixed growth advantage model. These models assume that new infections are determined by two essential parameters: the effective reproduction number which determines the average number of secondary infections generated over the course of a primary infection and the generation time which determines length of infection as well as

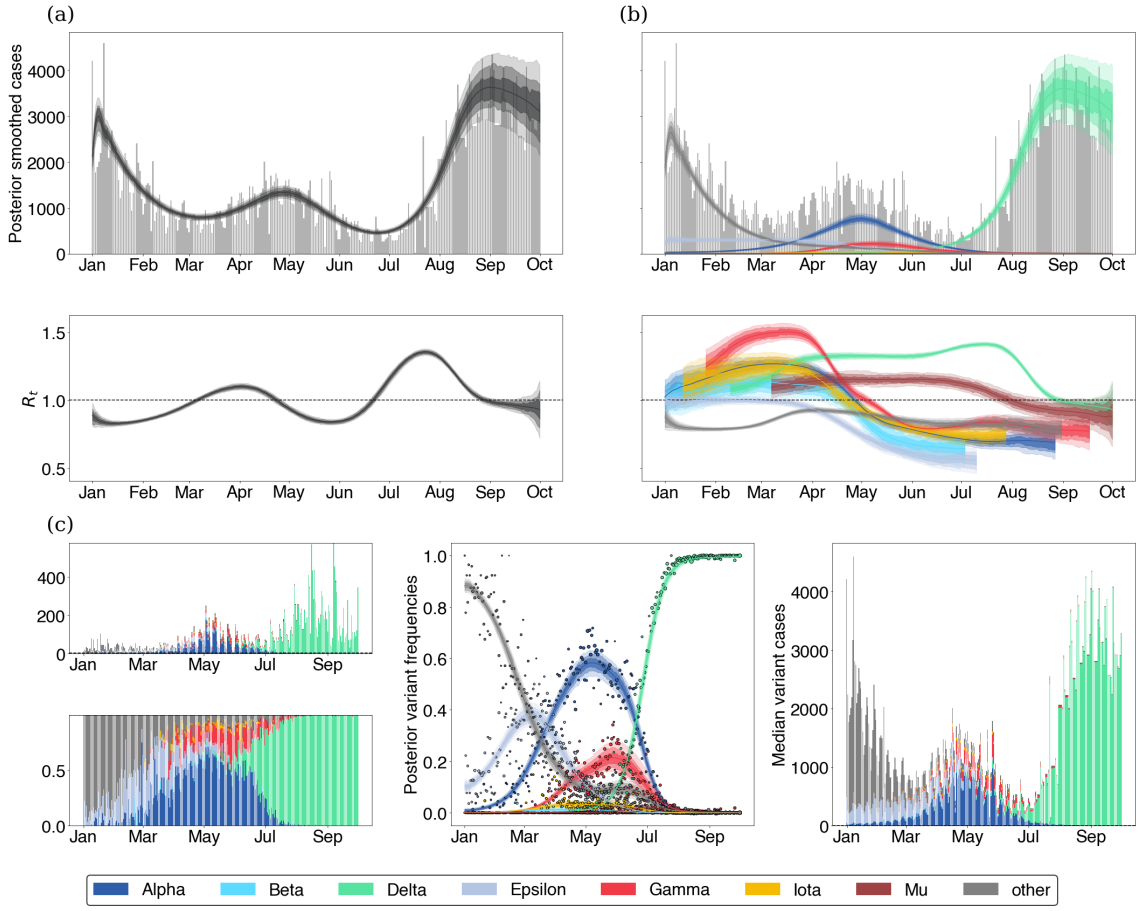

Figure 1. Fitting the free R_t model to Washington state data. (a) When assessing epidemic growth rates, we often compute a single effective reproduction number trajectory which is effectively an average over the all viruses in population. Gray bars are observed daily case counts, while blue lines are model inferences with 50%, 80% and 95% credible intervals. (b) Epidemics are made of different variants which may differ in fitness. Using case counts alongside sequences of different lineages allows us to understand the proportion of different variants in the population. (c) Using both case count and frequency data, we can estimate the effective reproduction numbers of different lineages.

their relative transmissibility over the course of their infection. In both models, variants generate infections independently of one another, but the sum of infections across variants is observed through surveillance data like case counts or hospitalizations. In order to disaggregate infections by variant we rely on frequency estimates which are informed by counts of sequenced samples using a Dirichlet-multinomial likelihood.

The transmission of each variant is modeled using a deterministic renewal equation which allows for realistic delay distributions between infection, transmission, and detection as a case. With this approach, we need only to determine the initial number of infections and the variant-specific effective reproduction numbers to estimate the frequency of each variant in the population over time. Due to this, the differences between the two models is determined in how each parameterizes variant-specific effective reproduction numbers.

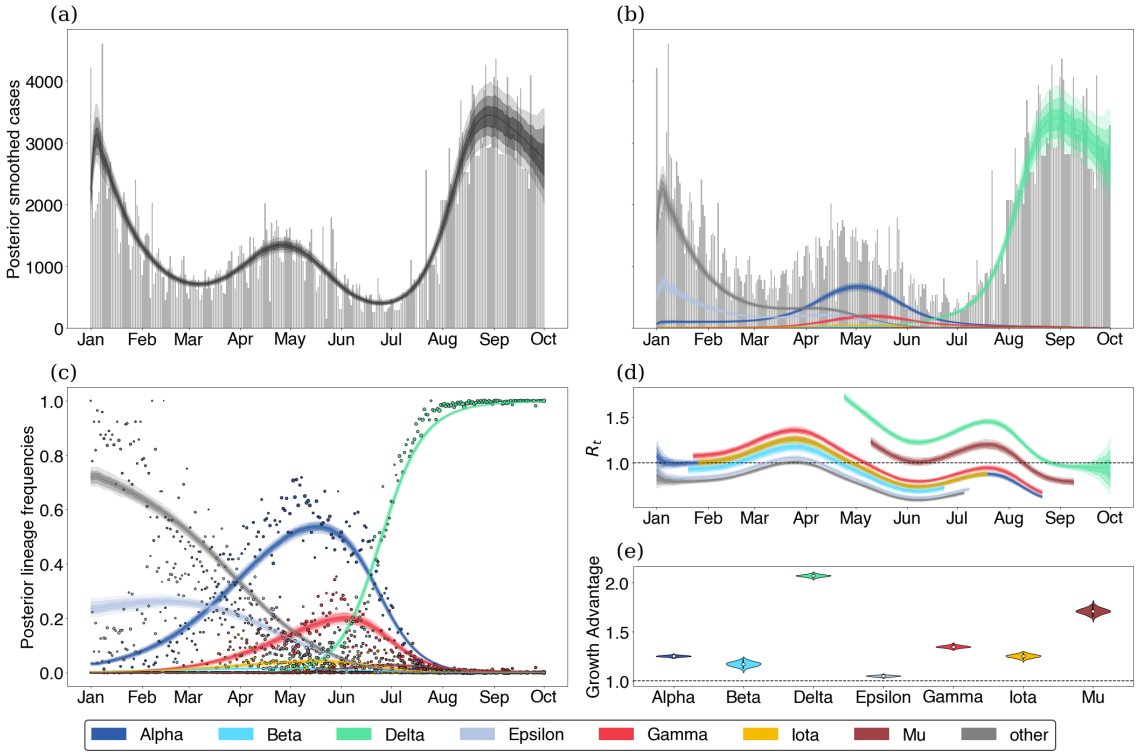

Figure 2. Fitting the fixed growth advantage model to Washington state data. (a) Posterior expected cases without weekly seasonality in reporting rate. Gray bars are observed daily case counts, while blue lines are model inferences with 50%, 80% and 95% credible intervals. (b) Posterior expected cases by lineage. Each colored lines is a different variant with intervals of varying opacity showing 50%, 80% and 95% credible intervals. (c) Posterior lineage frequency against observed sample frequency. Dots represent observed weekly frequencies in sequence data and each colored line is a different variant with shaded CIs. (d) Variant-specific effective reproduction numbers. (e) Posterior growth advantage by variant.

In the first model, we introduce a free variant R_t which infers the effective reproduction number of each variant independently from one another to allow for non-linear relationships between the growth rates of different variants over time. Each variant effective reproduction number is parameterized using an exponentiated spline basis, so that the log effective reproduction numbers are described by a linear basis expansion.

The second model is a fixed growth advantage model of variant R_t in which each variants has its own multiplicative growth advantage which acts as a scaling to a single non-variant R_t trajectory. With this fixed growth advantage model, we parameterize fitness of variants at the level of transmission by inferring variant-specific effective reproduction numbers. This differs from previous work on variant effective reproduction numbers which often parameterize these differences by assuming logistic growth of frequencies [11, 12]. Though, in general, our method allows one to estimate variant growth in the frequency domain in terms of effective reproduction number differences, we find that assuming a fixed advantage for variants results in estimates which are qualitatively similar to the aforementioned models which assume fixed growth advantages in frequency growth. This provides the

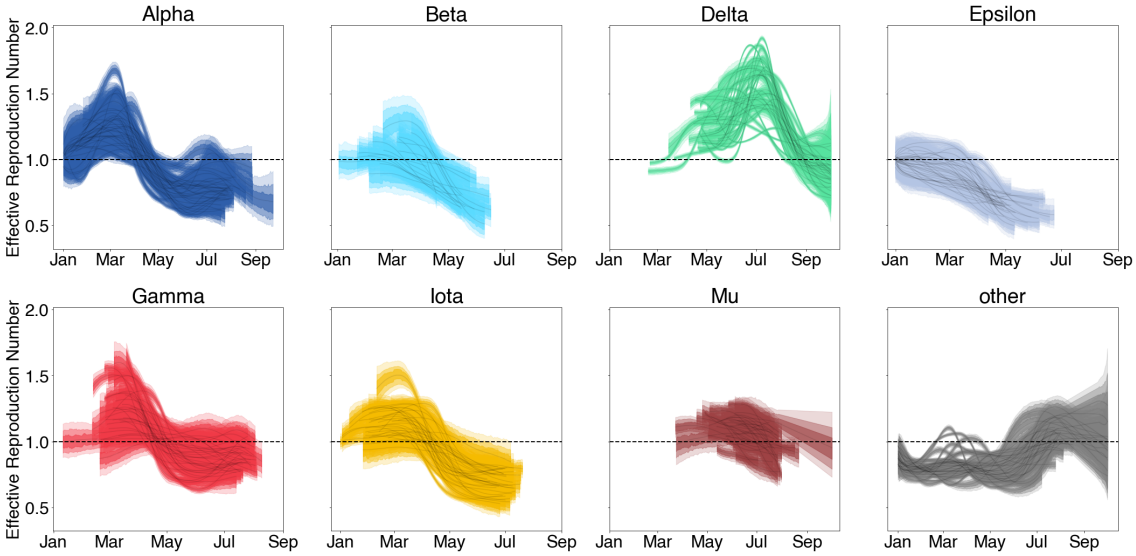

Figure 3. Inferred effective reproduction numbers from free R_t model in 36 states show consistent trends of variants across states. Each panel shows a series of 36 trajectories, representing R_t through time for this variant across states. Shaded intervals show 50%, 80% and 95% credible intervals.

additional benefit of the inferred parameters being interpretable as scaling the effective reproduction number.

In cases where a singular fixed growth advantage is insufficient to describe the data, we return to our first model in which the effective reproduction numbers of variants are modeled without assuming a fixed advantage, accounting for possible variation in variants' advantages over one another over time.

We demonstrate these models on data from Washington State. The free R_t model is shown in Figure 1 and the fixed growth advantage model is shown in Figure 2. Example model output from several other states is provided in the supplemental appendix.

Estimating growth advantages in the United States We estimate the effective reproduction numbers of SARS-CoV-2 Variant of Concern and Variant of Interest viruses in the United States using daily confirmed case counts obtained from the US CDC and sequence counts annotated by variant obtained from the Nextstrain-curated 'open' dataset [13] (see Data and code accessibility). Each sequence is labeled with a Nextstrain clade [13] and we partition clades into variants based on designated WHO VOC/VOI status [10]. Nextstrain clades annotated in the fashion correspond to a subset of major lineages designated by PANGO [14]. We consider the following 7 variants which have been flagged as variants of interest or concern and which circulated in the US during 2021: Alpha (PANGO lineage B.1.1.7, Nextstrain clade 20I), Beta (lineage B.1.351, clade 20H), Gamma (lineage P.1, clade 20J), Delta (lineage B.1.617.2, clade 21A), Epsilon (lineage B.1.427/429, clade 21C), Iota (lineage B.1.526, clade 21F), and Mu (lineage B.1.621, clade 21H). We use a cutoff of 2000 sequences from a particular variant across states to determine threshold of circulation. This eliminates Eta, Lambda, Kappa and Theta from consideration and groups

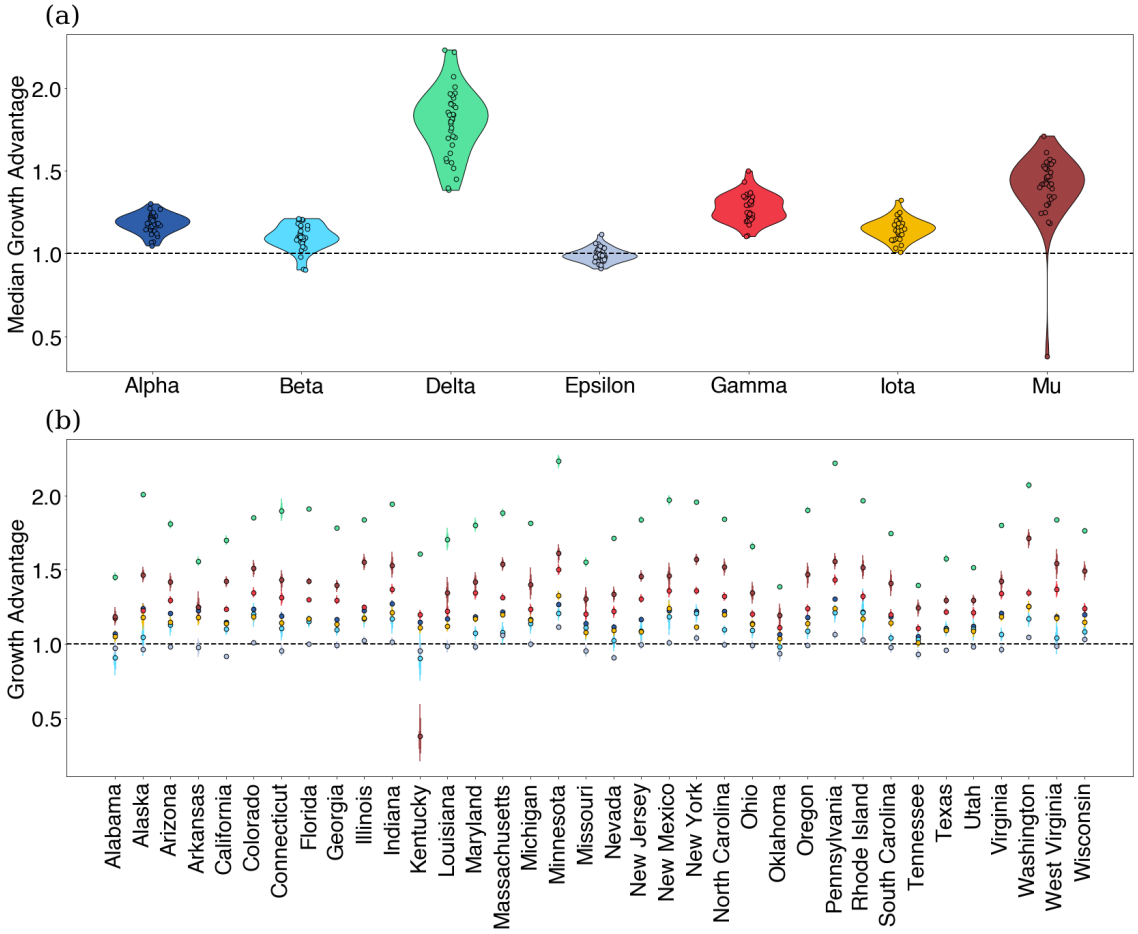

Figure 4. Using fixed growth advantage model, we infer growth advantages for 7 variants in 36 US states. (a) Growth advantages for variants of concern. Each point is the median growth advantage inferred from a single state. (b) Same as (a) but visualized by variant.

these variants along with ancestral 'non-variant' viruses into a single 'other' category. We use a cutoff of 5000 sequences from a particular state as basis for including the state in the dataset. This cutoff left 36 states available for inference.

In order to inform our estimates of the frequency of genetic variants, we divide sequences from each state into daily sample counts for each of the 7 variants above and a single 'other' category. We then use these counts alongside the daily case counts in each state to estimate the effective reproduction number for individual variants using our free R_t model. We find that overall there appears to be consistent trends in the effective reproduction numbers of variants across the United States (Fig. 3). We see that non-variant viruses were declining from January onwards, while initial VOCs Alpha and Gamma initially had $R_t > 1$, but saw R_t decline below one across most states in April and May respectively. Upon arrival in May, Delta shows significantly higher values of R_t that don't decline below 1 until September.

In order to transform these observed trends to a variant-specific growth advantage, we

rely on our fixed growth advantage model which infers a fixed variant-specific growth advantage as a multiplicative scaling of the effective reproduction number. Using the fixed growth advantage model, we find that most variants identified share some positive growth advantage with the exception of Epsilon. Further, these growth advantages appear to be consistent between the states analyzed (Fig. 4). Alpha, Beta, Gamma and Iota show modest growth advantage over largely ancestral 'other' viruses, while Mu and Delta show larger growth advantages. Mu has previously been associated with increased neutralization resistance to convalescent serum [15], and its advantage of 1.2–1.8 across states is perhaps partially driven by immune escape. Despite this, Mu's growth advantage whether from immune escape or otherwise was insufficient to outcompete Delta in any of the states analyzed. Delta's advantage of 1.6–2.0 across states is particularly significant. Given this large growth advantage was evident in May (Fig. 3), Delta's rapid rise in frequency and sizable epidemic should have been clear at the time. The significant growth advantage observed in Delta is recapitulated in other studies including Obermeyer et al. [6] and Vöhringer et al. [16].

Discussion

We find that a model that partitions case count data based on variant frequency in sequence data works well to describe SARS-CoV-2 variant dynamics in the United States from January to October 2021. In each state, spring waves are primarily driven by the arrival of Alpha, Beta, Gamma, and Iota variants. However, as these waves are subsiding, the arrival of Delta with a significantly greater growth advantage, drives a large summer wave. Importantly, we can directly estimate a variant-specific R_t, which for example, shows that Delta was a growing rapidly sub-epidemic across states in May, before its impact was noticeable in overall case counts. We imagine that this general approach could provide early warning of imminent epidemics driven by low-frequency but highly transmissible variants.

With this mind, this work is not without limitations. The underlying transmission model is deterministic and does not account for demographic stochasticity and over-dispersion in transmission which has been documented in SARS-CoV-2 transmission [17]. As with all methods which depend on parameterizations of the generation time, misspecification of the generation time can be lead to biased estimates of the effective reproduction number or growth advantages [18]. In order to quantify this source of error, we derive an equation relating our inferred growth advantages, the epidemic growth rates, and the mean and standard deviation of the generation time distribution. This source of error can be partially combatted by converting effective reproduction numbers to their corresponding epidemic growth rates under the generation time assumption. (see Supplement Appendix) There is also a general need to account for biases in the case data which may not faithfully describe the infection dynamics of SARS-CoV-2 due to changes in case ascertainment rate, as possibly caused by differences in testing intensity, infection severity among other reasons. However, we suspect that case ascertainment remained largely consistent from January to October 2021.

We do not explicitly model multiple introductions of variants which can play an important

role in variants establishing themselves in different geographies at low infection counts and could bias our estimates of the effective reproduction number if not properly accounted for [8,19]. This could be especially impactful early on when variant cases are driven by multiple importations from a large epidemic elsewhere in the world. However, we expect once local transmission is predominant that estimated R_t will reflect characteristics intrinsic to the variant in the local geography. Using hierarchical models of variants to jointly estimate growth advantages and pool estimates across locations could be a useful approach for analyzing consistency between growth advantages of variants geographically and beginning to combat the issue of multiple introduction events. That said, fully combating this issue would likely involve incorporating demographic stochastic into the model at the level of transmission and reduce its speed of inference, scalability, and limit available inference options.

Although there are several ways to improve these methods and expand their applicability, our current model does have utility as a way of assessing early claims of variant advantages and is able to show there is evidence of consistent variant advantages shared between different geographies. Additional work is needed to attribute these inferred advantages to biological mechanisms like immune escape and transmissibility [1]. Modeling the effect of changes in other factors such as contact patterns or non-pharmaceutical interventions can be done with the current formulation of the model by including quantities of interest as features in the R_t model as in Sharma et al. [20].

In general, the development of methods which can account for fitness differences between genetic variants is much needed in order for proper epidemic preparedness. Our method provides one way of analyzing the growth rates of SARS-CoV-2 variants without directly parameterizing how variants grow in terms of frequency by instead focusing on differences in the effective reproduction number. In cases where the assumption of a fixed growth advantage is warranted and justified, our fixed growth advantage model provides a way of quantifying variant growth advantages at the level of transmission which allow for various delays between infection, transmission, and sampling.

Our method can be extended to analyze the role of specific constituent mutations defining a variant or lineage in changing the effective reproduction number of specific variants directly, similar to the model formulation of Obermeyer et al. [6]. With this in mind, our method potentially has use for evolutionary forecasting of variants for SARS-CoV-2 as we inform the frequency dynamics of co-circulating variants by describing their population-level transmission dynamics. Extending the model further towards this aim will require methods for quantifying population immunity as well as escape potential for circulating and emerging SARS-CoV-2 variants.

With these issues in mind, surveillance of variants should be folded into standard epidemiological surveillance as knowledge of variant-specific growth advantages will be useful for forecasting growth of cases, hospitalization, deaths, vaccine effectiveness among other key metrics related to epidemic response.

Methods

Using sampled counts of sequences from different lineages as well as case data, we can infer jointly infer the proportion of variants in the larger population and the effective reproduction number of these variants.

Modeling the infection process We estimate the effective reproduction number of competing lineages using a deterministic renewal equation based framework. These equations arise as the expectation of a Bellman-Harris branching process [21] which is a type of Branching process in which offspring generation depends on the age of infection.

The renewal equation framework allows one to model infection processes in a way that is mathematically equivalent to standard epidemic models like the SEIR compartment model [22], but in a way that can be more suitable for estimating the effective reproduction number and forecasting using arbitrary generation times. This renewal equation can be written as

$$I(t) = R_t \int_0^t I(\tau) g_{t-\tau} d\tau, \tag{1}$$

where g is the generation time. In addition, we also include onset distribution o for symptoms which allows us to compute the prevalence, or the number of active infections, as

$$P(t) = \int_0^t I(\tau) o_{t-\tau} d\tau. \tag{2}$$

We bin the generation time g and the onset distribution o to nearest day, so that we estimate the daily incidence $I(t)$ and prevalence $P(t)$ as

$$I(t) = R_t \sum_{\tau<t} I(\tau) g_{t-\tau} \tag{3}$$

$$P(t) = \sum_{\tau<t} I(\tau) o_{t-\tau} \tag{4}$$

We parameterize the generation time g as having Gamma distribution with mean 5.2 and standard deviation 1.72 in line with the estimates of [23] and onset time o as having LogNormal with mean 6.8 and standard deviation 2.0 in line with [24]. We note that the choice of generation time can have strong effects on the inferred effective reproduction number and growth advantage under renewal equation model. The effect of generation time choice is quantifiable as shown in Figures S2 S4 and supplemental appendix (see Relating epidemic growth rates to relative effective reproduction numbers). Though converting the posterior effective reproduction numbers to epidemic growth rates may be more robust to changes in generation time as can be seen in Figure S3.

This method of using delays to represent lags between infection and observation can be extended to use multiple delays to better fit other data sources such as hospitalization or deaths.

Modeling variant frequencies In the case of V variants co-circulating in a population, we denote incidence of variant v at time t as $I_v(t)$ and prevalence as $P_v(t)$. In this case, we can compute the frequency of variant v in the population at time t under the infection process outlined above as

$$f_v(t) = \frac{P_v(t)}{\sum_{1 \leq v \leq V} P_v(t)}. \tag{5}$$

Since we've defined the frequency in terms of the transmission dynamics, the variant-specific effective reproduction numbers $R_{t,v}$ and initial infections $I_v(0)$ determine the frequency dynamics directly. Therefore, we do not need to impose a parametric form on $f_v(t)$ directly as in other models of variant frequency.

Observation process for cases As most case time series in the United States have a strong weekly seasonal effect, we estimate a reporting rate which varies weekly, so that $\rho = (\rho_1, \ldots, \rho_7)$ as in [9]. We then define the observation likelihood using a negative binomial distribution as follows

$$Y_t \sim \text{NegBinom}(\rho_{[t]} P(t), \alpha), \tag{6}$$

where $[t] = t \mod 7+1$, α is an over-dispersion parameter relative to the Poisson distribution and $\text{NegBinom}(\mu, \alpha)$ is the negative binomial distribution with mean μ and variance $\mu + \alpha\mu^2$. In the case of multiple variants, we use $P(t) = \sum_{1 \leq v \leq V} P_v(t)$. The negative binomial likelihood is often used for modeling observation noise for count data such as epidemic time series which are often over-dispersed relative to a Poisson distribution.

Observation process for lineage annotations Suppose we're tracking the growth of V variants, our data for a given day t takes the form of daily counts $C_t = (C_{t,1}, \ldots, C_{t,V})$ of sequences of each variant with daily total $N_t = \sum_{1 \leq v \leq V} C_{t,v}$. We then assume that the likelihood of observing these counts of each lineage is described by a Dirichlet-multinomial distribution, so that

$$C_t \sim \text{DirMultinomial}\left(N_t, f(t) \cdot \left(\frac{1-\xi}{\xi}\right)\right), \tag{7}$$

given lineage frequencies $f(t) = (f_1(t), \ldots, f_V(t))$ and over-dispersion parameter $0 < \xi < 1$. Here, we use a Dirichlet-multinomial distribution to account for possible over-dispersion in the counts relative to the standard Multinomial distribution.

Basis expansions of log effective reproduction numbers Instead of inferring R_t directly, we parameterize the log effective reproduction number using a basis of cubic splines. Each basis spline is written as a column in the design matrix $\boldsymbol{X}$, so that

$$\ln R_t = \boldsymbol{X}\boldsymbol{\beta}, \tag{8}$$

where the $\boldsymbol{\beta}$ are to be estimated to parameterize the effective reproduction number. We then use locally adaptive smoothing of order one with a Laplace prior on the coefficients $\boldsymbol{\beta}$ to promote smoothness on the inferred R_t trajectory [25]. This method also allows one to use other predictors such as vaccination proportion, intervention indicators, temperature, humidity, etc...

Modeling variant-specific effective reproduction numbers To model the variant-specific reproduction numbers, we can infer individual independent effective reproduction number trajectories for each variant

$$\ln R_{t,v} = \boldsymbol{X}\boldsymbol{\beta}_v, \tag{9}$$

where each lineage v gets its own vector of parameters $\boldsymbol{\beta}_v$ in this model. We use the same prior structure as above to promote smoothness on inferred trajectories. This is our "free R_t" model which is used to generate Figure 1.

Modeling variant-specific growth advantages In order to use our model to infer growth advantages for specific variants, we can instead parameterize the effective reproduction numbers as

$$\ln R_{t,v} = \boldsymbol{X}\boldsymbol{\beta} + \delta_v, \tag{10}$$

where the parameters β are shared between all variants and δ_v is the log-scale variant-specific growth advantage of variant v. We consider $\Delta_v = \exp(\delta_v)$ to be the variant-specific growth advantage which can be seen in Figure 4.

Estimating an average effective reproduction number for an epidemic Given variant-specific effective reproduction numbers $R_{t,v}$ and the frequency of variants in the population $f_v(t)$, we define the average effective reproduction number to be

$$R_t^{\text{ave}} = \sum_{1 \leq v \leq V} R_{t,v} f_v(t), \tag{11}$$

which is the sum of the variant-specific effective reproduction numbers weighted by their frequency. This quantity can be seen in Figure 1.

Decomposing variant-specific growth advantages Under the free R_t model, we can attempt to decompose the relative advantage of different lineages over time into increased transmissibility and immune escape. For example, given that a variant effective reproduction number can be written as an sum of these two contributions, we can write

$$R_{t,v} = (1 + \eta_T) R_0 S_t + \eta_E R_0 \phi_t,$$

where R_0 is the basic reproduction number of the baseline strain, S_t is the fraction of the population susceptible to first infection, ϕ_t is the fraction of the population with prior immunity due to vaccination or past infection. Assuming that the baseline variant has no immune escape, we can then write the difference in the reproduction number as

$$R_{t,v} - R_t = R_0(\eta_T S_t + \eta_E \phi_t).$$

Writing S_t as $1 - \phi_t$, we have that

$$\frac{R_{t,v} - R_t}{R_0} = \eta_T + (\eta_E - \eta_T)\phi_t.$$

Using this model, we can estimate the relative contribution of each component by estimating the fraction of the population with some immunity jointly with these to variant advantages.

Priors for Bayesian Inference For both models, we provide a Laplace random walk prior on the spline coefficients $\boldsymbol{\beta}$ with scale parameter γ which itself has a HalfCauchy$(0, 0.5)$ prior distribution. In the fixed growth advantage model, only a baseline R_t trajectory is parameterized by $\boldsymbol{\beta}$ and the variant advantages δ_v are given a Normal$(0, 1)$ prior. The initial infected individuals for each variant have a uniform prior between 0 and 300,000. The weekly reporting rates $\rho_{[t]}$ each follow a Beta$(5, 5)$ prior, and the case observation over-dispersion is given a HalfNormal$(0, 10)$ prior on $\alpha^{-\frac{1}{2}}$. Finally, the over-dispersion parameter ξ is given a Beta$(1, 99)$ prior to penalize high levels of over-dispersion in sequencing.

Inference The model is implemented in NumPyro [26] in Python and approximate Bayesian inference was conducted using Stochastic Variational Inference [27] using the ADAM optimizer [28] with a learning rate of 0.01. For the analyses presented, all models are fit using a Multivariate Normal autoguide as implemented in NumPyro [26] which transforms the entire parameter space (with appropriate constraints on the individual parameter spaces) into a multivariate normal distribution for fitting purposes.

Models for each individual state in the United States variants data set were fit for 50,000 iterations and 3000 posterior samples were produced under both the free R_t model and fixed growth advantage model.

Data and code accessibility

Case count data was obtained from the US CDC using the 'United States COVID-19 Cases and Deaths by State over Time' dataset available from data.cdc.gov. Sequence data including date and location of collection as well as clade annotation was obtained via the Nextstrain-curated 'open' dataset [13] that pulls from sequences shared to NCBI GenBank. Raw sequence data is available from data.nextstrain.org. Here, we subsetted to sequences with specimens collected from the USA between January 1, 2021 and October 1, 2021. We additionally dropped 80 sequences without an assigned Nextstrain clade. This subsetting resulted in 952,091 sequences for analysis. However, we reduced dataset to just the 36 states with more 5000 sequences available in this timeframe. Doing so reduced the full dataset to 801,435 sequences for analysis.

Derived data of sequence counts and case counts, along with all source code used to analyze this data and produce figures is available via the GitHub repository github.com/blab/rt-from-frequency-dynamics.

Competing interests

The authors declare no conflicting interests.

Author contributions

MF, TB conceived the study. TB gathered sequence and case count data. MF designed and implemented inference model. MF performed the analysis. MF, TB interpreted the results. MF, TB wrote the paper.

Acknowledgements

We thank John Huddleston, Eslam Abousamra and other members of the Bedford Lab for helpful feedback. MF is an ARCS Foundation scholar and was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1762114. TB is an Investigator of the Howard Hughes Medical Institute. This project was supported by funds from the HHMI COVID-19 Collaboration Initiative awarded to the Fred Hutchinson Cancer Research Center and the University of Washington.

References

[1] Tao K, Tzou PL, Nouhin J, Gupta RK, de Oliveira T, et al. (2021) The biological and clinical significance of emerging SARS-CoV-2 variants. Nat Rev Genet 22: 757–773.

[2] Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, et al. (2021) Detection of a SARS-CoV-2 variant of concern in South Africa. Nature : 438–443.

[3] Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, et al. (2021) Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science : eabg3055.

[4] Annavajhala MK, Mohri H, Wang P, Nair M, Zucker JE, et al. (2021) Emergence and expansion of SARS-CoV-2 b.1.526 after identification in new york. Nature 597: 703–708.

[5] Faria NR, Mellan TA, Whittaker C, Claro IM, da S Candido D, et al. (2021) Genomics and epidemiology of the p.1 SARS-CoV-2 lineage in manaus, brazil. Science 372: 815–821.

[6] Obermeyer F, Schaffner SF, Jankowiak M, Barkas N, Pyle JD, et al. (2021) Analysis of 2.1 million SARS-CoV-2 genomes identifies mutations associated with transmissibility. medRxiv .

[7] Ito K, Piantham C, Nishiura H (2021) Predicted dominance of variant delta of sars-cov-2 before tokyo olympic games, japan, july 2021. Eurosurveillance 26: 2100570.

[8] Cori A, Ferguson NM, Fraser C, Cauchemez S (2013) A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology 178: 1505–1512.

[9] Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, et al. (2020) Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Research 5: 112.

[10] Konings F, Perkins MD, Kuhn JH, Pallen MJ, Alm EJ, et al. (2021) SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. Nature Microbiology .

[11] Earnest R, Uddin R, Matluk N, Renzette N, Siddle KJ, et al. (2021) Comparative transmissibility of sars-cov-2 variants delta and alpha in new england, usa. medRxiv .

[12] Vöhringer HS, Sanderson T, Sinnott M, Maio ND, Nguyen T, et al. (2021) Genomic reconstruction of the SARS-CoV-2 epidemic in england. Nature .

[13] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, et al. (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34: 4121–4123.

[14] Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, et al. (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology 5: 1403–1407.

[15] Uriu K, Kimura I, Shirakawa K, Takaori-Kondo A, Nakada Ta, et al. (0) Neutralization of the sars-cov-2 mu variant by convalescent and vaccine serum. New England Journal of Medicine 0: null.

[16] Vohringer HS, Sanderson T, Sinnott M, De Maio N, Nguyen T, et al. (2021) Genomic reconstruction of the SARS-CoV-2 epidemic in England. Nature .

[17] Wong F, Collins JJ (2020) Evidence that coronavirus superspreading is fat-tailed. Proceedings of the National Academy of Sciences 117: 29416–29418.

[18] Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, et al. (2020) Practical considerations for measuring the effective reproductive number, rt. PLOS Computational Biology 16: 1-21.

[19] Müller NF, Wagner C, Frazar CD, Roychoudhury P, Lee J, et al. (2021) Viral genomes reveal patterns of the sars-cov-2 outbreak in washington state. Science Translational Medicine 13: eabf0202.

[20] Sharma M, Mindermann S, Rogers-Smith C, Leech G, Snodin B, et al. (2021) Understanding the effectiveness of government interventions against the resurgence of COVID-19 in europe. Nature Communications 12.

[21] Bellman R, Harris TE (1948) On the theory of age-dependent stochastic branching processes. Proceedings of the National Academy of Sciences 34: 601–604.

[22] Champredon D, Dushoff J, Earn DJD (2018) Equivalence of the erlang-distributed SEIR epidemic model and the renewal equation. SIAM Journal on Applied Mathematics 78: 3258–3278.

[23] Ganyani T, Kremer C, Chen D, Torneri A, Faes C, et al. (2020) Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, march 2020. Eurosurveillance 25.

[24] Cheng C, Zhang D, Dang D, Geng J, Zhu P, et al. (2021) The incubation period of COVID-19: a global meta-analysis of 53 studies and a chinese observation study of 11 545 patients. Infectious Diseases of Poverty 10.

[25] Faulkner JR, Minin VN (2018) Locally adaptive smoothing with markov random fields and shrinkage priors. Bayesian Analysis 13.

[26] Phan D, Pradhan N, Jankowiak M (2019) Composable effects for flexible and accelerated probabilistic programming in numpyro. arXiv preprint arXiv:191211554 .

[27] Hoffman MD, Blei DM, Wang C, Paisley J (2013) Stochastic variational inference. J Mach Learn Res 14: 1303–1347.

[28] Kingma DP, Ba J (2017). Adam: A method for stochastic optimization. 1412.6980.

[29] Wallinga J, Lipsitch M (2006) How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings of the Royal Society B: Biological Sciences 274: 599–604.

Supplemental Appendix

Supplemental Results

Relationship to multinomial logistic regression

Other papers have tried to infer growth advantages of variants from sequence data alone, we show that the multinomial logistic regression model typically used in these analysis is roughly equivalent to our fixed growth advantage model, but that inferring relative effective reproduction numbers between variants using multinomial logistic regression requires additional restrictions on the generation time. Multinomial logistic regression typically models the probability of a given observation belong to class v as

$$\text{Prob}(X = v) = f_v = \frac{p_v \exp(r_v t)}{\sum_{1 \leq u \leq V} p_u \exp(r_u t)}. \tag{12}$$

For our purpose, we can assume this probability is equivalent to the true frequency of variant v in the population and in this case, p_v is considered to be related to the prevalence on variant v in the population at $t = 0$ and r_v can be considered to be the growth advantage relative to a pivot class u_* which has $r_{k_*} = 1$. In order to see the connection between this above model and ours, we return to the original renewal equation of the form

$$I(t) = R_t \int_0^t I(t - \tau) g(\tau). \tag{13}$$

Assuming that g is a point mass at a mean generation time T_g, we have that

$$I(nT_g) = \left(\prod_{i=1}^{n} R_{iT_g} \right) I(0). \tag{14}$$

Assuming that there are several variants following these same dynamics, we have that the frequency of a given variant v can be written as

$$f_v(nT_g) = \frac{I_v(nT_g)}{\sum_{1 \leq u \leq V} I_u(nT_g)}. \tag{15}$$

If we assume a constant growth advantage as in our model, we then have that $R_{t,v} = \Delta_v R_t$, so that

$$f_v(nT_g) = \frac{\Delta_v^n I_v(0)}{\sum_{1 \leq u \leq V} \Delta_u^n I_u(0)}. \tag{16}$$

Writing $\Delta_v = \exp(\delta_v)$ and $t = nT_g$, allows us to see that

$$f_v(t) = \frac{I_v(0) \exp(\frac{\delta_v}{T_g} t)}{\sum_{1 \leq u \leq V} I_u(0) \exp(\frac{\delta_u}{T_g} t)}. \tag{17}$$

By fixing one pivot class so that $I_{u_*} = \delta_{u_*}/T_g = 1$, we can identify our model with the multinomial logistic regression by relating the parameters as

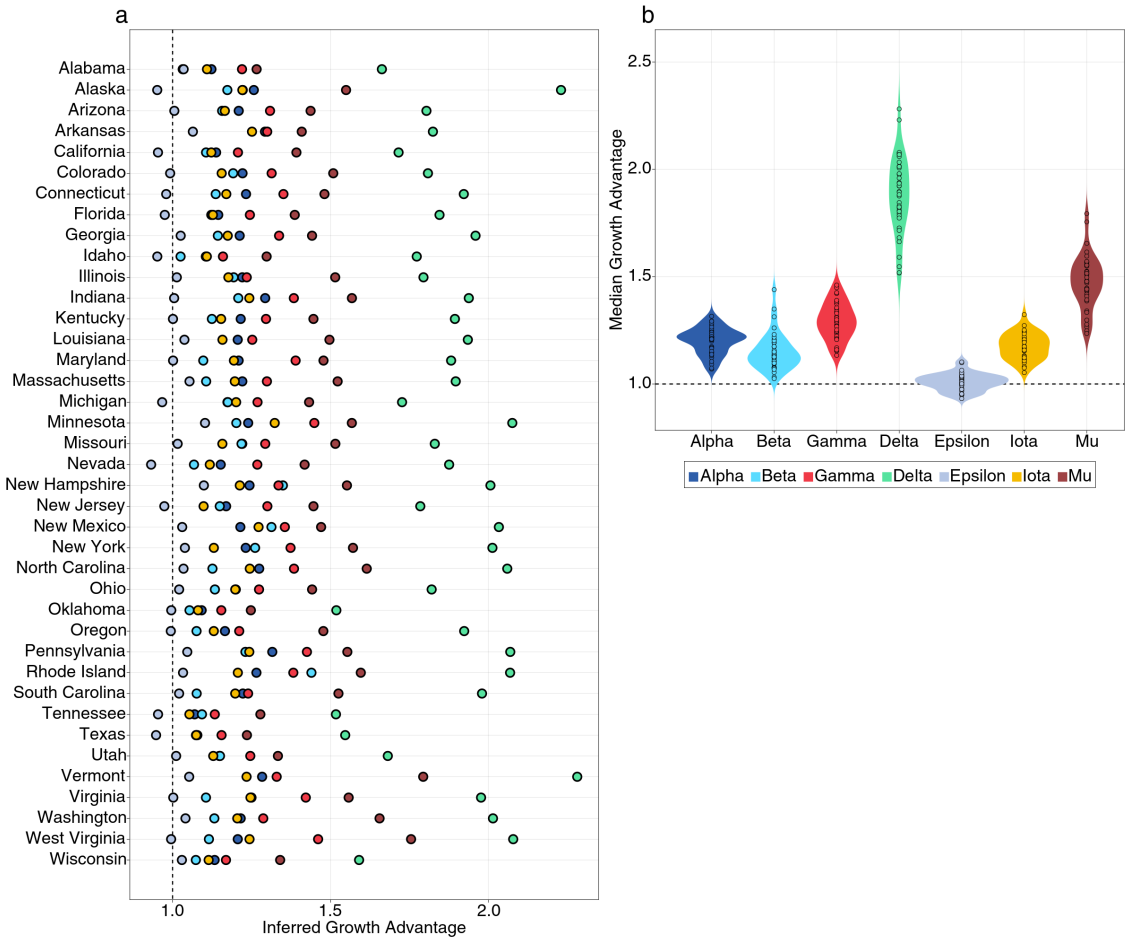

Figure S1. Assess growth advantages in various states using Multinomial Logistic Regression model assuming generation time $T_g = 5.2$. (a) Growth advantages visualized by state. (b) Same as (a) but grouped by variant.

$$\delta_v = r_v T_g \tag{18}$$

$$I_v(0) = p_v. \tag{19}$$

This shows that the multinomial logistic regression functions similarly to our fixed growth advantage model except with the additional assumption that the generation time is a point mass at T_g. This assumption additionally allows us to relate the epidemic growth rate r and the effective reproduction number as $R = \exp(rT_g)$ [29]. This means that the relative effective reproduction number for any two variants can be written as

$$\ln\left(\frac{R_{t,v}}{R_{t,u}}\right) = (r_v - r_u)T_g.$$

Relating epidemic growth rates to relative effective reproduction numbers

An important relationship of interest is between the epidemic growth rate of an epidemic and its effective reproduction number. In the case of our analysis, we are particularly interested in the ratios of variant-specific effective reproduction numbers. First, notice that the effective reproduction number and the epidemic growth rate of an epidemic are related by

$$R_t = \frac{1}{\int_0^\infty \exp(-r\tau)g(\tau)d\tau} = \frac{1}{M_g(-r)}$$

according to the Lotka-Euler equation [29] where r is the epidemic growth rate and M_g is the moment-generating function of the generation time g. This allows us to write the relative reproduction number of two variants v and u as a function of their epidemic growth rates, so that

$$\frac{R_{t,v}}{R_{t,u}} = \frac{M_g(-r_u)}{M_g(-r_v)}.$$

We'll consider three common generation time assumptions. First, we consider the case where the generation time is a point mass at T_g. In which case, $M_g(-r) = \exp(-rT_g)$ and we recover the relationship

$$\frac{R_{t,v}}{R_{t,u}} = \exp((r_v - r_u)T_g).$$

In this case, the relative effective reproduction number depends on only the difference between the epidemic growth rates and therefore, is commonly used when converting epidemic growth rates to relative reproduction numbers in the case of logistic growth models.

Second, we consider the case where the the generation time is an exponential distribution with mean T_g. This assumption is often implicit and common in models of infectious diseases such as ODEs and their stochastic variants. Using the corresponding moment-generating function, we see that

$$\frac{R_{t,v}}{R_{t,u}} = \frac{1 + r_v T_g}{1 + r_u T_g}.$$

Next, we consider the Gamma distributed generation times with mean T_g and standard deviation s. This is often used in models of infectious diseases via the chain trick in which multiple compartments are chained together to obtain non-exponential generation or waiting times. Re-parameterizing the Gamma distribution in terms of its mean and standard deviation, we have that

$$\frac{R_{t,v}}{R_{t,u}} = \left(\frac{1 + r_v\left(\frac{s^2}{T_g}\right)}{1 + r_u\left(\frac{s^2}{T_g}\right)}\right)^{T_g^2/s^2}.$$

From this equation, we can see that increases in the mean of the generation time lead to higher inferred variant advantages given the same growth rates. On the other hand, increases in the standard deviation lead to lower inferred variant advantages. This effect is also visualized in Figure S2.

Taking a logarithm, we can also evaluate the sensitivity of our inferred growth advantages from our fixed growth advantage model with respect to the generation time assuming it is Gamma distributed as

$$\delta_v = \ln\left(\frac{R_{t,v}}{R_{t,0}}\right) = \left(\frac{T_g^2}{s^2}\right) \ln\left(\frac{1 + r_v\left(\frac{s^2}{T_g}\right)}{1 + r_0\left(\frac{s^2}{T_g}\right)}\right).$$

The behavior here is analogous to that discussed above when the mean T_g and standard deviation s are changed although these growth advantages appear to relatively stable under varying standard deviation in Figure S4. Although the effective reproduction number and the growth advantage appear to have strong dependence on generation time parameters, we find that the epidemic growth rate r is more robust to changes in generation time (see Figure S3).

The cases of exponential and Gamma-distributed generation times highlight that for non-deterministic generation times there is no guarantee that the relative effective reproduction number depends on only the difference in epidemic growth rates. In fact, these estimates based on the deterministic generation times correspond to the case in which the standard deviation shrinks zero, they are likely overestimates of variant advantages given the observed variation in the serial interval of SARS-CoV-2 infections.

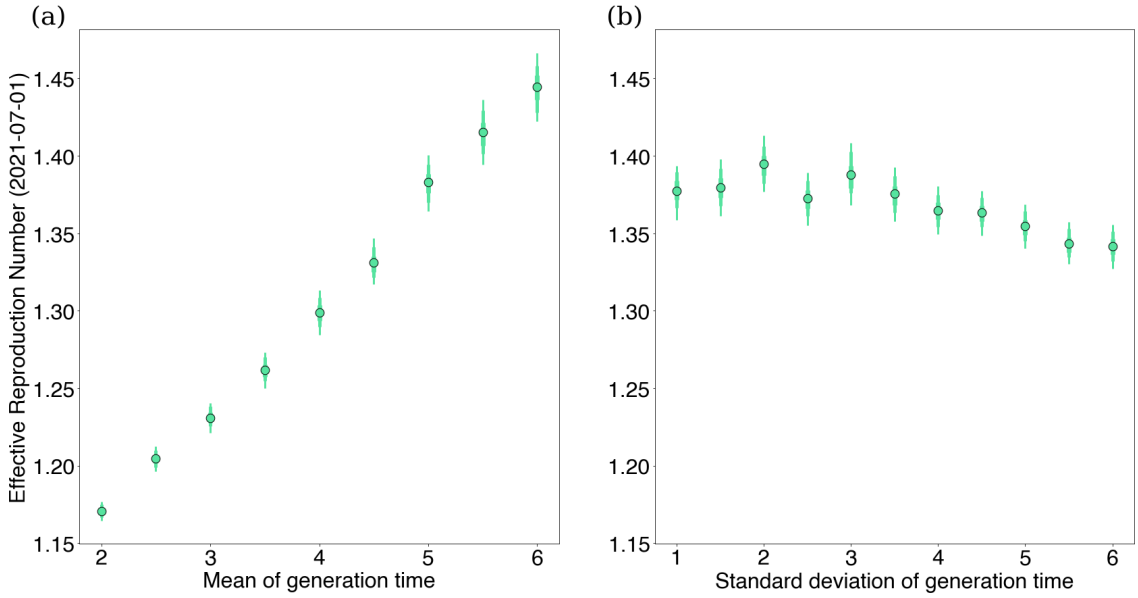

Figure S2. Sensitivity of effective reproduction number to changes in generation time. (a) We vary the mean of the generation time keeping a constant standard deviation 1.72 and plot against effective reproduction number estimates for Delta in Washington state on July 1st, 2021 using our free R_t model. (b) The same as (a), but we instead vary the standard deviation of the generation time keeping a constant mean 5.2.

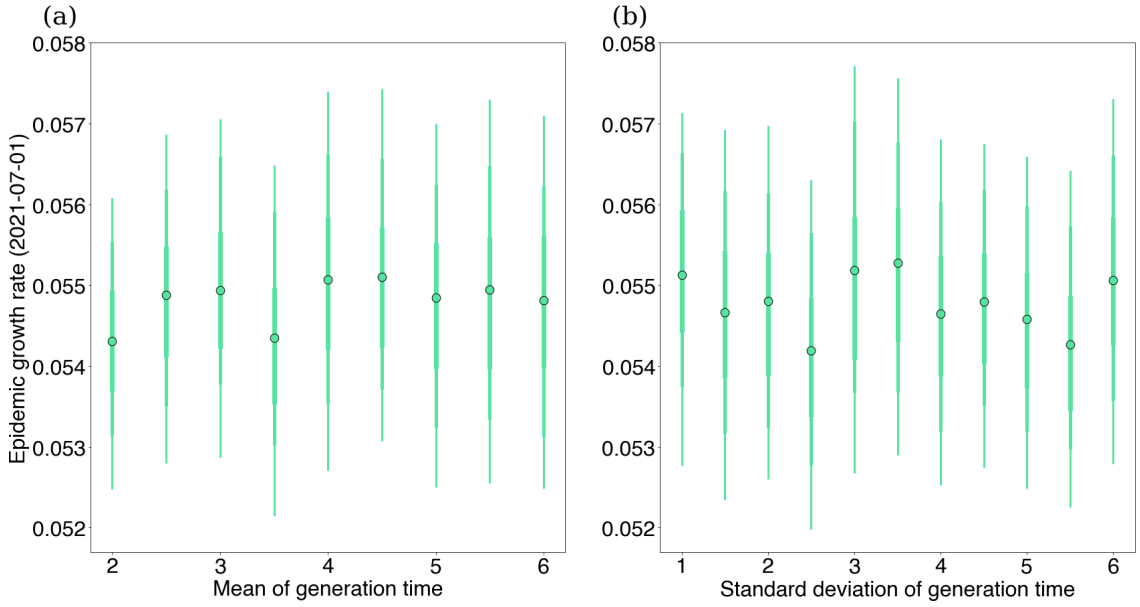

Figure S3. Sensitivity of epidemic growth rates to changes in generation time. (a) We vary the mean of the generation time keeping a constant standard deviation 1.72 and plot against exponential growth rates for Delta in Washington state on July 1st, 2021 using our free R_t model and assuming a Gamma-distributed generation time. (b) The same as (a), but we instead vary the standard deviation of the generation time keeping a constant mean 5.2.

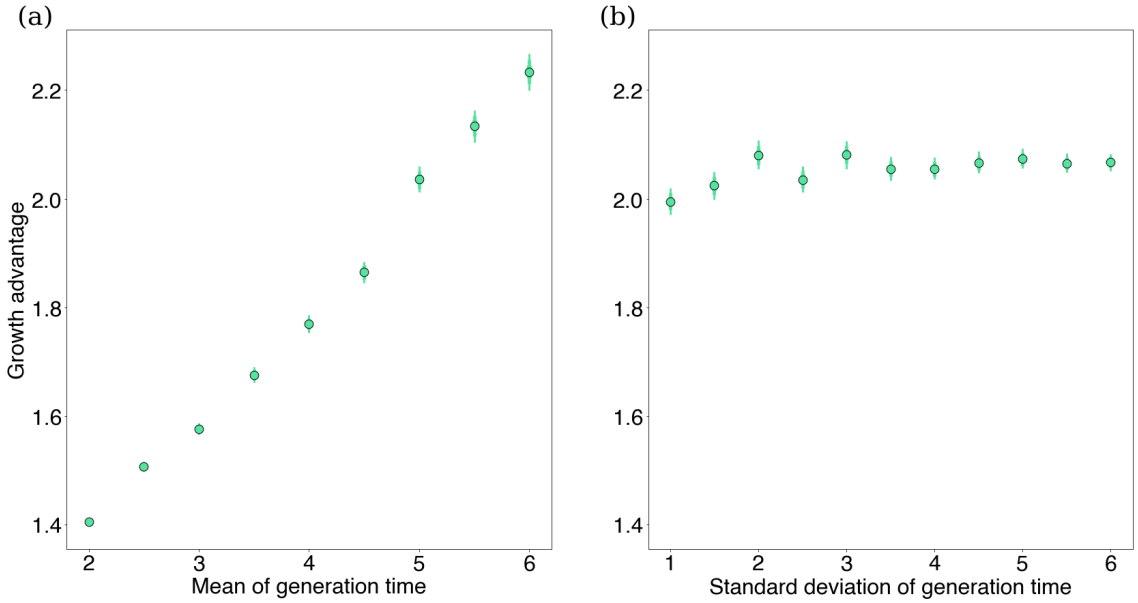

Figure S4. Sensitivity of growth advantages to changes in generation time. (a) We vary the mean of the generation time keeping a constant standard deviation 1.72 and plot against exponential growth rates for Delta in Washington state on July 1st, 2021 using our fixed growth model. (b) The same as (a), but we instead vary the standard deviation of the generation time keeping a constant mean 5.2.

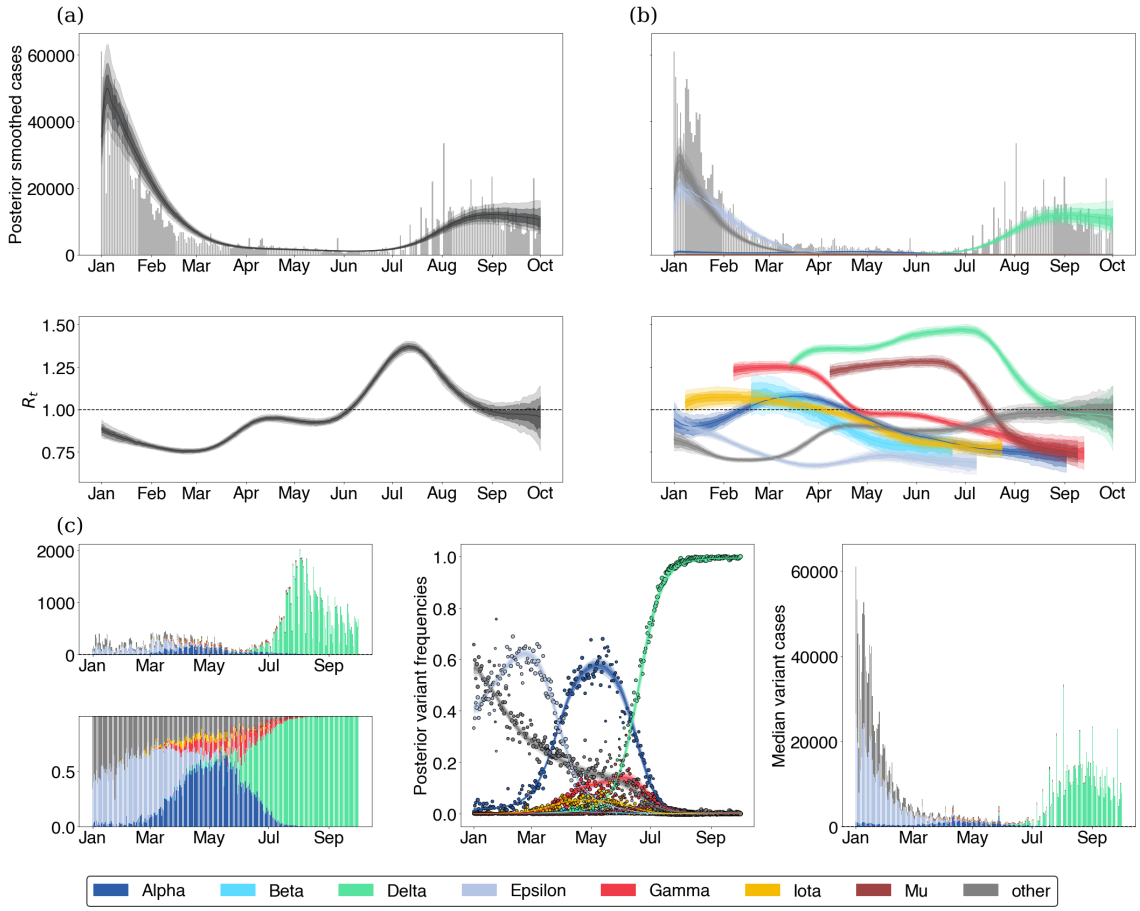

Figure S5. Fitting the free R_t model to California data.

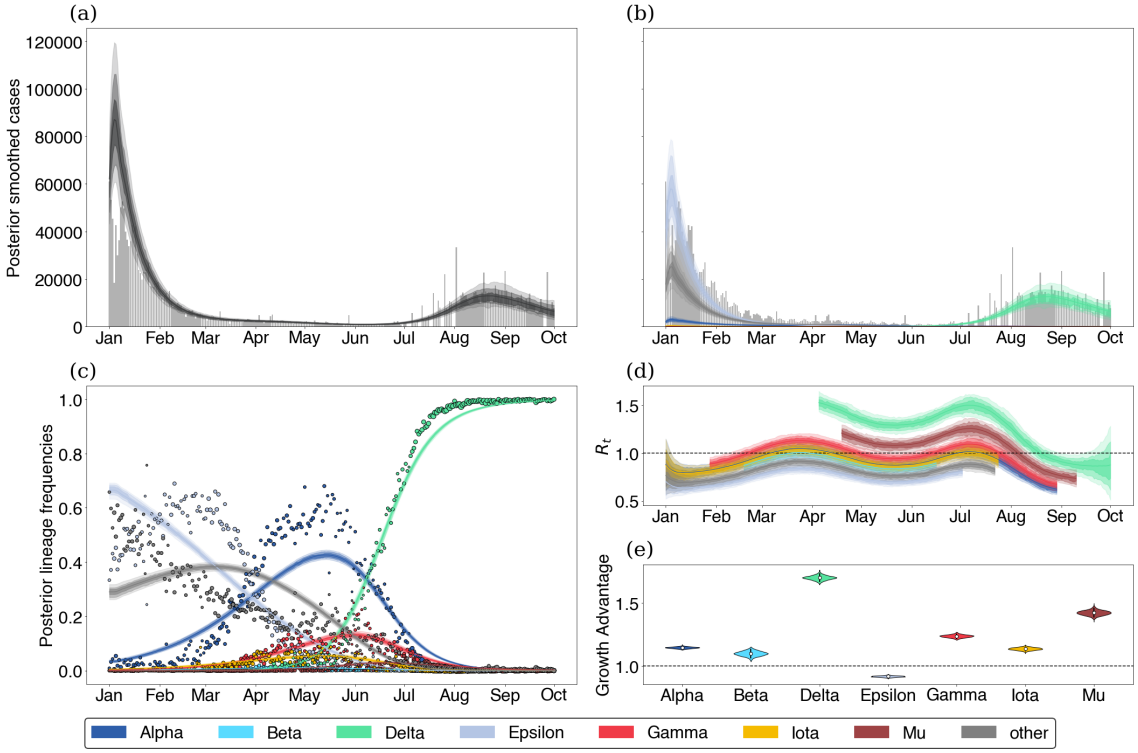

Figure S6. Fitting the fixed growth advantage model to California data.

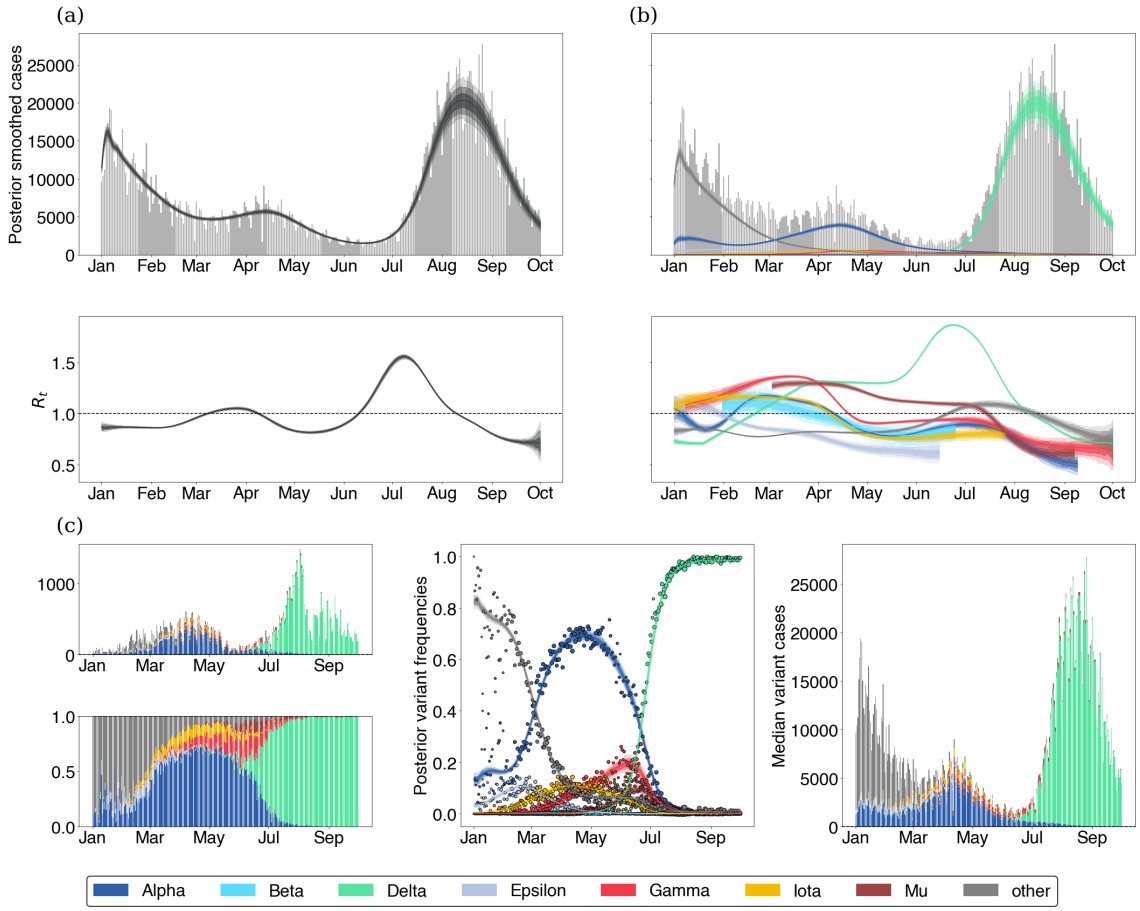

Figure S7. Fitting the free R_t model to Florida data.

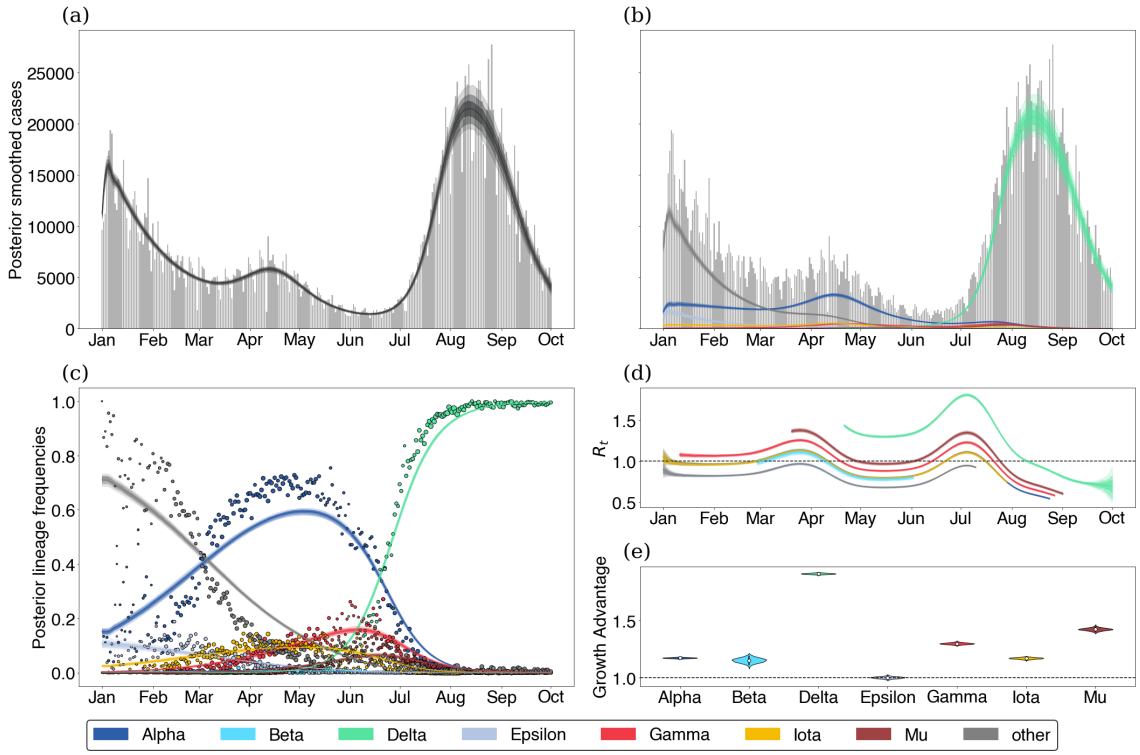

Figure S8. Fitting the fixed growth advantage model to Florida data.

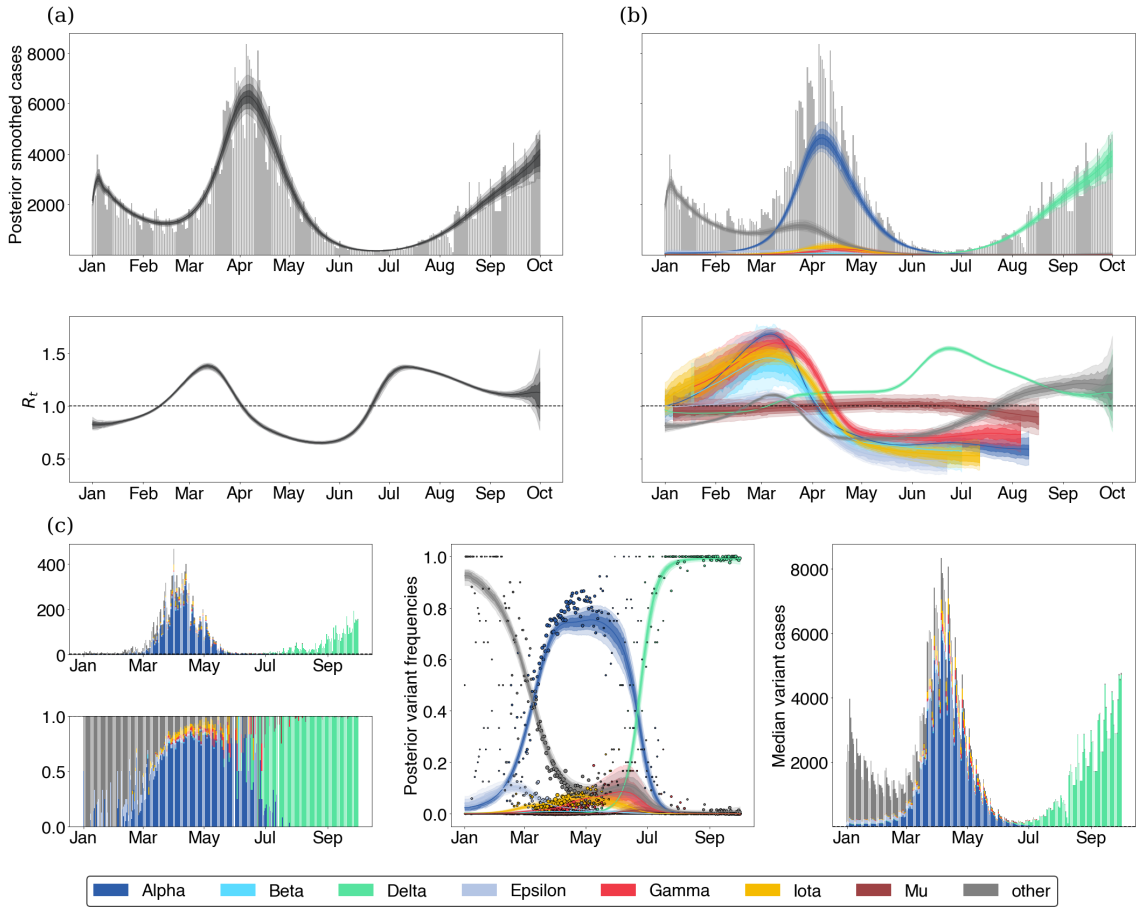

Figure S9. Fitting the free R_t model to Michigan data.

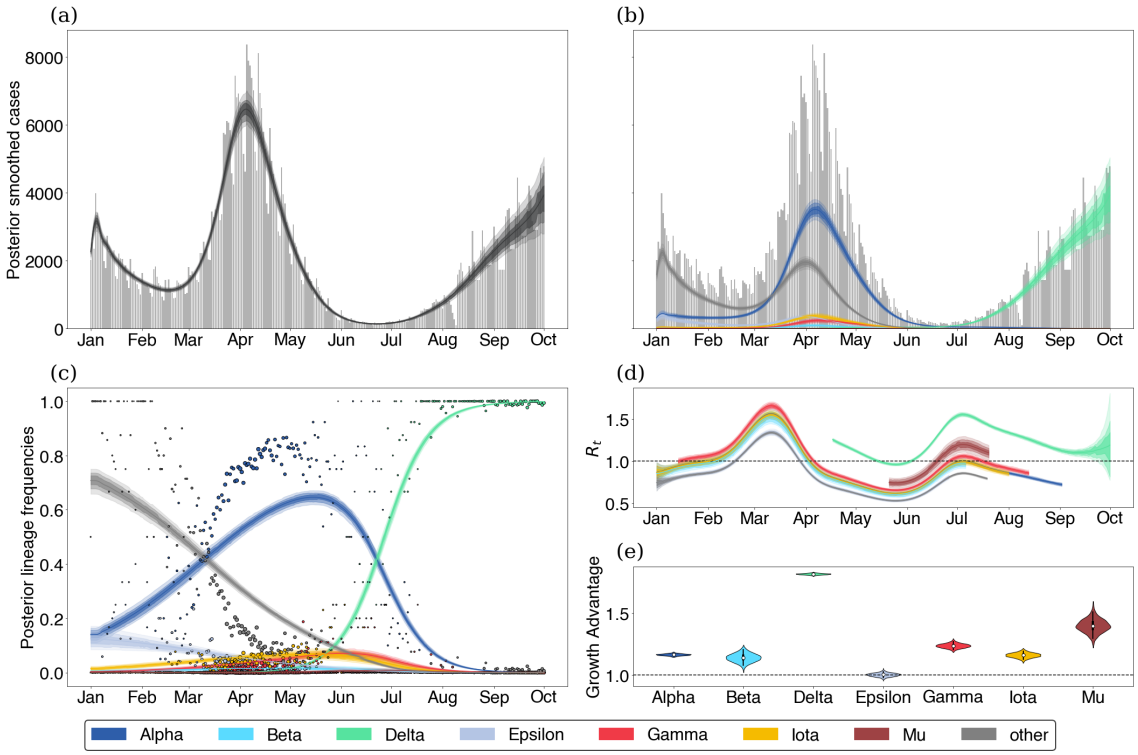

Figure S10. Fitting the fixed growth advantage model to Michigan data.

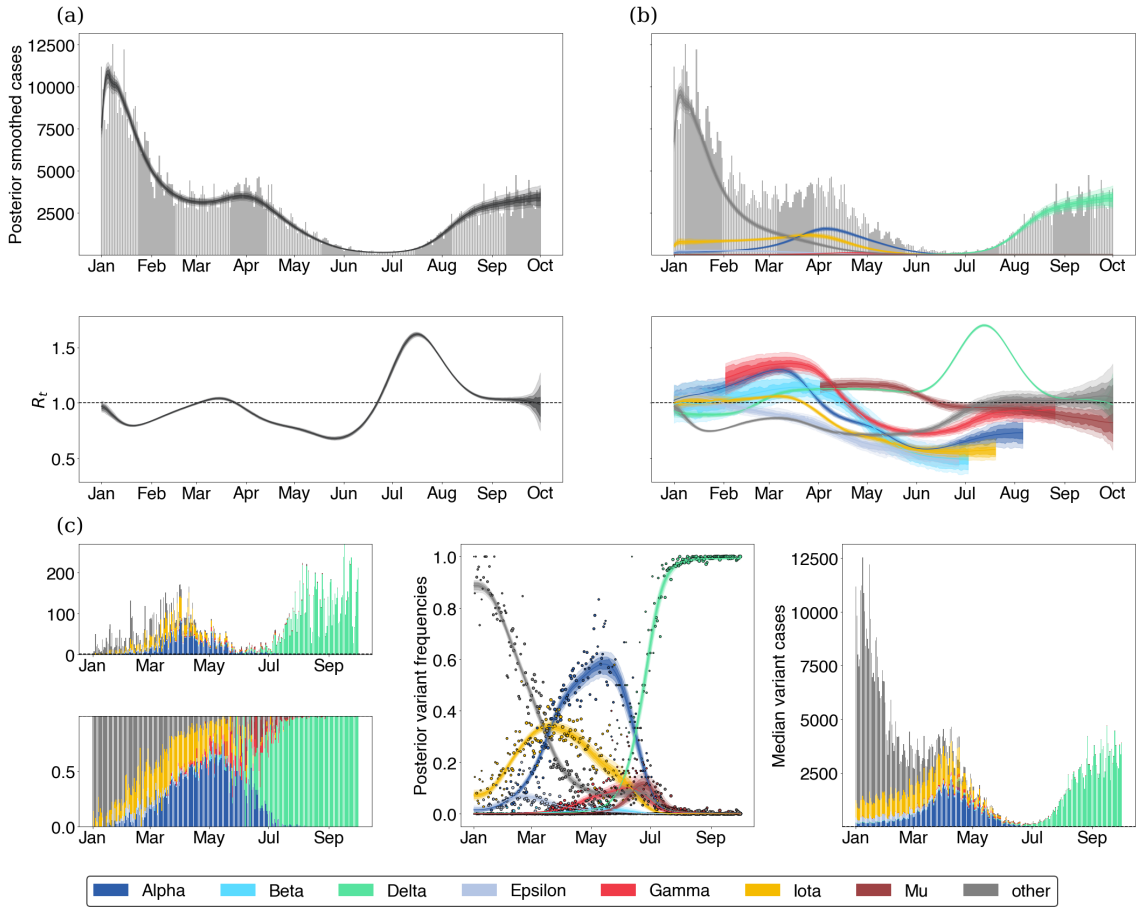

Figure S11. Fitting the free R_t model to New York state data.

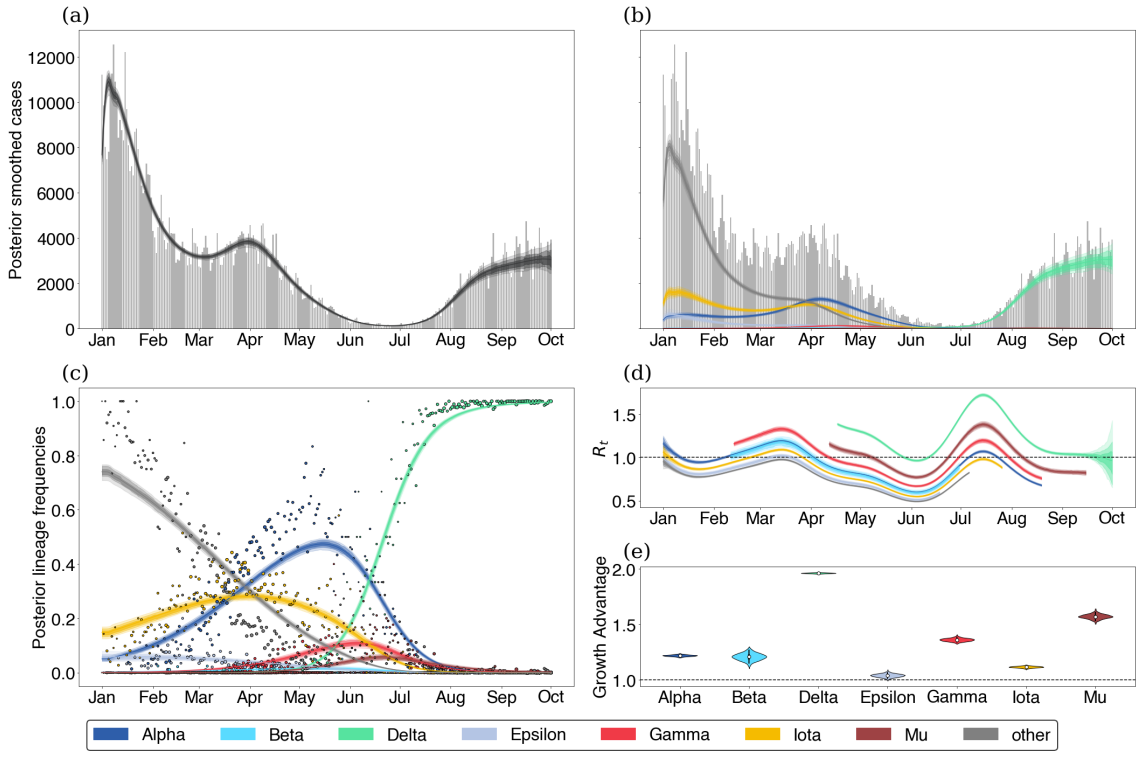

Figure S12. Fitting the fixed growth advantage model to New York state data.