Lineage frequency time series reveal elevated levels of genetic drift in SARS-CoV-2 transmission in England

QinQin Yu[1,9,*], Joao Ascensao[2,‡], Takashi Okada[1,3,4,5,‡], The COVID-19 Genomics UK (COG-UK) consortium[6*], Olivia Boyd[7], Erik Volz[6,7], and Oskar Hallatschek[1,3,8,†]

[1]Department of Physics, University of California, Berkeley, United States
[2]Department of Bioengineering, University of California, Berkeley, United States
[3]Department of Integrative Biology, University of California, Berkeley, United States
[4]Institute for Life and Medical Sciences, Kyoto University, Kyoto, Japan
[5]RIKEN iTHEMS, Wako, Saitama, Japan
[6]https://www.cogconsortium.uk
[7]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom
[8]Peter Debye Institute for Soft Matter Physics, Leipzig University, Leipzig, Germany
[9]Present affiliation: Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States
[‡]These authors contributed equally.
[*]Full list of consortium names and affiliations are in the appendix

November 21, 2022

Abstract

Random genetic drift in the population-level dynamics of an infectious disease outbreak results from the randomness of inter-host transmission and the randomness of host recovery or death. The strength of genetic drift has been found to be high for SARS-CoV-2 due to superspreading, and this is expected to substantially impact the disease epidemiology and evolution. Noise that results from the measurement process, such as biases in data collection across time, geographical areas, etc., can potentially confound estimates of genetic drift as both processes contribute "noise" to the data. To address this challenge, we develop and validate a method to jointly infer genetic drift and measurement noise from time-series lineage frequency data. We apply this method to over 490,000 SARS-CoV-2 genomic sequences from England collected between March 2020 and December 2021 by the COVID-19 Genomics UK (COG-UK) consortium. We find that even after correcting for measurement noise, the strength of genetic drift is consistently, throughout time, higher than that expected from the observed number of COVID-19 positive individuals in England by 1 to 3 orders of magnitude. Corrections taking into account epidemiological dynamics (susceptible-infected-recovered or susceptible-exposed-infected-recovered models) do not explain the discrepancy. Moreover, the levels of genetic drift that we observe are higher than the estimated levels of superspreading found by modeling studies that incorporate data on actual contact statistics in England. We discuss how even in the absence of superspreading, high levels of genetic drift can be generated via community structure in the host contact network. Our results suggest that further investigations of heterogeneous host contact structure may be important for understanding the high levels of genetic drift observed for SARS-CoV-2 in England.

[*]qinqinyu@berkeley.edu
[†]ohallats@berkeley.edu

Introduction

Random genetic drift is the change in the composition of a population over time due to the randomness of birth and death processes. In pathogen transmission, births occur as a result of transmission of the pathogen between hosts and deaths occur as a result of infected host recovery or death. The strength of genetic drift in pathogen transmission is determined by the disease prevalence, the disease epidemiology parameters [1], the variance in offspring number (the number of secondary infections that result from an infected individual) [2], as well as host contact patterns [3]. Many diseases have been found to exhibit high levels of genetic drift, such as SARS, MERS, tuberculosis, and measles [2, 4, 5]. The strength of genetic drift affects how the disease spreads through the population [2, 3, 6] how new variants emerge [7, 8, 9, 10, 11], and the effectiveness of interventions [12], making it an important quantity to accurately estimate for understanding disease epidemiology, evolution, and control.

The effective population size is often used to quantify the strength of genetic drift; it is the population size in an idealized Wright-Fisher model (with discrete non-overlapping generations, a constant population size, and offspring determined by sampling with replacement from the previous generation) that would reproduce the observed dynamics [13]. If the effective population size is lower than the true population size, it is an indication that there are additional sources of stochasticity beyond random sampling with replacement; thus, a lower effective population size indicates a higher level of genetic drift.

Transmission of SARS-CoV-2 has been shown to exhibit high levels of superspreading (high variance in offspring number) [14, 15, 16] and high levels of genetic drift (low effective population sizes) [17, 18, 19] (see also Supplementary table S1). However, studies have focused on particular times and locations, and we lack systematic studies over time and space (see Ref. [20] for a recent first study that uses contact tracing data to infer changes in SARS-CoV-2 superspreading over time in Hong Kong). Performing a systematic study may be most feasible with a large-scale surveillance dataset, such as that from the COVID-19 Genomics UK (COG-UK) consortium, which has sequenced almost 3 million cases of SARS-CoV-2 in both surveillance and non-surveillance capacities as of October 5, 2022. We focus specifically on this dataset, and specifically on England, due to its consistently large number of sequenced SARS-CoV-2 cases since early in the pandemic.

A challenge to performing a systematic study of the strength of genetic drift for SARS-CoV-2 and other pathogens is how to handle measurement noise, or noise from the data collection process [21]. Measurement noise can arise from a variety of factors, including variability in the testing rate across time, geographic locations, demographic groups, and symptom status, and biases in contact tracing. Methods exist to infer measurement noise from time-series lineage or allele frequencies [22, 23, 24] (see the Supplementary information for a summary of other methods used for inferring genetic drift, and additional references). Note that here we use the term "lineage" to refer to a group of sequences that are genetically similar to one another, which are not necessarily the same as the lineages defined by the Pango nomenclature. Intuitively, in time-series frequency data, genetic drift leads to frequency fluctuations whose magnitudes scale with time, whereas measurement noise leads to frequency fluctuations whose magnitudes do not scale with time (Figure 1a). Thus, this system has been mapped onto a Hidden Markov Model (HMM) with continuous hidden and observed states (similar to a Kalman filter), where the hidden states are the true frequencies and the observed states are the observed frequencies (Figure 1b), and the processes of genetic drift and measurement noise determine the transition and emission probabilities, respectively [25, 26]. Methods often assume uniform sampling of infected individuals from the population [27, 22, 23], but this assumption does not usually hold outside of surveillance studies. A recent study accounted for overdispersed sampling of sequences in the inference of fitness coefficients of SARS-CoV-2 variants, but assumes constant overdispersion over time [28]; in reality, the observation process may change over time due to changes in testing intensity between locations and subpopulations.

In this study, we develop a method to jointly infer genetic drift and measurement noise that allows measurement noise to be overdispersed (rather than uniform) and for the strength of overdispersion to vary over time (rather than stay constant). By fitting this model to observed lineage frequency trajectories from simulations, we show that the effective population size and the strength of measurement noise can be accurately determined in most situations, even when both quantities are varying over time. We then apply our validated method to estimate the strengths of genetic drift and measurement noise for SARS-CoV-2 in England across time (from March 2020 until December 2021) and space using over 490,000 SARS-CoV-2 genomic sequences from COG-UK. We find high levels of genetic drift for SARS-CoV-2 consistently

throughout time that cannot be explained by literature values of superspreading. We discuss how community structure in the host contact network may partially explain these results. Additionally, we observe that sampling of infected individuals from the population is mostly uniform for this dataset, and we also find evidence of spatial structure in the transmission dynamics of B.1.177, Alpha, and Delta.

Results

Method for jointly inferring genetic drift and measurement noise from time-series lineage frequency data

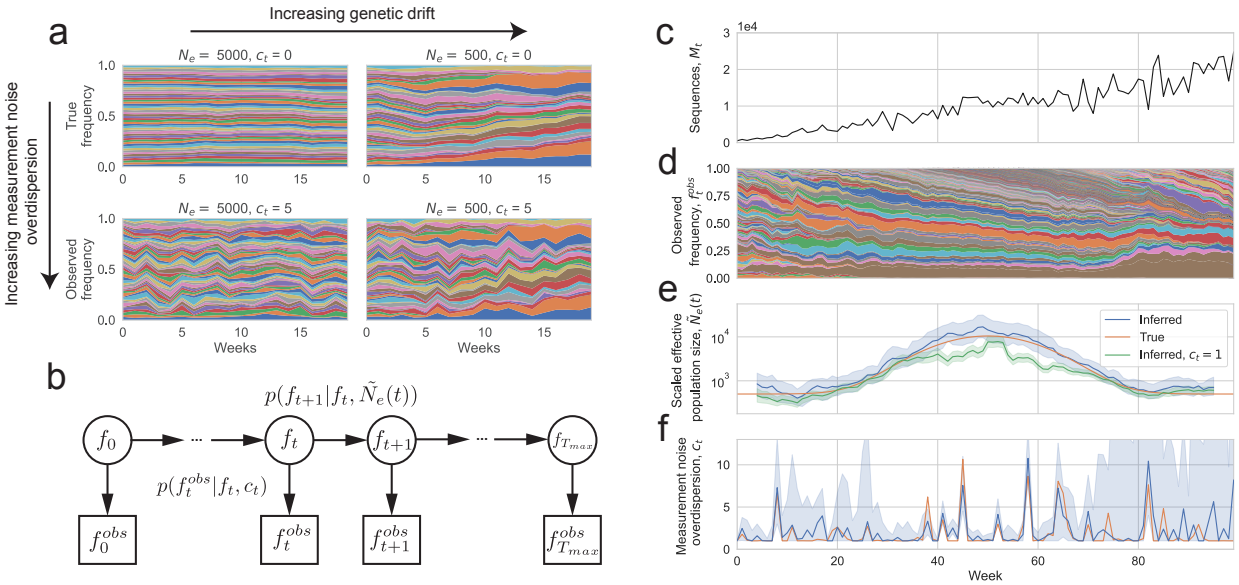

Figure 1: A Hidden Markov Model with continuous hidden and observed states (similar to a Kalman filter) for inferring genetic drift and measurement noise from lineage frequency time series. (a) Illustration of how genetic drift and measurement noise affect the observed frequency time series. Muller plot of lineage frequencies from Wright-Fisher simulations with effective population size 500 and 5000, with and without measurement noise. In simulations with measurement noise, 100 sequences were sampled per week with the measurement noise overdispersion parameter $c_t = 5$ (parameter defined in text). All simulations were initialized with 50 lineages at equal frequency. A lower effective population size leads to larger frequency fluctuations whose variances add over time, whereas measurement noise leads to increased frequency fluctuations whose variances do not add over time. (b) Schematic of Hidden Markov Model describing frequency trajectories. f_t is the true frequency at time t (hidden states) and f_t^{obs} is the observed frequency at time t (observed states). The inferred parameters are $\tilde{N}_e(t) \equiv N_e(t)\tau(t)$, the effective population size scaled by the generation time, and c_t, the overdispersion in measurement noise ($c_t = 1$ corresponds to uniform sampling of sequences from the population). (c-f) Validation of method using Wright-Fisher simulations of frequency trajectories with time-varying effective population size and measurement noise. (c) Simulated number of sequences. (d) Simulated lineage frequency trajectories. (e) Inferred scaled effective population size ($\tilde{N}_e(t)$) on simulated data compared to true values. (f) Inferred measurement noise (c_t) on simulated data compared to true values. In (e) the shaded region shows the 95% confidence interval calculated using the posterior, and in (f) the shaded region shows the 95% confidence interval calculated using bootstrapping (see Methods).

We first summarize the statistical inference method that we developed to infer time-varying effective population sizes from neutral lineage frequency time series that are affected by overdispersed measurement noise (more variable than uniform sampling). We explain the method more extensively in the Methods.

Briefly, we use a Hidden Markov Model (HMM) with continuous hidden and observed states (similar to a Kalman filter), where the hidden states are the true frequencies (f_t, where t is time), and the observed states are the observed frequencies (f_t^{obs}) (Figure 1b) (see Methods). The transition probability between hidden states is set by genetic drift, where the mean true frequency is the true frequency at the previous time $\mathrm{E}(f_{t+1}|f_t) = f_t$, and when the frequencies are rare the variance in frequency is proportional to the mean, $\mathrm{Var}(f_{t+1}|f_t) = \frac{f_t}{\tilde{N}_e(t)}$. $\tilde{N}_e(t) = N_e(t)\tau(t)$ where $N_e(t)$ is the effective population size and $\tau(t)$ is the generation time, and both quantities can vary over time; however, we are only able to infer the compound parameter $N_e(t)\tau(t)$. The emission probability between hidden and observed states is set by measurement noise, where the mean observed frequency is the true frequency $\mathrm{E}(f_t^{obs}|f_t) = f_t$ and when the frequencies are rare the variance in the observed frequency is proportional to the mean, $\mathrm{Var}(f_t^{obs}|f_t) = c_t \frac{f_t}{M_t}$. $c_t \geq 1$ describes the time-varying deviation from uniform sampling ($c_t = 1$), and M_t is the number of sequences at time t. Our model assumes that the number of individuals and frequency of a lineage is high enough such that the central limit theorem applies; to meet this condition, we created "superlineages" where we randomly and exclusively grouped lineages together such that the sum of their abundances and frequencies was above a threshold (see Methods).

Using the transition and emission probability distributions (see Methods) and the HMM structure, we determine the likelihood function (Equation 13 in Methods) describing the probability of observing a particular set of lineage frequency time-series data given the unknown parameters, namely the scaled effective population size across time $\tilde{N}_e(t)$ and the strength of measurement noise across time c_t. We then maximize the likelihood over the parameters to determine the most likely parameters that describe the data. Because we are relying on a time-series signature in the data for the inference, we need to use a sufficiently large number of timesteps of data, but on the other hand, the longer the time series, the more parameters would need to be inferred (since both $\tilde{N}_e(t)$ and c_t are allowed to change over time). To balance these two factors, we assumed that the effective population size stays constant over a time period of 9 weeks (a form of "regularization"). We then shift this window of 9 weeks across time to determine how $\tilde{N}_e(t)$ changes over time (see Methods), but this effectively averages the inferred $\tilde{N}_e(t)$ over time. c_t is still allowed to vary weekly.

To validate our model, we ran Wright-Fisher simulations with time-varying effective population size and time-varying measurement noise (Figure 1c-f). Because a substantial number of lineages would go extinct over the simulation timescale of 100 weeks, we introduced new lineages with a small rate (a rate of 0.01 per week per individual of "mutating" to start a new lineage) to prevent the number of lineages from becoming too low. We then did inference on the simulated time-series frequency trajectories (Figure 1d). The inferred $\tilde{N}_e(t)$ and c_t closely follow the true values (Figure 1e-f), and the 95% confidence intervals (see Methods for how they are calculated) include the true value in a median (across timepoints) of 95% of simulation realizations (Figure S5). The error in c_t is higher when the variance contributed to the frequency trajectories by measurement noise is lower than that of genetic drift, which occurs when the effective population size is low or number of sequences is high (more clearly seen in Figure S6, where the effective population size is held constant). However, the error on $\tilde{N}_e(t)$ seems to be unchanged or even slightly decrease when the error on c_t is increased because the contribution to the variance due to genetic drift is higher. We also observe that the inferred $\tilde{N}_e(t)$ is smoothed over time due to the assumption of constant $\tilde{N}_e(t)$ over 9 weeks (Figure S7); this is a potential drawback when there are sharp changes in the effective population size over time. Importantly, we observed that the inferred $\tilde{N}_e(t)$ will be underestimated if sampling is assumed to be uniform when it is actually overdispersed (Figure 1e). This is because variance in the frequency trajectories due to measurement noise is incorrectly being attributed to genetic drift. The underestimation is strongest when the variance contributed due to measurement noise is high, either due to high measurement noise overdispersion, a low number of sampled sequences, or a high effective population size. In this situation, joint inference of measurement noise and $\tilde{N}_e(t)$ from the data is necessary for accurate inference of $\tilde{N}_e(t)$.

In summary, we developed a method to infer the strength of genetic drift and measurement noise from lineage frequency time series data and validated the accuracy of the method with simulations.

Application to COG-UK data in England

We next applied this method to study the effective population size and strength of measurement noise for SARS-CoV-2 in England. Because our method assumes that lineages are neutral with respect to one another (no selection), we performed separate analyses on groups of lineages that have been shown to exhibit fitness

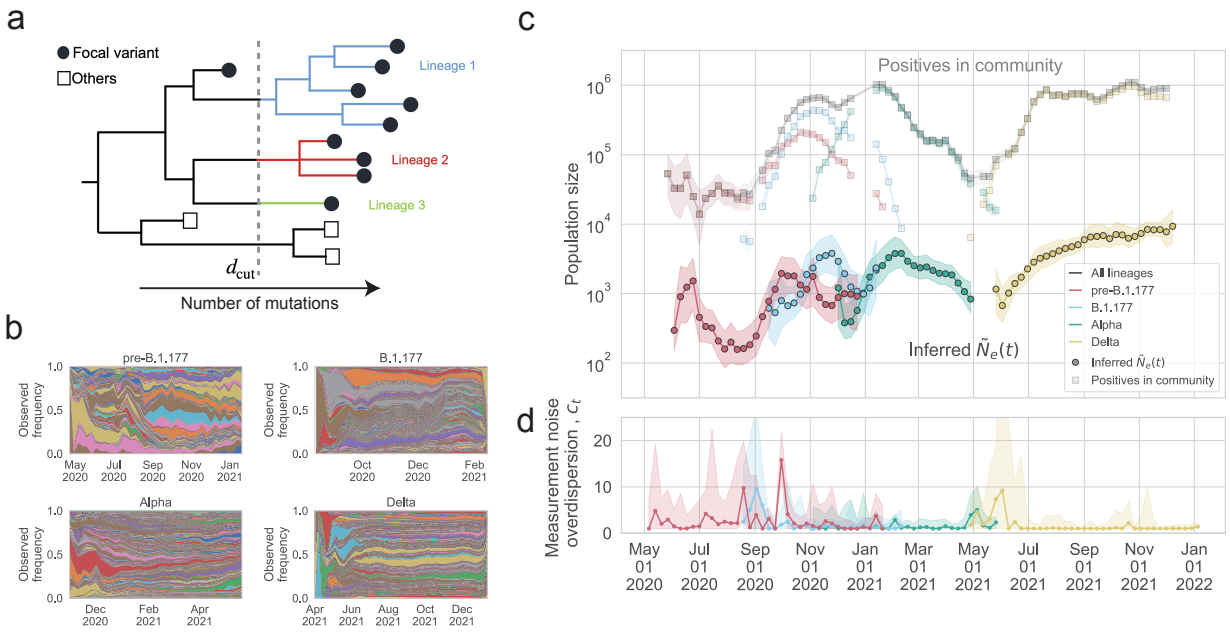

Figure 2: The inferred effective population size and overdispersion of measurement noise in England compared with the number of positive individuals. (a) Schematic of lineage construction for B.1.177, Alpha, and Delta from the COG-UK phylogenetic tree. The filled circles represent the sequences of a focal variant sampled in England, while the unfilled squares represent other sequences, which are of other variants or sampled in other countries. The phylogenetic tree is cut at a certain depth $d = d_{\text{cut}}$, and each branch cut by the line $d = d_{\text{cut}}$ defines a lineage. Lineages pre-B.1.1.7 are defined using the pango nomenclature [29, 30]. (b) Muller plot of lineage frequency time series for lineages pre-B.1.177, of B.1.177, of Alpha, of Delta. (c) Inferred scaled effective population size ($\tilde{N}_e(t) \equiv N_e(t)\tau(t)$) for pre-B.1.177 sequences, B.1.177, Alpha, and Delta, compared to the estimated number of people testing positive for SARS-CoV-2 in England at the community level, as measured by the COVID-19 Infection Survey [31], for all lineages and by variant or group of lineages. To simplify the plot, only data where the number of positive individuals for a given variant or group of lineages was higher than 10^3 in a week are shown. The inferred $\tilde{N}_e(t)$ is considerably lower than the number of positive individuals for all times and for all variants or group of lineages. (d) Inferred measurement noise overdispersion (c_t) for pre-B.1.177 sequences, B.1.177, Alpha, and Delta.

differences or deterministic changes in frequency: lineages pre-B.1.177, B.1.177, Alpha, and Delta [28, 17, 32, 33]. We did not find any studies in the literature claiming detectable fitness differences between lineages within each of these groups; thus, we assumed that our neutral model should be valid when analyzing lineages only within a single group. We checked that this assumption is valid and describe the results below.

To obtain lineage frequency time series data for SARS-CoV-2 in England, we downloaded genomic metadata from the COVID-19 Genomics UK Consortium (COG-UK) [34] (Figure 2b) and the associated phylogenetic trees that were created at different points in time. To minimize potential bias, we used only surveillance data (labeled as "pillar 2"). For sequences pre-B.1.177, we used the pangolin lineages assignments from COG-UK [29, 30]. However, B.1.177, Alpha, and Delta were subdivided into few or only one pangolin lineage, since a new lineage is defined by sufficiently many mutations and evidence of geographic importation. However, these requirements are not important for our purposes and instead we only need resolution of neutral lineages within a variant. Thus, we created our own neutral lineages by grouping sequences together based on phylogenetic distance in the tree (see Figure 2a and Methods), and cutting the tree at a particular point. Most sequenced samples were included in the trees (Figure S8), and any downsampling was done by preserving genetic diversity. Most sequences in the tree were assigned to lineages (see Methods), and we corrected for the fraction of sequences that were not assigned to lineages in our inference of $\tilde{N}_e(t)$

(see Methods). This yielded 486 lineages for pre-B.1.177, 4083 lineages for B.1.177, 6225 lineages for Alpha, 24867 lineages for Delta.

The inferred effective population size is shown in Figure 2c. The inferred effective population size was lower than the number of positive individuals in the community by a factor of 16 to 1055 at different points in time. The most notable differences between the changes over time in the number of positives in the community and that of the effective population size were: the inferred effective population size of lineages pre-B.1.177 peaked slightly before the number of pre-B.1.177 positives peaked, the inferred effective population size of Alpha decreased slower than the number of positives decreased after January 2021, and the shoulder for the inferred effective population size of Delta occurred earlier than in the number of positives. We checked that the inferred effective population size is not sensitive to the depth at which the trees are cut to create lineages (Figure S9), the threshold counts for creating superlineages (Figure S10), or the number of weeks in the moving time window (Figure S11).

The inferred measurement noise for each group of lineages is shown in Figure 2d. The inferred measurement noise overdispersion was mostly indistinguishable from 1 (uniform sampling), but at times was above 1 (sampling that is more variable than uniform sampling). There were also at times differences in the strength of measurement noise between variants when they overlapped in time. In particular, measurement noise for lineages pre-B.1.177 peaked in October 2020 despite measurement noise being low for B.1.177 at that time.

To better understand the observed levels of genetic drift, we compared the inferred $\tilde{N}_e(t)$ to that of an SIR null model, which includes a susceptible, infectious, and recovered class. The $\tilde{N}_e(t)$ for an SIR model was derived in Ref. [35, 36, 37] and is given by

$$\tilde{N_e}^{\text{SIR}}(t) = \frac{I(t)}{2R_t\gamma_I} \tag{1}$$

where $I(t)$ is number of infectious individuals, R_t is the effective reproduction number, and γ_I is the rate at which infectious individuals recover. For the number of infectious individuals, we used the number of positive individuals estimated from the UK Office for National Statistics' COVID-19 Infection Survey [31], which is a household surveillance study that reports positive PCR tests, regardless of symptom status. We used the measured effective reproduction number in England reported by the UK Health Security Agency [38]. We used $\gamma_I^{-1} = 5.5$ days [39, 40], and our results are robust to varying γ_I within a realistic range of values (Figure S12). We found that $\tilde{N_e}^{\text{SIR}}(t)$ is very similar to the number of positives because the effective reproduction number in England was very close to 1 across time and γ_I is also very close to 1 in units of weeks^{-1}. To calculate $\tilde{N_e}^{\text{SIR}}(t)$ for each variant or group of lineages, we rescaled the population-level $I(t)$ and R_t based on the fraction of each variant in the population and the relative differences in reproduction numbers between variants (see Methods). We then calculated the scaled true population size, $\tilde{N}(t) \equiv N(t)\tau(t)$, for the SIR model by multiplying by the variance in offspring number, σ^2, for the SIR model [41]

$$\tilde{N}^{\text{SIR}}(t) = \tilde{N_e}^{\text{SIR}}(t)\{\sigma^2\}^{\text{SIR}} \tag{2}$$

$$\{\sigma^2\}^{\text{SIR}} = 2. \tag{3}$$

Overall, the inferred $\tilde{N}_e(t)$ is lower than $\tilde{N}^{\text{SIR}}(t)$ by a time-dependent factor that varies between 16 and 589 (Figures 3c and S13), suggesting high levels of genetic drift in England across time. We find similar results when using an SEIR rather than an SIR model which additionally includes an exposed class and may be more realistic (Methods, Supplementary information, and Figure S14). The ratio of $\tilde{N}^{\text{SIR}}(t)$ to the inferred $\tilde{N}_e(t)$ was similar across variants anda cross time, except that for Alpha the ratio initially peaked and then decreased over time.

Because non-neutral lineages could potentially bias the the inferred effective population size to be lower in a model that assumes all lineages are neutral, we checked the assumption that lineages are neutral with respect to one another within a group or variant (pre-B.1.177, B.1.177, Alpha, and Delta) using methods that detect deterministic changes in lineage frequency. We used two methods: a more conservative, deterministic method that ignores genetic drift, and a more accurate method that accounts for time-varying genetic drift. To detect any lineages that could possibly be non-neutral, we used the conservative method, and found that about 20% of lineages in each group were significantly non-neutral at a significance level of 5%. Very likely, some of these lineages are detected as non-neutral simply because the model does not correctly account for

strong genetic drift. Excluding these lineages from the analysis of the inferred effective population size leads to slightly different values of the effective population size, but by less than an order of magnitude and mostly by less than the 95% confidence intervals (Figure S15). This result shows that conservatively excluding lineages that could be non-neutral does not change the result that the inferred effective population size is one to two order of magnitudes lower than the SIR or SEIR model effective population size. Additionally, when we used these new estimates of effective population size in the more accurate method for inferring fitness coefficients that accounts for time-varying effective population size, we find that only about 1% of lineages are non-neutral at a significance level of 5%, which further supports that lineages under constant selection are unlikely to explain our results.

We also probed the spatial structure of transmission by inferring the scaled effective population size separately for each region within England. We find that the scaled effective population size in the regions of England is substantially smaller than that in England as a whole for B.1.177, Alpha, and Delta (Figure S1), suggesting that the transmission was not well-mixed at that time. Additionally, the discrepancy between the inferred regional scaled effective population size and the observed number of positive individuals in a region was comparable to that seen in England as a whole (Figure S3), which is consistent with spatially segregated dynamics with similar levels of genetic drift in each region. We further describe these results in the Supplementary Information.

Potential mechanisms that can contribute to the high levels of genetic drift

Two potential mechanisms that can contribute to the observed high levels of genetic drift are: (1) variability at the individual level through superspreading (Figure 3a), and (2) host population structure (Figure 3b). We investigate each of these mechanisms in turn and compare it to our results. While in reality, both mechanisms (and others not explored here) are likely at play, it is challenging to tease them apart given our limited data. Therefore, we consider the extreme situations where only one mechanism at a time is driving the dynamics.

Superspreading occurs due to overdispersion in the number of secondary cases, which decreases the effective population size. If superspreading were the only mechanism at play, then the variance in offspring number that would explain our results would be the same as the ratio between the SIR null model $\tilde{N}^{\mathrm{SIR}}(t)$ and the inferred $\tilde{N}_e(t)$ (16-589) (Figure 3c). Current estimates of the variance in offspring number measured by contact tracing and modeling across a wide range of times and locations are from around 0.7 in one study to 65 in another (Table S1). We found two studies that apply to the UK, one which used a model that incorporated the empirical viral load trajectories and contact numbers to estimate superspreading [43] and another which used a branching process model of the number of imported and local cases [42]. The literature estimates whose time window overlapped with our time windows found substantially lower levels of superspreading than what we observe (Figure 3c and corresponding overdispersion parameter shown in Figure S16). It is possible that contact tracing and modeling over- or under-estimates overdispersion due to missed contacts [20]. However, on the other hand, it may be the case that superspreading is not the only mechanism at play.

Host deme structure is another mechanism that can lead to decreased effective population size. In such a model, individuals within a deme are very well-connected to one another (i.e. households or friend groups, also known as "communities" in network science [44]), but there are few connections between demes (Figure 3c). It is possible for deme structure to occur without superspreading. For instance, in the schematic in Figure 3c, the number of contacts is either 4 or 5; if every contact led to a transmission, this would be an extremely narrow offspring number distribution (i.e. no superspreading). Because individuals are very well-connected within a deme, once the pathogen spreads to a susceptible deme, it will spread rapidly in a deme until all individuals are infected (a jackpot event). In this way, deme structure can lower the effective population size by lowerering the effective number of stochastic transmissions. For instance, in the example in Figure 3c, there are 20 individuals, but only 3 potential stochastic transmissions. Deme structure may also arise from correlations in the number of secondary infections over a series of hosts (i.e. a series of high numbers of secondary infections in a transmission chain, or conversely low numbers of secondary infections in a transmission chain) [45]. This may arise, for instance, if individuals in a transmission chain have similar behavior, due to geographical proximity, or similar value systems on risk aversion. A recent study has found that individuals infected by superspreading tend to be superspreaders themselves more than expected by

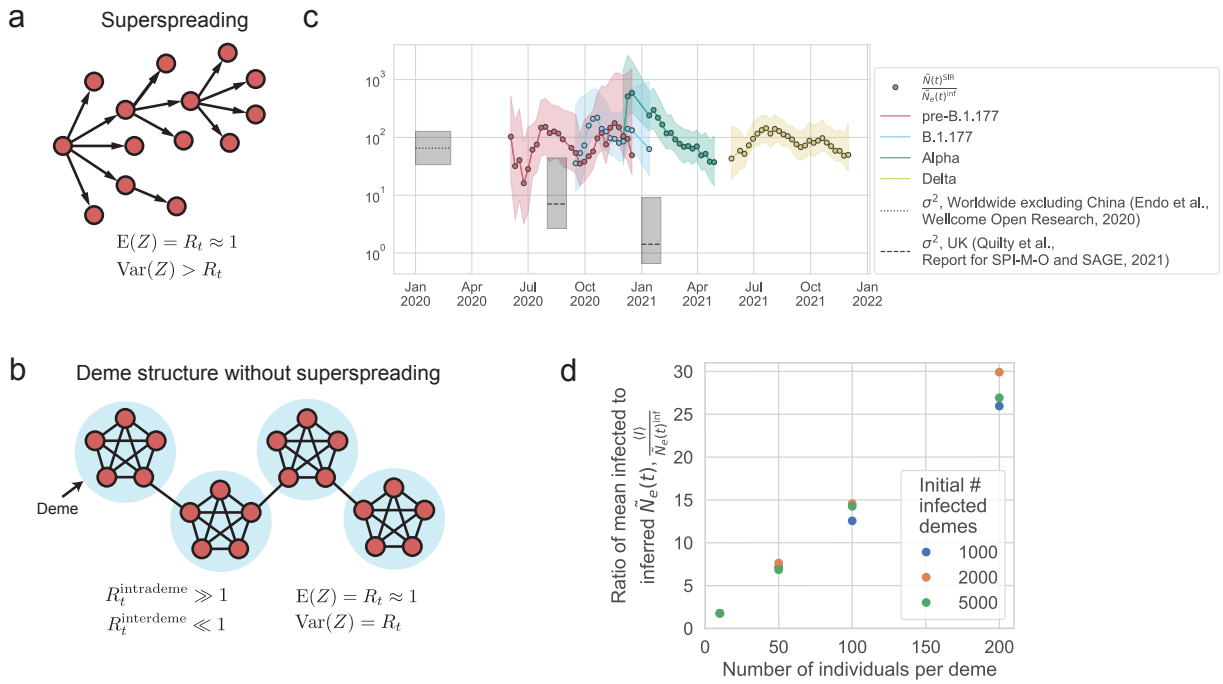

Figure 3: Potential mechanisms that can generate a low effective population size. (a) Superspreading, where the distribution of the number of secondary cases (Z) from a single infected individual is broadly distributed (variance greater than mean). (b) Deme structure without superspreading, due to heterogeneity in the host network structure, where the distribution of the number of secondary cases is not broadly distributed (variance approximately equal to mean). (c) The ratio between the $\tilde{N}^{\text{SIR}}(t)$ (the scaled population size calculated from an SIR model using the number of observed positive individuals and the observed effective reproduction number) and the inferred $\tilde{N}_e(t)$ for each variant. Only data where the error in the SIR model $\tilde{N}^{\text{SIR}}(t)$ is less than 3 times the value are shown, because larger error bars make it challenging to interpret the results. The inferred $\tilde{N}_e(t)$ is lower than the $\tilde{N}^{\text{SIR}}(t)$ (which assumes well-mixed dynamics and no superspreading) by a factor of 16 to 589, indicating high levels of genetic drift. The variance in offspring number from the literature [42, 43] does not entirely explain the discrepancy between the true and effective population sizes. (d) Simulations of deme structure without superspreading can generate high levels of genetic drift via jackpot events. SEIR dynamics are simulated within demes (with $R_t = 10$, i.e. deterministic transmission) and Poisson transmission is simulated between demes ($R_t \ll 1$, i.e. stochastic transmission) such that the population $R_t \sim 1$ (see Methods). Simulation parameters are: mean transition rate from exposed to infected $\gamma_E = (2.5 \text{ days})^{-1}$, mean transition rate from infected to recovered $\gamma_I = (6.5 \text{ days})^{-1}$, total number of demes $D_{total} = 5.6 \times 10^5$. The ratio between the number of infected individuals and the inferred effective population size is found to scale linearly with the deme size and not with the number of infected demes. This scaling results because of jackpot events where a lineage that happens to infect a susceptible deme grows rapidly until all susceptible individuals in the deme are infected.

chance [46], which would be consistent with this phenomenon.

To check our intuition that deme structure can decrease the effective population size and increase genetic drift, we ran simulations of a simplified deme model (see Methods): all demes have the same number of individuals, and there is a sufficiently large enough number of demes that the total number of demes does not matter. Initially some number of demes are infected, and transmission occurs such that the overall effective reproduction number in the population is around 1. From our simulations, we find that when the number of individuals in a deme increases, the ratio between the number of infected individuals and the inferred effective population size increases (Figure 3d); in other words, the more individuals there are in a

deme, the higher the level of genetic drift we observe compared what is expected from the number of infected individuals. This is because while the number of infected individuals increases when the deme size increases (Figure S17a), the inferred effective population size (and thus the level of stochasticity) stays the same as a function of deme size (it is more dependent on the number of infected demes) (Figure S17b). However, the exact ratio of the number of infected individuals to the inferred effective size depends on the parameters of the model.

Discussion

Here, we systematic studied the strength of genetic drift of SARS-CoV-2 in England across time and spatial scales. To do this, we developed and validated a method for jointly inferring time-varying genetic drift and overdispersed measurement noise using lineage frequency time series data (Figure 1), allowing these two effects to be disentangled, which overcomes a major challenge in the ability to infer the strength of genetic drift from time-series data. We find that the effective population size of SARS-CoV-2 in England was lower than that of an SIR null model true population size (using the observed number of positives) by a time-dependent factor ranging from 16 to 589 (Figure 3c), suggesting that there were higher levels of genetic drift than expected from uniform transmission. We also find evidence for spatial structure in the transmission dynamics during the B.1.177, Alpha, and Delta waves, as the inferred $\tilde{N}_e(t)$ was substantially lower in regions compared to that of all England (Figure S1). These findings are consistent with other studies that have found spatial structure in transmission of B.1.177 [47], Alpha [48], and Delta [49].

The levels of genetic drift that we observe are higher than literature values of superspreading; while this may be partly due to the challenges of accurately estimating the extent of superspreading from data, it also suggests that additional mechanisms may be leading to increased stochasticity. In particular, we explore a simplified deme model with groups of individuals that are well-connected to each other (demes) but not to individuals from other demes, and find that such a simple model can generate a low effective population size even in the absence of superspreading, due to jackpot events. In reality, both superspreading and host structure are likely at play. Additionally, they could interact with each other. For instance, there could be superspreading within a deme. Future work can try to tease apart the contribution of these two mechanisms, which for instance may be possible with better transmission network data, building on previous work on transmission networks [50], or with time-resolved contact tracing data [20]. This will be important because the relative contributions of the two mechanisms of superspreading and host population structure to genetic drift can affect the establishment of new variants in the population [3].

Accurately estimating the strength of genetic drift allows us to better understand disease spread and extinction, as well as to better parameterize evolutionary models and understand how mutations will establish in the population. We observed that with a few exceptions, the amount by which genetic drift was elevated compared to the number of positives did not change much over time or across variants (Figure 3c), despite changes in lockdowns and restrictions (which we may expect to decrease behavior that leads to superspreading). On the other hand, this may not be so surprising given the findings that restrictions affect the mobility network structure in a complex way, decreasing some types of mobility while increasing others [51]. One exception was that Alpha had significantly higher genetic drift compared to Delta and the strength of genetic drift in Alpha first peaked then slowly decreased over time. This may be either due to differences in the properties of the virus or differences in host behavior. For instance, it may suggest that the stochasticity in the transmission of Alpha sharply increased then slowly decreased over time. Alternatively, this may be driven by Alpha's expanding geographic range combined with reimported cases of Alpha into the UK (observed from February 2021 onwards), which could both also decrease the effective population size [52].

We observe that measurement noise of SARS-CoV-2 is mostly indistinguishable from uniform sampling, but data from some variants at some times do exhibit more elevated measurement noise than uniform sampling. Thus, we expect that assuming uniform sampling, as many methods do, or constant overdispersion will lead to accurate estimates for this dataset [27, 22, 23, 28]. The number of SARS-CoV-2 sequences from England is extremely high and sampling biases are expected to be low, because of efforts to reduce sampling biases by sampling somewhat uniformly from the population through the COVID-19 Infection Survey [31] (from which a subset of positives are sequenced and included in the COG-UK surveillance sequencing data that we use). On the other hand, other countries may have higher sampling biases, so jointly estimating

measurement noise and genetic drift may be more crucial in those settings. It may also be interesting to use this method to test whether genomics data taken from wastewater has lower levels of measurement noise as compared to sequenced cases.

We find that constant selection is unlikely to explain our results, as excluding potentially non-neutral lineages does not significantly change the inferred effective population size. Our method allows for the inference of constant selection while accounting for time-varying effective population size, and thus may be interesting in future applications of tests for selection. However, the model ignores other processes that may lead to deterministic changes in frequency, such as migration and mutation (discussed below); thus, we recommend combining it with complementary methods (for instance, phylogenetic methods) that test for selection as somewhat independent tests for selection.

In summary, we find that the strength of genetic drift in SARS-CoV-2 transmission in England is higher than expected based on the number of positive individuals and the levels of superspreading reported in the literature. Our results suggest that additional models and methods will be needed to better understand the mechanisms behind the observed elevated levels of genetic drift.

Limitations of the study and opportunities for future directions

First, in our analysis, we focused on standing variation that existed at a particular depth in the phylogenetic tree and ignored de novo mutations subsequently arising during the time series. However, we don't think this should substantially affect our results because introducing mutations in the form of new lineages with a small rate in the simulations did not have a large effect on the method performance (Figure 1e). We also ignored importation of SARS-CoV-2 into England and exportation of SARS-CoV-2 out of England. Migration can substantially change frequencies that are locally rare, but we expect importations to only weakly influence the frequency fluctuations of abundant variants, on which we have focused in this work. Additionally, we did not test for more complex forms of selection, such as fluctuating selection. More generally, future work should explore joint inference of selection, migration, and/or mutation in the model, as is appropriate for the pathogen of interest, building on previous work in this area [53, 54, 26].

Second, there may be biases in the way that data are collected that are not captured in our model. While our method does account for sampling biases that are uncorrelated in time, sampling biases that remain over time cannot be identified as such (i.e. if one geographical region was dominated by a particular lineage and it consistently had higher sequencing rates compared to another geographical region), and this can potentially bias the inferred effective population size; although, this is also a problem in phylogenetic methods. One approach to this problem that was utilized by some early methods during the pandemic is to develop sample weights based on geography, time, and number of reported cases. Future work should study the effect of different sampling intensities between regions on uncorrelated and correlated sampling noise. Additionally, we assume that the measurement noise overdispersion is identical for all lineages within a variant; in reality, there may be differences in sampling between lineages. Future work should explore whether this is the case in the actual dataset, as well as the effect of lineage-specific measurement noise overdispersion on overall method performance.

Third, the quantity of effective population size is a summary statistic that is influenced by many factors, making its interpretation challenging. The effective population size describes the population size under a well-mixed Wright-Fisher model, whereas in reality, this assumption is broken by many effects, including host structure and broad offspring number distributions. Thus, in our study we are careful to interpret effective population size only in the broadest terms of genetic drift, without being able to determine what mechanisms lead to the inferred effective population size (although we do explore some possibilities above). While within-host dynamics may in principle impact the lineage frequency trajectories, this effect is likely small for our analysis because we focus on acute infections (infections in the community rather than in hospitals and nursing homes). This is because acute infections of SARS-CoV-2 are thought to generate little within-host diversity that is passed on due to the short infection duration and small bottleneck size between hosts [55, 56]; while new mutations arising within acute hosts have been observed to be transmitted, these events are rare [55].

Fourth, the use of a sliding window of 9 weeks on the lineage frequency data will lead to smoothing of sharp changes in effective population size. It may be interesting in future work to develop a continuous method that uses a prior to condition on changes in effective population size, similar to those that have been

developed for coalescence-based methods [1, 57]. This would allow us to infer continuous changes in effective population size without needing to use a sliding window.

While we have focused on SARS-CoV-2 in this study, the method developed here can be extended to study genetic drift in other natural populations that are influenced by measurement noise and where genomic frequency data are available, for instance other pathogens, field studies, and ancient DNA [58, 59, 60]. More generally, ongoing methods development that integrates genomics, epidemiological, and other data sources is crucial for being able to harness the large amounts of data that have been generated to better understand and predict evolutionary dynamics.

Materials and Methods

Data sources and processing

We downloaded sequence data from the COVID-19 Genomics UK Consortium (COG-UK) [34]. We only used surveillance data (labeled as "pillar 2"); this dataset is composed of a random sample of the positive cases from the COVID-19 Infection Survey, which is a surveillance study of positive individuals in the community administered by the Office for National Statistics (see below). For lineages that appeared before B.1.177, we downloaded the metadata from the COG-UK Microreact dashboard [61], which included the time and location of sample collection (at the UTLA level), as well as the lineage designation using the Pango nomenclature [29, 30]. For B.1.177, Alpha, and Delta sequences, because the Pango nomenclature classified them into very few lineages, we created our own lineages from the phylogenetic trees (see below). We downloaded the publicly available COG-UK tree on February 22, 2021 for B.1.177; June 20, 2021 for Alpha; and January 25, 2022 for Delta. We also downloaded the COG-UK metadata for all lineages on January 16, 2022, which included the time and location (at the UTLA level) of sample collection. For the data of B.1.177, Alpha, and Delta, the data was deduplicated to remove reinfections in the same individual by the same lineage, but reinfections in the same individual by a different lineage were allowed. This yielded a total of 490,291 sequences.

The lineage frequency time-series is calculated separately for each variant or group of lineages (pre-B.1.177, B.1.177, Alpha, and Delta). First, the sequence metadata are aggregated by epidemiological week (Epiweek) to average out measurement noise that may arise due to variations in reporting within a week. Then, the lineage frequency is calculated by dividing the number of sequences from that lineage in the respective tree by the total number of sequences of that variant (or group of lineages) that were assigned to any lineage in the respective tree.

Because our model describes birth-death processes when the central limit theorem can be applied, we need the lineage frequencies to be sufficiently high. Thus, we randomly combine rare lineages into "superlineages" that are above a threshold number of counts and threshold frequency in the first and last timepoint of each trajectory. For the threshold, we chose of 20 counts and frequency of 0.01. Sensitivity analyses showed that the choice of the superlineage count threshold does not substantially affect the results (Figure S10). Superlineages are non-overlapping (i.e. each sequence belongs to exactly one superlineage).

The estimated number of people testing positive for COVID-19 in England and each region of England was downloaded from the UK Office for National Statistics' COVID-19 Infection Survey [31]. The COVID-19 Infection Survey includes households that are semi-randomly chosen, and individuals are tested regardless of whether they are reporting symptoms. Infections reported in hospitals, care homes, and other communal establishments are excluded. Thus the dataset provides a representative number of positive individuals in the community setting. The reported date of positive cases is the date that the sample was taken. The error on the number of positive individuals from April 17, 2020 to July 5, 2020 is reported as the 95% confidence interval, and after July 5, 2020 is reported as the 95% credible interval. The regional data reported the positivity rate over two week intervals. To get the number of positives, we multiplied by the number of individuals in the community setting in the region (excluding hospitals, care homes, and other communal establishments). As the data was reported over two week intervals, we obtained the number of positives for each week using linear interpolation.

The observed effective reproduction numbers for England and each region of England were downloaded from the UK Health Security Agency [38]. Only times where the certainty criteria are met and the inference is not based on fewer days or lower quality data are kept. The error on the effective reproduction number is reported as the 90% confidence interval. Although not reported in the dataset, we choose the point estimate of the effective reproduction number to be the midpoint between the upper and lower bounds of the 90% confidence interval.

Creating lineages in B.1.177, Alpha, and Delta

For B.1.177, Alpha, and Delta, we divided each of them into neutral lineages based on phylogenetic distance. Specifically, for B.1.177 and Alpha, we cut a phylogenetic tree (in units of number of mutations) at a certain depth, $d = d_{\text{cut}}$. Each of the internal or external branches that are cut by the line $d = d_{\text{cut}}$ defines a lineage (Figure 2a). The (observed) frequency of a lineage at a given time point in England was computed by

counting the number of England sequences (leaf nodes) belonging to the lineage and by normalizing it by the total number of sequences in all assigned lineages of the focal variant in England at that time point. Lineage frequencies at the regional level were similarly computed by counting the number of sequences separately for each region.

The choice of d_{cut} is arbitrary to some extent. Because we wanted a sufficiently high resolution of lineages from the early phase of spreading of a variant and because the evolutionary distance correlates with the actual sample date (Figure S18), for each focal variant, we chose the depth d_{cut} that roughly corresponds to the time point when it began to spread over England.

For the Delta variant, the sequences form two distinct groups along the depth direction, as seen from the last panel of Figure S18. Therefore, to divide the Delta variant into lineages with small frequencies, we cut the phylogenetic tree at two depths sequentially; we first cut the tree at $d_{\mathrm{cut}}^{(1)}$, which resulted in lineages with small frequencies plus a lineage with $\mathcal{O}(1)$ frequency. Then, to divide the latter lineage further, we took the subtree associated with this lineage and cut the subtree at $d_{\mathrm{cut}}^{(2)}$.

For the results presented in the main text, we used (in units of substitutions per site, with the reference d=0 being the most recent common ancestor) $d_{\mathrm{cut}} = 2.323 \cdot 10^{-2}$ for B.1.177, $d_{\mathrm{cut}} = 2.054 \cdot 10^{-3}$ for Alpha, and $d_{\mathrm{cut}}^{(1)} = 1.687 \cdot 10^{-3}$ and $d_{\mathrm{cut}}^{(2)} = 1.954 \cdot 10^{-3}$ for Delta. We confirmed that our results are robust to the choice of d_{cut} as well as the choice of the phylogenetic tree data we used (Figure S9).

Model for inferring effective population size from lineage frequency time series

We use a Hidden Markov Model with continuous hidden and observed states to describe the processes of genetic drift and sampling of cases for sequencing (similar to a Kalman filter) (Figure 1A). The hidden states describe the true frequencies of the lineages and the observed states describe the observed frequencies of the lineages as measured via sequenced cases.

The transition probability between the true frequencies f_t (the hidden states) due to genetic drift when $0 \ll f \ll 1$ has been shown in [62] to be well-described by the following expression, which we use as our transition probability,

$$p(f_{t+1}|f_t, \tilde{N}_e(t)) = \frac{1}{2}\sqrt{\frac{2f_t^{1/2}}{\pi f_{t+1}^{3/2}(\tilde{N}_e(t))^{-1}}}\exp\left(-\frac{2(\sqrt{f_{t+1}}-\sqrt{f_t})^2}{(\tilde{N}_e(t))^{-1}}\right). \tag{4}$$

$\tilde{N}_e(T) \equiv N_e(t)\tau(t)$ where $N_e(t)$ is the time-dependent effective population size and $\tau(t)$ is the time-dependent generation time, which is defined as the mean time between two subsequent infections per individual (i.e. the time between when an individual becomes infected and infects another individual, or the time between two subsequent infections caused by the same individual). This transition probability gives the correct first and second moments describing genetic drift when $f \ll 1$, $\mathrm{E}(f_{t+1}|f_t) = f_t$ and $\mathrm{Var}(f_{t+1}|f_t) = \frac{f_t}{\tilde{N}_e(t)}$, and is a good approximation when the central limit theorem can be applied, which is the case when $f \gg 0$. By assuming that $f_{t+1} \approx f_t$, and defining $\phi_t \equiv \sqrt{f_t}$, Equation 4 can be approximated as a simple normal distribution

$$p(\phi_{t+1}|\phi_t, \tilde{N}_e(t)) = \mathcal{N}\left(\phi_t, \frac{1}{4\tilde{N}_e(t)}\right). \tag{5}$$

We describe the emission probability from the true frequency f_t to the observed frequency f_t^{obs} (the observed states), defining $\phi_t^{obs} \equiv \sqrt{f_t^{obs}}$, as

$$p(\phi_t^{obs}|\phi_t, c_t) = \mathcal{N}\left(\phi_t, \frac{c_t}{4M_t}\right) \tag{6}$$

where M_t is the number of input sequences. Again, this distribution is generically a good description when the number of counts is sufficiently large, due to the central limit theorem. The first and second moments of this emission probability are $\mathrm{E}(f_t^{obs}|f_t) = f_t$ and $\mathrm{Var}(f_t^{obs}|f_t) = \frac{c_t}{M_t}f_t$, or equivalently considering the number of sequences $n_t^{obs} = f_t^{obs}M_t$ and the true number of positive individuals n_t, $\mathrm{E}(n_t^{obs}|n_t) = n_t$ and $\mathrm{Var}(n_t^{obs}|n_t) = c_t n_t$. Thus, c_t describes the strength of measurement noise at time t. When $c_t = 1$, the emission probability approaches that describing uniform sampling of sequences from the population

of positive individuals (i.e. can be described by a Poisson distribution in the limit of a large number of sequences), namely $\mathrm{Var}(n_t^{obs}|n_t) = n_t$ or equivalently $\mathrm{Var}(f_t^{obs}|f_t) = \frac{f_t}{M_t}$. This is the realistic minimum amount of measurement noise. When $c_t > 1$, it describes a situation where there is bias (that is uncorrelated in time) in the way that sequences are chosen from the positive population. The case of $0 < c_t < 1$ describes underdispersed measurement noise, or noise that is less random than uniform sampling. The case of $c_t = 0$ describes no measurement noise (for instance, when all cases are sampled for sequencing). These last two situations are unlikely in our data, and thus as we describe below, we constrain $c_t \geq 1$ in the inference procedure. In addition to being a good description of measurement noise, defining the emission probability in the same normal distribution form as the transmission probability allows us to easily derive an analytical likelihood function, described below (Note: see Ref. [26] for a method to derive an analytical likelihood function for arbitrary forms of the transition and emission probabilities).

We derive the likelihood function (up to a constant) for the the Hidden Markov Model using the forward algorithm, although it can alternatively be derived by marginalizing over all hidden states. We assume an (improper) uniform prior on ϕ_0 (i.e. no information about the initial true frequency of the lineage).

$$p(\phi_0, \phi_0^{obs}, \theta_0) = p(\phi_0^{obs}|\phi_0, c_0)p(\phi_0) \tag{7}$$

$$p(\phi_0) \propto 1 \tag{8}$$

$$p(\phi_t, \phi_{0:t}^{obs}, \theta_{0:t}) = p(\phi_t^{obs}|\phi_t, c_t)\int_{-\infty}^{\infty} p(\phi_t|\phi_{t-1}, \tilde{N}_e(t))p(\phi_{t-1}, \phi_{0:t-1}^{obs}, \theta_{0:t-1})d\phi_{t-1}, \quad 0 < t \leq T \tag{9}$$

$$p(\phi_{0:T}^{obs}, \theta_{0:T}) = \int_{-\infty}^{\infty} p(\phi_T, \phi_{0:T}^{obs}, \theta_{0:T})d\phi_T \tag{10}$$

$$\mathcal{L}(\vec{\phi}_{0:T}^{obs}|\theta_{0:T}) = \prod_{\alpha} p(\{\phi_{0:T}^{obs}\}_\alpha, \theta_{0:T})p(\theta_{0:T}) \tag{11}$$

$$p(\theta_{0:T}) \propto 1 \tag{12}$$

$$\mathcal{L}(\vec{\phi}_{0:T}^{obs}|\theta_{0:T}) = \prod_{\alpha} p(\{\phi_{0:T}^{obs}\}_\alpha, \theta_{0:T}). \tag{13}$$

where $\phi_{0:t}^{obs} \equiv \{\phi_0^{obs}, ..., \phi_t^{obs}\}$, $\theta_{0:t} \equiv \{\tilde{N}_e(0), ..., \tilde{N}_e(t), c_0, ..., c_t\}$, and the subscript α indicates a particular lineage. We use a uniform prior on the parameters. The parameters $\theta_{0:T}$ are inferred by maximizing the likelihood (described below).

The forward algorithm has an analytical form for the simple case of Gaussian transition and emission probabilities. We use the identity for the product of two normal distributions $N(x, \mu, v)$, where μ is the mean and v is the variance:

$$N(x, \mu_1, v_1)N(x, \mu_2, v_2) = N(\mu_1, \mu_2, v_1 + v_2)N(x, \mu_{12}, v_{12}) \tag{14}$$

$$\mu_{12}(\mu_1, \mu_2, v_1, v_2) = \frac{\mu_1 v_2 + \mu_2 v_1}{v_1 + v_2} \tag{15}$$

$$v_{12}(v_1, v_2) = \frac{1}{\frac{1}{v_1} + \frac{1}{v_2}}. \tag{16}$$

Solving the forward algorithm recursively, we have

$$p(\phi_{0:T}^{obs}, \theta_{0:T}) = \prod_{i=1}^{T} N(\phi_i^{obs}, \mu_i, \frac{c_i}{4M_i} + v_i) \tag{17}$$

$$\tag{18}$$

where

$$\mu_1 = \phi_0^{obs} \tag{19}$$

$$v_1 = \frac{\frac{1}{\tilde{N}_e(t)} + \frac{c_0}{M_0}}{4} \tag{20}$$

$$\mu_{i+1} = \mu_{12}(\mu_i, \phi_i^{obs}, v_i, \frac{c_i}{4M_i}) \tag{21}$$

$$v_{i+1} = v_{12}(\frac{c_i}{4M_i}, v_i) + \frac{1}{4\tilde{N}_e(t)}. \tag{22}$$

$$\tag{23}$$

Equation 17 can be substituted into Equation 13 to obtain the full analytical likelihood function.

Fitting the model to data

We split the time series data into overlapping periods of 9 Epiweeks, over which the effective population size is assumed to be constant. We first use the moments of the probability distributions combined with least squares minimization to get an initial guess for the parameters. Then, we perform maximum likelihood estimation using the full likelihood function. To capture uncertainties that arise from the formation of superlineages from lineages, we create superlineages randomly 100 times (except where indicated otherwise). We infer the strength of measurement noise and the effective population size for each superlineage combination (described below).

Determining the initial guess for the parameters using method of moments approach

Combining the transition and emission probabilities, and marginalizing over the hidden states we have

$$p(f_j^{obs}|f_i^{obs}) \propto \sqrt{\frac{1}{(f_j^{obs})^{3/2}}} \exp\left(- \frac{2\left(\sqrt{f_j^{obs}} - \sqrt{f_i^{obs}}\right)^2}{\kappa_{i,j}} \right) \tag{24}$$

$$p(\phi_j^{obs}|\phi_i^{obs}) = \mathcal{N}(\phi_i^{obs}, \kappa_{i,j}) \tag{25}$$

$$\kappa_{i,j} \equiv \frac{c_i}{4M_i} + \frac{c_j}{4M_j} + \frac{(j-i)}{4\tilde{N}_e(t)}. \tag{26}$$

The first two terms of $\kappa_{i,j}$ are the contribution to the variance from measurement noise at times i ad j, and the third term is the contribution to the variance from genetic drift.

We calculate the maximum likelihood estimate of $\kappa_{i,j}$, $\hat{\kappa}_{i,j}$, which is simply the mean squared displacement

$$\hat{\kappa}_{i,j} = \left\langle (\phi_j^{obs} - \phi_i^{obs})^2 \right\rangle. \tag{27}$$

The standard error is given by

$$\Delta\hat{\kappa}_{i,j} = \sqrt{\frac{\left\langle \left[(\phi_j^{obs} - \phi_i^{obs})^2 - \hat{\kappa}_{i,j} \right]^2 \right\rangle}{Z}} \tag{28}$$

where Z is the number of superlineages.

By looking across all pairs of timepoints i and j, we get a system of linear equations in $\kappa_{i,j}$ that depend on the parameters c_t and $\tilde{N}_e(t)$. To determine the most likely values of the parameters, we minimize

$$\ln \sum_{i,j} \frac{(\hat{\kappa}_{i,j} - Ac)^2}{\Delta\hat{\kappa}_{i,j}} \tag{29}$$

using scipy.optimize.minimize with the L-BFGS-B method and the bounds $1 \leq c_t \leq 100$ and $1 \leq \tilde{N}_e(t) \leq 10^7$. While underdispersed measurement noise ($c_t < 1$) is in principle possible, we constrain $c_t \geq 1$ because

realistically, the lowest amount of measurement noise will be from uniform sampling of sequences. An example of inferred parameters using the methods of moments approach on simulated data is shown in Figure S19.

Maximum likelihood estimation of the parameters

For each set of superlineages, we use the inferred measurement noise values (c_t) and inferred scaled effective population size from above ($\tilde{N}_e(t)$) as initial guesses in the maximization the likelihood function in Equation 13 over the parameters. For the optimization, we use scipy.optimize.minimize_scalar with the Bounded method and the bounds $1 \leq c_t \leq 100$ and $1 \leq \tilde{N}_e(t) \leq 10^{11}$. The time t in the inferred $\tilde{N}_e(t)$ is taken to be the midpoint of the 9 Epiweek period. The reported $\tilde{N}_e(t)$ is the median inferred $\tilde{N}_e(t)$ across all superlineage combinations where $\tilde{N}_e(t) < 10^5$ (values above 10^5 likely indicate non-convergence of the optimization). The reported errors on $\tilde{N}_e(t)$ are the 95% confidence intervals (again taking the median across all superlineage combinations where $\tilde{N}_e(t) < 10^5$) which are calculated by using the likelihood ratio to get a p-value [63, 64]. We replace the likelihood with the profile likelihood, which has the nuisance parameters $c_{0:T}$ profiled out:

$$p > 0.05 \tag{30}$$

$$p = \int I\left[\frac{\mathcal{L}_{\tilde{N}_e}(\hat{c}_{0:T}|\vec{\phi}^{obs}_{0:T})}{\mathcal{L}_{\tilde{N}'_e}(\hat{c}_{0:T}|\vec{\phi}^{obs}_{0:T})} > 1\right] P_{\tilde{N}'_e}(\hat{c}_{0:T}|\vec{\phi}^{obs}_{0:T}) d\tilde{N}'_e \tag{31}$$

$$\hat{c}_{0:T} = \arg\max_{c_{0:T}} \mathcal{L}_{\tilde{N}_e}(c_{0:T}|\vec{\phi}^{obs}_{0:T}) \tag{32}$$

$$P_{\tilde{N}'_e}(\hat{c}_{0:T}|\vec{\phi}^{obs}_{0:T}) \propto \mathcal{L}_{\tilde{N}'_e}(\hat{c}_{0:T}|\vec{\phi}^{obs}_{0:T}) p(\tilde{N}_e) \tag{33}$$

$$p(\tilde{N}_e) \propto 1 \tag{34}$$

where I is an indicator function that equals one when the argument is true and zero otherwise, $\mathcal{L}_{\tilde{N}_e}(\hat{c}_{0:T}|\vec{\phi}^{obs}_{0:T})$ is the profile likelihood with the nuisance parameters (in this case) $c_{0:T}$ profiled out, $P_{\tilde{N}'_e}(\hat{c}_{0:T}|\vec{\phi}^{obs}_{0:T})$ is the posterior where we have used a uniform prior. We also tried a Jeffreys prior which is used for variance parameters, but it gave similar results on simulated data because it looked relatively flat over the values of $\tilde{N}_e(t)$ of interest. As the Jeffreys prior was more computationally expensive than the uniform prior and the two priors gave similar results, we used the uniform prior for the analyses.

The reported values of c_t are the median across all superlineage combinations and across all time series segments where the timepoint appears. The reported errors on c_t are the 95% confidence intervals as calculated by the middle 95% of values across superlineage combinations and time series segments.

An example of inferred parameters on simulated data using the maximum likelihood estimation approach, compared to the initial guesses of the parameters from the methods of moments approach, is shown in Figure S19.

Correcting for the number of sequences assigned to lineages

Because some sequences occur before the cut point in the tree that is used for creating lineages, they are not included in any lineages. As a result, the number of sequences assigned to lineages is lower than the number of sequences in the tree. To correct for the bias in inferred effective population size that results from leaving out sequences from parts of the tree, we divide the inferred effective population size by the fraction of sequences in the tree that are assigned to a lineage. We note that while the number of sequences in the tree is less than the total number of sampled sequences, the sequences in the tree were chosen to be a representative fraction of the total sampled sequences. Thus, we do not need to additionally correct for the downsampling of sequences that were included in the tree. To test that randomly subsampling sequences for the analysis does not affect the results, we randomly subsampled half of the Delta sequences, and reran the analyses; the inferred effective population size was very similar to that from the full number of sequences (Figure S20).

Simulations for validating method

For the model validation, we perform simulations of the lineage trajectories using a discrete Wright-Fisher model. 500 lineages are seeded initially, and the initial frequency of lineages is taken to be the same across all lineages. In each subsequent Epiweek, the true number of counts for a lineage is drawn from a multinomial distribution where the probabilities of different outcomes are the true frequencies of the lineages in the previous Epiweek and the number of experiments is the effective population size. The true frequency is calculated by dividing the true number of counts by N. The observed counts are drawn from a negative binomial distribution,

$$p(n_t^{obs}|f_t) = NB(r,q) \equiv \binom{n_t^{obs}+r-1}{r-1} q^r (1-q)^{n_t^{obs}} \tag{35}$$

$$r = \frac{f_t M_t}{c_t - 1} \tag{36}$$

$$q = \frac{1}{c_t} \tag{37}$$

which has the same mean and variance as the emission probability in Equation 6. The total number of observed sequences in each timepoint is calculated empirically after the simulation is completed, as it may not be exactly M_t. The simulation is run for 10 weeks of "burn-in" time before recording to allow for equilibration. Superlineages are created in the same way as described above.

For long time series simulations, some lineages will go extinct due to genetic drift, making it challenging to have sufficient data for the analysis. To be able to have a high enough number of lineages for the entire time series, we introduce mutations to a new lineage with a small rate $\mu = 0.01$ per generation per individual.

Calculating the effective population size for an SIR or SEIR model

The effective population size times the generation time in an SIR model is given by Refs. [41, 35]

$$\tilde{N_e}^{\text{SIR}}(t) \equiv N_e^{\text{SIR}}(t)\tau(t) = \frac{I(t)}{2R_t\gamma_I}. \tag{38}$$

The variance in offspring number for an SIR model is approximately 2.

For an SEIR model, we calculated $\tilde{N}_e(t)$ following the framework from Ref. [36]. Using this framework, we were only able to consider a situation where the epidemic is in equilibrium. We test how well this approximates the situation out of equilibrium using simulations (see Supplementary Information).

We first considered how the mean number of lineages, A, changes going backwards in time, s, which is given by

$$\frac{dA}{ds} = -fp_c \tag{39}$$

where f is the number of transmissions per unit time and p_c is the probability that a transmission results in a coalescence being observed in our sample. p_c is given by the number of ways of choosing two lineages divided by the number of ways of choosing two infectious individuals

$$p_c = \frac{\binom{A(s)}{2}}{\binom{N(s)}{2}} \underset{\lim_{N(s)\to\infty}}{=} \binom{A(s)}{2}\frac{2}{N(s)^2}. \tag{40}$$

where the limit assumes that the number of infectious individuals, $N(s)$, is large. In the Kingman coalescent we also have

$$\frac{dA}{ds} = -\binom{A(s)}{2}\frac{1}{\tilde{N}_e(t)}. \tag{41}$$

Combining Equations 39, 40, and 41, we have

$$\tilde{N}_e(t) = \frac{N(s)^2}{2f}. \tag{42}$$

Thus by determining the number of transmissions per unit time, f, and the number of infectious individuals, $N(s)$, in an SEIR model, we can find an expression for $\tilde{N}_e(t)$.

These quantities can be derived from the equations describing the number of susceptible (S), exposed (E), infectious (I), and recovered (R) individuals in an SEIR model

$$\frac{dS}{dt} = -\beta I \frac{S}{N_H} \tag{43}$$

$$\frac{dE}{dt} = \frac{\beta I S}{N_H} - \gamma_E E - \delta_E E \tag{44}$$

$$\frac{dI}{dt} = \gamma_E E - \gamma_I I - \delta_I I \tag{45}$$

$$\frac{dR}{dt} = \gamma_I I \tag{46}$$

where β is the number of transmissions per infectious individual per unit time (the number of contacts made by an infectious individual per unit time multiplied by the probability that a contact results in a transmission), N_H is the total population size ($N_H = S + E + I + R$), γ_E is the rate that an exposed individual becomes infectious, δ_E is the rate of death for an exposed individual, γ_I is the rate than an infectious individual recovers, and δ_I is the rate of death for an infectious individual.

The number of infectious individuals in a generation, $N(s)$, is given by the instantaneous number of infectious individuals plus the number of exposed individuals that will become infectious in that generation [41]. Thus,

$$N(s) = \frac{\gamma_E}{\gamma_E + \delta_E} E + I. \tag{47}$$

The number of transmissions per unit time is given by

$$f = \beta I \frac{S}{N_H}. \tag{48}$$

We rewrite f in terms of the effective reproduction number (for which data are available) which is given by the number of transmissions per unit time (f) divided by the number of recoveries and deaths per unit time

$$R_t = \frac{f}{(\gamma_I + \delta_I) I + \delta_E E}. \tag{49}$$

Putting everything together, we have that $\tilde{N}_e(t)$ for an SEIR model in equilibrium is given by

$$\tilde{N_e}^{\text{SEIR,eq}}(t) = \frac{\left[\left(\frac{\gamma_E}{\gamma_E + \gamma_I}\right) E + I\right]^2}{2R_t[(\gamma_I + \delta_I) I + \delta_E E]}. \tag{50}$$

For SARS-CoV-2, the death rates are much lower than the rate at which exposed individuals become infectious and the rate at which infectious individuals recover ($\delta_E, \delta_I \ll \gamma_E, \gamma_I$). In this limit, Equation 50 simplifies to

$$\tilde{N_e}^{\text{SEIR,eq}}(t) = \frac{(E + I)^2}{2R_t \gamma_I I}. \tag{51}$$

To calculate the $\tilde{N_e}$ for an SIR or SEIR model, we use the estimated number of positives from the COVID-19 Infection Survey for $I(t)$. This number is an estimate of the number of positive individuals in the community as measured by surveillance and includes both symptomatic and asymptomatic individuals. While the estimated number of positives does not include cases from hospitals, care homes, and other communal establishments, community cases likely contribute the most to transmission. We used the measured effective reproduction number from the UK Health Security Agency for R_t.

To calculate the number of exposed individuals for the SEIR model, we solved for E in Equation 45 (taking $\delta_E \ll \gamma_E$)

$$E = \frac{1}{\gamma_E}\left(\frac{dI}{dt} + \gamma_I I\right). \tag{52}$$

$\frac{dI}{dt}$ was calculated numerically as $\frac{I(t+\Delta t)-I(t-\Delta t)}{2\Delta t}$ where $\Delta t = 1$ week. The parameter values used were $\gamma_E^{-1} = 3$ days and $\gamma_I^{-1} = 5.5$ days [39, 40]. We checked that varying the value used for γ_I does not substantially affect the results (Figure S12). The error on E was calculated by taking the minimum and maximum possible values from the combined error intervals of $I(t+\Delta t)$ and $I(t-\Delta t)$ (note that this does not correspond to a specific confidence interval size).

The error on $\tilde{N}_e(t)$ for the SIR or SEIR model was calculated similarly by taking the minimum and maximum possible values from the combined error intervals of E, I, and R_t. Only time points where the error interval of $\tilde{N}_e(t)$ was less than 3 times the point estimate were kept.

Calculating the effective population size for an SIR or SEIR model by variant

To calculate the effective population size for an SIR or SEIR model by variant, we needed to determine the variant-specific: number of infectious individuals $I(t)$, number of exposed individuals $E(t)$, effective reproduction number R_t, and rate than an infectious individual recovers γ_I. We assumed that γ_I is constant between variants. We calculated the number of infectious individuals $I(t)$ by multiplying the total number of positives by the fraction of each variant in the reported sequences. This should be a good representation of the fraction of the variant in the population as the sequences are a random sample of cases detected via surveillance. We calculated the number of variant-specific exposed individuals $E(t)$ in the same way as described above using the variant-specific number of infectious individuals. We assumed that the rate an exposed individual becomes infectious γ_E is constant between variants.

We calculated the variant-specific effective reproduction number by rescaling the measured effective reproduction number for the whole population

$$R_t^v = R_t \frac{R_0^v}{\sum_w R_0^w f^w} \tag{53}$$

where R_0^w is the basic reproduction number of the variant w and f^w is the fraction of the infectious population with variant w. The values of R_0 when rescaled to $R_0^{pre-B.1.1.7}$ that are used for the data presented in the main text are $\frac{R_0^{\text{pre-B.1.1.7}}}{R_0^{\text{pre-B.1.1.7}}} = 1$, $\frac{R_0^{\text{Alpha}}}{R_0^{\text{pre-B.1.1.7}}} = 1.7$ (Ref. [17]), $\frac{R_0^{\text{Delta}}}{R_0^{\text{pre-B.1.1.7}}} = 1.97$ (Ref. [65]). Varying the variant R_0 within the ranges reported in the literature does not substantially affect the results (Figure S21).

Inference of fitness from lineage frequency time series

We sought to infer the fitness effects of individual lineages, so that we could then determine if putatively selected lineages are influencing the estimation of the time-varying effective population sizes. We first used a deterministic method to estimate lineage fitness effects, similar to the method described in [66].

On average, when the frequency of lineage i is sufficiently small $f_{t,i} \ll 1$, the frequency dynamics will exponentially grow/decay according to the lineage fitness effect, s_i,

$$\langle f_{t,i} \rangle = f_{0,i} e^{s_i t}$$

The two sources of noise–genetic drift and measurement noise–both arise from counting processes, so the combined noise will follow $\text{var}\left(f_{t,i}\right) \propto \langle f_{t,i} \rangle$. To account for the inherent discreteness of the number of cases in a lineage–especially important to accurately model lineages at low frequencies–we modeled the observed counts at Epiweek t of lineage i, $r_{t,i}$, as a negative binomial random variable,

$$r_{t,i} | s_i, f_{0,i} \sim \text{NB}\left(\mu_{t,i}, \zeta_t\right) \tag{54}$$

$$\langle r_{t,i} \rangle = \mu_{t,i} \tag{55}$$

$$\text{var}\left(r_{t,i}\right) = \zeta_t \langle r_{t,i} \rangle \tag{56}$$

$$\mu_{t,i} = M_t f_{0,i} e^{s_i t} \tag{57}$$

Where M_t is the total number of sequences, and ζ_t is a dispersion parameter. We took ζ_t as the total marginal variance at a given time-point, i.e. $\zeta_t = c_t + M_t/N_e(t)$, where we computed estimates of c_t and N_e

as previously described (section "Maximum likelihood estimation of the parameters"). The final likelihood for the fitness, s_i, of lineage i is obtained by combining the data from all the relevant the time-points,

$$P(\boldsymbol{r}_i|s_i, f_{0,i}) = \prod_t \frac{\Gamma\left(r_{t,i} + \frac{\mu_{t,i}}{\zeta_t - 1}\right)}{\Gamma\left(\frac{\mu_{t,i}}{\zeta_t - 1}\right)\Gamma\left(r_{t,i}+1\right)} \frac{(\zeta_t - 1)^{r_{t,i}}}{\zeta_t^{r_{t,i} + \frac{\mu_{t,i}}{\zeta_t - 1}}} \tag{58}$$

The point estimate of the lineage fitness, $\hat{s}_i$, is then numerically computed as the maximum likelihood,

$$\hat{s}_i = \underset{s_i}{\operatorname{argmax}} \log P(\boldsymbol{r}_i|s_i, f_{0,i}) \tag{59}$$

For ease of computation and generality we compute the p-value as the posterior probability that the likelihood ratio between null and alternative hypotheses is greater than 1, i.e. the probability that the data more strongly support the null hypothesis (the lineage is neutral) over the alternative (the lineage is non-neutral),

$$p_i = P_{s_i|\boldsymbol{r}}\left(\frac{\mathcal{L}(0|\boldsymbol{r}_i)}{\mathcal{L}(s_i|\boldsymbol{r}_i)} > 1\right)$$
$$P(s_i|\boldsymbol{r}) \propto \mathcal{L}(s_i|\boldsymbol{r}_i)$$

Where $\mathcal{L}(s_i|\boldsymbol{r}_i)$ is the profile likelihood. This convenient definition has been shown to be equivalent to the frequentist definition of the p-value using a likelihood ratio test statistic (if the distribution is invariant under transformation) [63, 64], and does not require asymptotic approximations. After obtaining p-values for all the lineages, we performed a standard Benjamini-Hochberg FDR on all the p-values. As mentioned in the main text, the N_e inference method was then re-run using all lineages, except those that were significant at FDR< 0.05 using the above method.

Subsequently, we sought to use a more accurate method to detect if lineages have constant, non-neutral fitness effects. So, we turned back to our HMM method and modified it slightly. We wanted to measure the fitness effects of lineages that may be very small, with counts at some time points close to or at 0, which then would not satisfy the central limit theorem, upon which our main HMM method rests. Thus, we used a different transition [62] and emmission probability for the HMM, which are more valid for small population sizes and counts:

$$p(f_{t+1}|f_t, N_e(t), s_i) = \frac{1}{2}\sqrt{\frac{2(e^{s_i t} f_t)^{1/2}}{\pi f_{t+1}^{3/2}(\tilde{N}_e(t))^{-1}}} \exp\left(-\frac{2(\sqrt{f_{t+1}} - \sqrt{e^{s_i t} f_t})^2}{(\tilde{N}_e(t))^{-1}}\right) \tag{60}$$

$$f_t^{obs}|f_t, c_t \sim \mathrm{NB}\left(f_t, c_t\right) \tag{61}$$

$$\langle f_t^{obs} \rangle = f_t \tag{62}$$

$$\mathrm{var}\left(f_t^{obs}\right) = c_t f_t. \tag{63}$$

We used the point estimates of N_e and c_t that were calculated after we had excluded the selected lineages detected by the conservative method. To get the maximum likelihood estimate of s for each lineage, we used the standard HMM forward algorithm, numerically integrating over the intermediates. We then calculated the p-value for each lineage by calculating the log-likelihood ratio $\mathrm{LLR}_i = 2(\mathcal{L}_i(\hat{s}_i) - \mathcal{L}_i(0))$, and then comparing it to a chi-squared ratio with d.f.=1. After obtaining p-values for all the lineages, we performed a standard Benjamini-Hochberg FDR on all the p-values.

Stochastic simulations of SEIR model

The stochastic simulations of an SEIR model were performed using a Gillespie simulation with 4 states: susceptible, exposed, infectious, and recovered, where the number of individuals in each state are denoted by $S(t)$, $E(t)$, $I(t)$, and $R(t)$ respectively. There are 3 types of events that lead to the following changes in the number of individuals in each state

1. Infection of an susceptible individual with probability $\frac{\beta I(t)S(t)}{N(t)}$

$$S(t) = S(t) - 1 \tag{64}$$
$$E(t) = E(t) + 1 \tag{65}$$

2. Transition of an exposed individual to being infectious with probability $\gamma_E E(t)$

$$E(t) = E(t) - 1 \tag{66}$$
$$I(t) = I(t) + 1 \tag{67}$$

3. Recovery of an infectious individual with probability $\gamma_I I(t)$

$$I(t) = I(t) - 1 \tag{68}$$
$$R(t) = R(t) + 1 \tag{69}$$

where $\beta \equiv R_0\gamma_I$, R_0 is the basic reproduction number, γ_E is the rate that exposed individuals become infectious, and γ_I is the rate that infectious individuals recover. As in the rest of this work, we assume that the birth rate of susceptible individuals, background death rate, and the death rate due to disease are much slower compared to the rates of the above processes and thus can be neglected from the dynamics.

The time until the next event is drawn from an exponential distribution with rate given by the inverse of the sum of the above probabilities, and the type of event is randomly drawn weighted by the respective probabilities.

Because the time of the events occurs in continuous time, but the inference method of the effective population size works in discrete time, we must convert from continuous to discrete time. To perform this conversion, we calculate the net number of events of each type in each chosen unit of discrete time (1 week) and perform the changes in the number of individuals of each state as described above. Thus, for example, if within the same week an individual becomes exposed and then becomes infectious, it will cause the number of susceptible individuals to decrease by 1, no change in the number of exposed individuals, and the number of infectious individuals to increase by 1.

The infected (or infected and exposed) individuals are randomly assigned a lineage at a given time after the start of the epidemic. For our simulations, we chose the lineage labeling time as 75 days or 10.7 weeks since the approximate number of infectious individuals was high enough at that time to generate sufficient diversity in lineages, and we chose the number of different types of lineages as 100. The other parameters that we used for the simulations were $R_0 = 2$, $\gamma_E^{-1} = 3$ days, $\gamma_I^{-1} = 5.5$ days, $N(t) = S(t)+E(t)+I(t)+R(t) = 10^6$. The initial condition of the simulation is $S(t) = N(t) - 1$, $E(t) = 1$, and $I(t) = R(t) = 0$.

To test the sensitivity of the results to whether the reported PCR positive individuals are infectious or whether they can also be from the exposed class, we recorded the results in two ways. In the first case, only the infectious individuals we recorded as positive (Figure S22), and in the second case both the exposed and infectious individuals were recorded as positive (Figure S23). Inference of $\tilde{N}_e(t)$ was subsequently done on the lineage frequency trajectories of the recorded positive individuals. The SIR or SEIR model $\tilde{N}_e(t)$ were calculated analytically using the true numbers of infectious and exposed individuals and numerically using the number of positive individuals as described above in "Calculating the effective population size for an SIR or SEIR model".

Deme simulations

To better understand the effect of host population structure on the effective population size, we simulated a simple situation where there are "demes", or groups, of individuals with very high rates of transmission between individuals in that deme, but the rate of transmission between individuals from different demes is very low. In a given simulation, all demes have the same number of individuals (10, 50, 100, or 200). The total number of demes is chosen to be very high (5.6×10^6). Initially, a certain number of demes (100, 1000, 2000, or 5000) are each seeded by a single infectious individual infected by a randomly chosen lineage (200 different lineages). We simulated deterministic SEIR dynamics within demes with $R_0 = 10$, $\gamma_E = (2.5 \text{ days})^{-1}$, $\gamma_I = (6.5 \text{ days})^{-1}$. We simulated Poisson transmission dynamics between demes. In

order to calibrate the overall population dynamics to be roughly in equilibrium (the number of infectious individuals is not deterministically growing or shrinking), we draw the number of between-deme infections caused by a given deme from a Poisson distribution with mean 1. The time of the between-deme infection event is randomly chosen, weighted by the number of infected individuals within a deme at a given time. The number of infectious individuals in each lineage is recorded every 1 week, and the frequency of the lineage is calculated by dividing by the total number of infectious individuals from all lineages in that week. The lineage frequency data from a period of 9 weeks starting in week 42 is used for the inference of effective population size. The effective population size inference is performed as above except in the absence of measurement noise, so there is no emission step in the HMM.

Data and code availability

Data and code to reproduce the analyses in this manuscript are available at https://github.com/qinqin-yu/sars-cov-2_genetic_drift.

Acknowledgements

We are grateful to the Hallatschek lab for helpful discussions, feedback, and comments on earlier versions of this manuscript, particularly Giulio Issachini and Valentin Slepukhin. We are grateful to Aditya Prasad for advice on computing. We thank Mike Boots, Vince Buffalo, Katia Koelle, Priya Moorjani, Rasmus Nielsen, Daniel Reeves, and Daniel Weissman for helpful discussions and feedback. We thank Hernan G. Garcia and Yun S. Song for helpful comments on earlier versions of this manuscript. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1106400 (to QY), and JSPS KAKENHI (#22K03453, #JP22K06347, to TO). JAA acknowledges support from an NSF Graduate Research Fellowship and a Berkeley Fellowship. OH acknowledges support by a Humboldt Professorship of the Alexander von Humboldt Foundation. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC BER-ERCAP0019907. COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) [grant code: MC_PC_19027], and Genome Research Limited, operating as the Wellcome Sanger Institute. The authors acknowledge use of data generated through the COVID-19 Genomics Programme funded by the Department of Health and Social Care. The views expressed are those of the author and not necessarily those of the Department of Health and Social Care or UKHSA. We thank the COG-UK consortium and all partners and contributors who are listed at https://www.cogconsortium.uk/about/.

References

1. Volz EM and Didelot X. Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. Systematic Biology 2018; 67:719–28
2. Lloyd-Smith JO, Schreiber SJ, Kopp PE, and Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005; 438:355–9
3. Alexander H and Day T. Risk factors for the evolutionary emergence of pathogens. Journal of The Royal Society Interface 2010; 7:1455–74
4. Kucharski A and Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Eurosurveillance 2015; 20:21167
5. Melsew YA, Gambhir M, Cheng AC, McBryde ES, Denholm JT, Tay EL, and Trauer JM. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. BMC Infectious Diseases 2019; 19:1–9

6. Pastor-Satorras R, Castellano C, Van Mieghem P, and Vespignani A. Epidemic processes in complex networks. Reviews of Modern Physics 2015; 87:925

7. Otto SP, Day T, Arino J, Colijn C, Dushoff J, Li M, Mechai S, Van Domselaar G, Wu J, Earn DJ, et al. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. Current Biology 2021; 31:R918–R929

8. Goyal A, Reeves DB, and Schiffer JT. Early super-spreader events are a likely determinant of novel SARS-CoV-2 variant predominance. medRxiv 2021

9. Day T, Gandon S, Lion S, and Otto SP. On the evolutionary epidemiology of SARS-CoV-2. Current Biology 2020; 30:R849–R857

10. Tasakis RN, Samaras G, Jamison A, Lee M, Paulus A, Whitehouse G, Verkoczy L, Papavasiliou FN, and Diaz M. SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial Founder Events and mutational bursts. PloS ONE 2021; 16:e0255169

11. Ghafari M, Liu Q, Dhillon A, Katzourakis A, and Weissman DB. Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern. Frontiers in Virology 2022 :76

12. Sneppen K, Nielsen BF, Taylor RJ, and Simonsen L. Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. Proceedings of the National Academy of Sciences 2021; 118

13. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics 2009; 10:195–205

14. Lakdawala SS and Menachery VD. Catch me if you can: superspreading of COVID-19. Trends in Microbiology 2021; 29:919–29

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

16. Goyal A, Reeves DB, Cardozo-Ojeda EF, Schiffer JT, and Mayer BT. Viral load and contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events. eLife 2021; 10:e63537

17. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, Hinsley WR, Laydon DJ, Dabrera G, O'Toole Á, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 2021; 593:266–9

18. Ragonnet-Cronin M, Boyd O, Geidelberg L, Jorgensen D, Nascimento FF, Siveroni I, Johnson RA, Baguelin M, Cucunubá ZM, Jauneikaite E, et al. Genetic evidence for the association between COVID-19 epidemic severity and timing of non-pharmaceutical interventions. Nature Communications 2021; 12:1–7

19. Du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, Raghwani J, Ashworth J, Colquhoun R, Connor TR, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science 2021; 371:708–12

20. Adam D, Gostic K, Tsang T, Wu P, Lim WW, Yeung A, Wong J, Lau E, Du Z, Chen D, et al. Time-varying transmission heterogeneity of SARS and COVID-19 in Hong Kong (preprint). 2022

21. Frost SD, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, and Bedford T. Eight challenges in phylodynamic inference. Epidemics 2015; 10:88–92

22. Bollback JP, York TL, and Nielsen R. Estimation of $2N_e s$ from temporal allele frequency data. Genetics 2008; 179:497–502

23. Buffalo V and Coop G. The linked selection signature of rapid adaptation in temporal genomic data. Genetics 2019; 213:1007–45

24. Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, and Sherlock G. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature 2015; 519:181–6

25. Williamson EG and Slatkin M. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics 1999; 152:755–61

26. Steinrücken M, Bhaskar A, and Song YS. A novel spectral method for inferring general diploid selection from time series genetic data. The Annals of Applied Statistics 2014; 8:2203

27. Stadler T. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. Journal of Theoretical Biology 2009; 261:58–66

28. Dorp CHv, Goldberg EE, Hengartner N, Ke R, and Romero-Severson EO. Estimating the strength of selection for new SARS-CoV-2 variants. Nature Communications 2021; 12:1–13

29. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, Colquhoun R, Ruis C, Abu-Dahab K, Taylor B, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evolution 2021; 7:veab064

30. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, Plessis L du, and Pybus OG. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology 2020; 5:1403–7

31. UK Office for National Statistics. Coronavirus (COVID-19) Infection Survey: England. `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/coronaviruscovid19infectionsurveydata`. Accessed: 2021-12-10

32. Challen R, Dyson L, Overton CE, Guzman-Rincon LM, Hill EM, Stage HB, Brooks-Pollock E, Pellis L, Scarabel F, Pascall DJ, et al. Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England. medRxiv 2021

33. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, Pearson CA, Russell TW, Tully DC, Washburne AD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science 2021; 372:eabg3055

34. The COVID-19 Genomics UK (COG-UK) consortium. An integrated national scale SARS-CoV-2 genomic surveillance network. The Lancet Microbe 2020; 1:e99

35. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, and Frost SD. Phylodynamics of infectious disease epidemics. Genetics 2009; 183:1421–30

36. Frost SD and Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. Philosophical Transactions of the Royal Society B: Biological Sciences 2010; 365:1879–90

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

38. UK Health Security Agency. The R value and growth rate. `https://www.gov.uk/guidance/the-r-value-and-growth-rate#:~:text=The%20R%20range%20for%20the,as%20of%2019%20March%202021.&text=The%20R%20range%20for%20the,as%20of%2012%20March%202021.`. Accessed: 2021-12-10

39. Miller D, Martin MA, Harel N, Tirosh O, Kustin T, Meir M, Sorek N, Gefen-Halevi S, Amit S, Vorontsov O, et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nature Communications 2020; 11:1–10

40. He X, Lau EH, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nature medicine 2020; 26:672–5

41. Koelle K and Rasmussen DA. Rates of coalescence for common epidemiological models at equilibrium. Journal of the Royal Society Interface 2012; 9:997–1007

42. Endo A et al. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Research 2020; 5

43. Quilty BJ, Chapman LA, Wong KL, Gimma A, Pickering S, JD S, Neil RPG, Jarvis CI, and Kucharski AJ. Reconstructing the secondary case distribution of SARS-CoV-2 from heterogeneity in viral load trajectories and social contacts. Report for SPI-M-O and SAGE 2021

44. Obadia T, Silhol R, Opatowski L, Temime L, Legrand J, Thiébaut AC, Herrmann JL, Fleury E, Guillemot D, Boelle PY, et al. Detailed contact data and the dissemination of Staphylococcus aureus in hospitals. PLoS Computational Biology 2015; 11:e1004170

45. Steiner MC and Novembre J. Population genetic models for the spatial spread of adaptive variants: A review in light of SARS-CoV-2 evolution. PLoS Genetics 2022; 18

46. Taube JC, Miller PB, and Drake JM. An open-access database of infectious disease transmission trees to explore superspreader epidemiology. PLoS Biology 2022; 20:e3001685

47. Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KH, Althaus CL, Reichmuth ML, Bowen JE, Walls AC, Corti D, et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature 2021; 595:707–12

48. Kraemer MU, Hill V, Ruis C, Dellicour S, Bajaj S, McCrone JT, Baele G, Parag KV, Battle AL, Gutierrez B, et al. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. Science 2021; 373:889–95

49. McCrone JT, Hill V, Bajaj S, Pena RE, Lambert BC, Inward R, Bhatt S, Volz E, Ruis C, Dellicour S, et al. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. Nature 2022; 610:154–60

50. Nande A, Adlam B, Sheen J, Levy MZ, and Hill AL. Dynamics of COVID-19 under social distancing measures are driven by transmission network structure. PLoS computational biology 2021; 17:e1008684

51. Edsberg Møllgaard P, Lehmann S, and Alessandretti L. Understanding components of mobility during the COVID-19 pandemic. Philosophical Transactions of the Royal Society A 2022; 380:20210118

52. Wakeley J. Coalescent theory. Roberts & Company 2009

53. Feder AF, Kryazhimskiy S, and Plotkin JB. Identifying signatures of selection in genetic time series. Genetics 2014; 196:509–22

54. Mathieson I and McVean G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics 2013; 193:973–84

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

56. Martin MA and Koelle K. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". Science translational medicine 2021; 13:eabh1803

57. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, and Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Molecular biology and evolution 2013; 30:713–24

58. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, et al. A draft sequence of the Neandertal genome. Science 2010; 328:710–22

59. Dehasque M, Ávila-Arcos MC, Dıez-del-Molino D, Fumagalli M, Guschanski K, Lorenzen ED, Malaspinas AS, Marques-Bonet T, Martin MD, Murray GG, et al. Inference of natural selection from ancient DNA. Evolution Letters 2020; 4:94–108

60. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 2010; 468:1053–60

61. COG-UK and Microreact. UK SARS-CoV-2. `https://microreact.org/project/mxgBucHEZCZgsSWNrnQQRo-%20uk-sars-cov-2-2020-02-052021-04-20`. Accessed: 2021-04-20

62. Ba ANN, Cvijović I, Echenique JIR, Lawrence KR, Rego-Costa A, Liu X, Levy SF, and Desai MM. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. Nature 2019; 575:494–9

63. Aitkin M, Boys RJ, and Chadwick T. Bayesian point null hypothesis testing via the posterior likelihood ratio. Statistics and Computing 2005; 15:217–30

64. Smith I and Ferrari A. Equivalence between the posterior distribution of the likelihood ratio and a p-value in an invariant frame. Bayesian Analysis 2014; 9:939–62

65. Campbell F, Archer B, Laurenson-Schafer H, Jinnai Y, Konings F, Batra N, Pavlin B, Vandemaele K, Van Kerkhove MD, Jombart T, et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. Eurosurveillance 2021; 26:2100509

66. Ascensao J, Wetmore K, Good B, Arkin A, and Hallatschek O. Quantifying the Local Adaptive Landscape of a Nascent Bacterial Community. 2022

67. Pybus OG, Rambaut A, and Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 2000; 155:1429–37

68. Ho SY and Shapiro B. Skyline-plot methods for estimating demographic history from nucleotide sequences. Molecular Ecology Resources 2011; 11:423–34

69. Zinger T, Gelbart M, Miller D, Pennings PS, and Stern A. Inferring population genetics parameters of evolving viruses using time-series data. Virus Evolution 2019; 5:vez011

70. Ferrer-Admetlla A, Leuenberger C, Jensen JD, and Wegmann D. An approximate Markov model for the Wright–Fisher diffusion and its application to time series data. Genetics 2016; 203:831–46

71. Lumby CK, Zhao L, Breuer J, and Illingworth CJ. A large effective population size for established within-host influenza virus infection. eLife 2020; 9:e56915

72. Sobel Leonard A, Weissman DB, Greenbaum B, Ghedin E, and Koelle K. Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. Journal of Virology 2017; 91:e00171–17

73. Lau MS, Grenfell B, Thomas M, Bryan M, Nelson K, and Lopman B. Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. Proceedings of the National Academy of Sciences 2020; 117:22430–5

74. Kirkegaard JB and Sneppen K. Variability of individual infectiousness derived from aggregate statistics of COVID-19. medRxiv 2021

75. Riou J and Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Eurosurveillance 2020; 25:2000058

76. Laxminarayan R, Wahl B, Dudala SR, Gopal K, Mohan B C, Neelima S, Jawahar Reddy K, Radhakrishnan J, and Lewnard JA. Epidemiology and transmission dynamics of COVID-19 in two Indian states. Science 2020; 370:691–7

77. Adam DC, Wu P, Wong JY, Lau EH, Tsang TK, Cauchemez S, Leung GM, and Cowling BJ. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nature Medicine 2020; 26:1714–9

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

79. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, Liu X, Wei L, Truelove SA, Zhang T, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. The Lancet Infectious Diseases 2020; 20:911–9

Supplementary Information

Summary of existing methods for inferring the strength of genetic drift

There are currently four main types of methods for estimating the strength of genetic drift in pathogen transmission, which we summarize here for giving context to this study.

1. **Contact tracing** can directly measure superspreading by following the close contacts of infected individuals to measure the distribution of the number of secondary cases (the offspring number distribution) [2]. However, some secondary cases may be missed which can lead to measurement bias [20]. Additionally, it is challenging to trace multiple generations of transmission, so we miss important information on host contact network structure.

2. Another type of method fits disease prevalence over time to **branching process models** [42]. These models assume a particular distribution for the offspring number distribution (often a negative binomial distribution) and estimate the combination of parameters of the offspring number distribution along with growth rate that best fit the observed disease prevalence. External information about the growth rate can be used to constrain the parameters of the offspring number distribution.

3. **Phylogenetics** methods arrange genomics sequences into a tree based on genomic distance and either measure the distribution of lineage sizes (number of sequences in different parts of the tree) [19] or fit the rate at which branches in the tree coalescence to determine the effective population size [27, 67, 1, 68]. The effective population size is the population size that would reproduce the observed population dynamics under the idealized conditions of Wright-Fisher dynamics (discrete non- overlapping generations, a constant population size, and offspring determined by sampling with replacement from the previous generation). A lower effective population size indicates a higher level of genetic drift.

4. **Time series frequency methods** make use of a signature that genetic drift leaves in time series data, which is that it causes fluctuations in the lineage abundances. Higher amounts of genetic drift (lower effective population size) lead to larger fluctuations, and the magnitude of the fluctuations can be fit to determine the effective population size [69, 24] (Figure 1a). Time series methods have also been used extensively in population genetics [22, 70, 53, 23, 26, 25] and to estimate within-host effective population size [71] and between-host transmission bottleneck sizes [72].

Comparison to SEIR null model

In the main text, we compared the inferred $\tilde{N}_e(t)$ to an SIR model. However, there are likely more complex epidemiological dynamics describing SARS-CoV-2. Here we check the results for an SEIR model which includes a susceptible, exposed, infectious, and recovered class. The SEIR model is a good representation of the epidemiology of SARS-CoV-2 when PCR test positivity is closely associated with an infected host being infectious; the literature suggests that this is a good assumption for SARS-CoV-2 [16], but we also test this assumption below. The exposed class thus represents individuals before they are infectious and test positive. $\tilde{N}_e(t)$ for an SEIR model in equilibrium (number of infectious individuals is constant over time) is given by (see Methods for derivation):

$$\tilde{N_e}^{\text{SEIR,eq}}(t) \equiv \{N_e(t)\tau(t)\}^{\text{SEIR,eq}} = \frac{(E(t)+I(t))^2}{2R_t\gamma_I(t)}. \tag{70}$$

where $E(t)$ is the number of exposed individuals, $I(t)$ is the number of infectious individuals, R_t is the effective reproduction number, and γ_I is the rate at which infectious individuals stop being infectious. While this equation is derived under equilibrium conditions, we show using simulations that this equation accurately estimates $\tilde{N}_e(t)$ in non-equilibrium conditions after the peak of the pandemic (Figure S22); before the pandemic peak, this equation overestimates $\tilde{N}_e(t)$ but by less than one order of magnitude. Additionally, we show that calculating the $\tilde{N}_e(t)$ using the equation for an SIR model (Equation 1) when the dynamics are actually described by an SEIR model provides a lower bound on the actual $\tilde{N}_e(t)$. Thus, if the true dynamics of SARS-CoV-2 in England are actually SEIR dynamics, then the inference results shown in Figure 3c using

the SIR model should be an underestimate of the level of genetic drift; thus our main result that the literature values of superspreading do not sufficiently explain our results should still hold.

In reality, it may also be the case that some people test positive in a PCR test before they become infectious. To test the impact of this possibility on our results, in our simulations we recorded both exposed and infectious individuals as testing positive. We then calculated the SEIR model $\tilde{N}_e(t)$ numerically as described in "Calculating the effective population size for an SIR or SEIR model" assuming that $I(t)$ includes both infectious and exposed individuals (Figure S23). We find that the numerical solutions give slightly higher $\tilde{N}_e(t)$ as compared with the true analytical solutions; however, the numerical solutions to the SEIR and SIR models bound the inferred $\tilde{N}_e(t)$. Thus we also expect that our main result that the literature values of superspreading do not sufficiently explain our results should still hold in this scenario.

To calculate the SEIR model $\tilde{N}_e(t)$ for the actual data, for the number of infectious individuals, we used the number of positive individuals estimated from the UK Office for National Statistics' COVID-19 Infection Survey [31], which is a household surveillance study that reports positive PCR tests, regardless of symptom status. We used the measured effective reproduction number in England reported by the UK Health Security Agency [38]. We found that $\tilde{N_e}^{\text{SEIR}}(t)$ is very similar to the number of positives because the effective reproduction number in England was very close to 1 across time. To calculate $\tilde{N_e}^{\text{SEIR}}(t)$ for each variant or group of lineages, we rescaled the population-level $I(t)$ and R_t based on the fraction of each variant in the population and the relative differences in reproduction numbers between variants (see Methods). We then calculated the scaled true population size, $\tilde{N}(t) \equiv N(t)\tau(t)$, for the SEIR model by multiplying by the variance in offspring number, σ^2, for the SEIR model [41]

$$\tilde{N}^{\text{SEIR}}(t) = \tilde{N_e}^{\text{SEIR}}(t)\{\sigma^2\}^{\text{SEIR}} \tag{71}$$

$$\{\sigma^2\}^{\text{SEIR}} = 2. \tag{72}$$

Overall, the inferred $\tilde{N}_e(t)$ is lower than $\tilde{N}^{\text{SEIR}}(t)$ by a time-dependent factor that varies between 70 and 2000 (Figure S14), suggesting high levels of genetic drift in England across time, which is consistent with what we find with an SIR model (Figures 2 and S13). Also similarly to in the case with an SIR model, the ratio of $\tilde{N}^{\text{SEIR}}(t)$ to the inferred $\tilde{N}_e(t)$ for Alpha decreased over time, suggesting that the stochasticity in the transmission of Alpha decreased over time.

Application to COG-UK data by regions in England

The inference of effective population size can also reveal information about the well-mixed or spatially-structured nature of transmission dynamics within England. This can be done by inferring effective population size at smaller geographical scales within England. If the transmission dynamics were completely well-mixed, then we would expect $\tilde{N}_e(t)$ to be the same across regions and compared to England. On the other hand, if the transmission dynamics were completely spatially segregated (i.e. transmission only occurs within the defined geographical areas, but not between them) and the dynamics were the same in each region, we would expect that the ratio $\tilde{N_e}^{\text{SIR}}(t)/\tilde{N_e}^{\text{inf}}(t)$ to be the same across regions.

The geographical areas that we used were the 9 regions of England: East Midlands, East of England, London, North East, North West, South East, South West, West Midlands, and Yorkshire and The Humber. We looked at sequences from each region, repeating the analysis described above, and inferred the scaled effective population size (Figure S1). We observe a lower $\tilde{N}_e(t)$ for in the region than in England for Delta in all regions, for Alpha in all regions except North East (where there was not enough data), and for B.1.177 in all regions except North East. For lineages pre-B.1.177, the inferred $\tilde{N}_e(t)$ is not significantly differnt in the region than in England. These results suggest that the dynamics are not well-mixed during the B.1.177, Alpha, and Delta waves.

The calculated SIR model $\tilde{N_e}^{\text{SIR}}(t)$ (Figure S2) and the number of positive individuals in each region (Figure S3) were 1-2 orders of magnitude higher than the inferred $\tilde{N}_e(t)$, suggesting high levels of genetic drift. The ratios of the the SIR model $\tilde{N}_e(t)$ and the number of positives to the inferred $\tilde{N}_e(t)$ in the regions were similar to one another and to that seen in England as a whole, consistent with a scenario where the dynamics are spatially-structured and the extent of stochasticity in transmission is similar across regions.

Similarly to in England as a whole, the inferred measurement noise in each region was mostly indistinguishable from uniform sampling except for in a few timepoints (Figure S4).

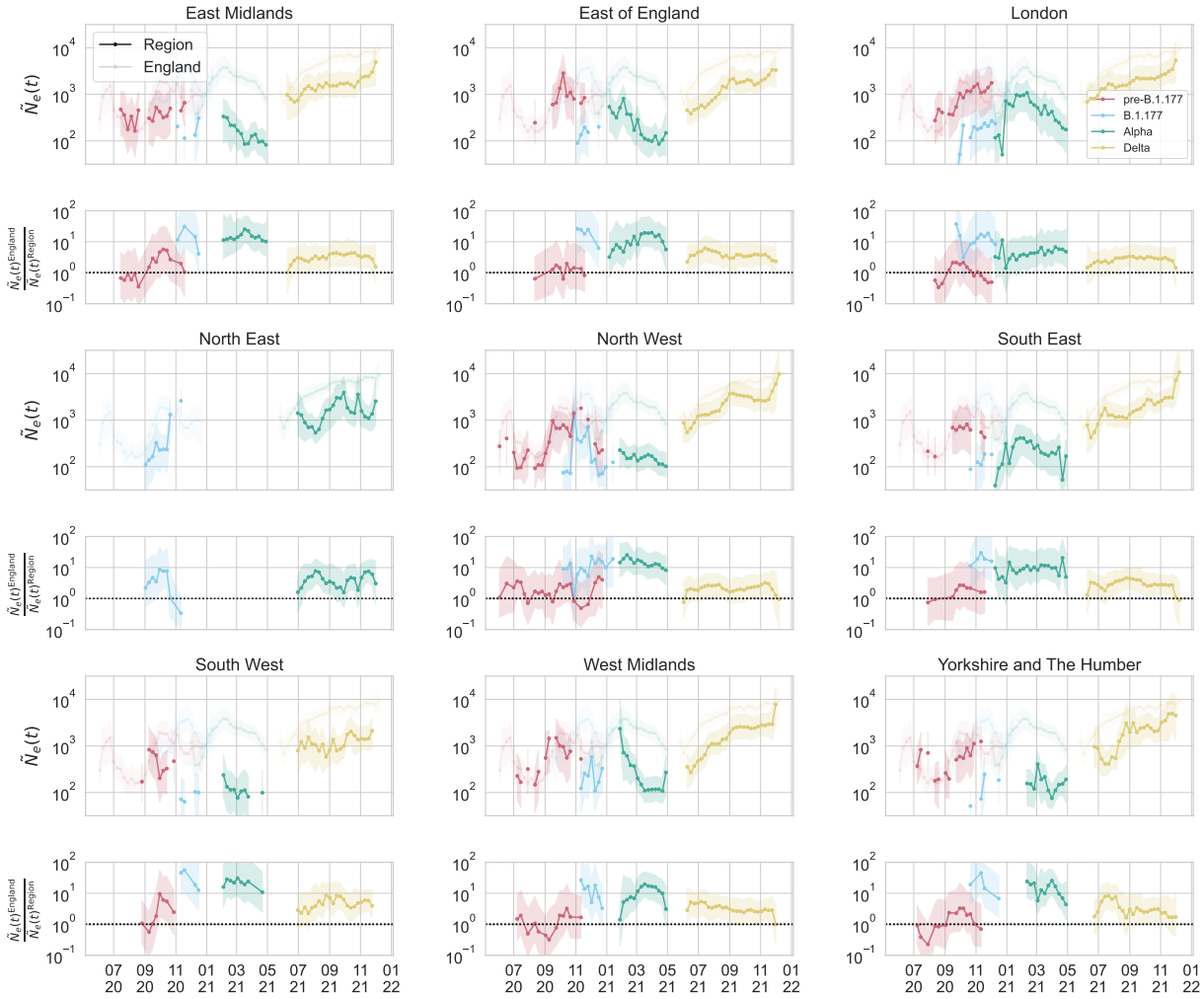

Figure S1: Inferred effective population size in regions of England. (Top panels) Inferred $\tilde{N}_e(t)$ of pre-B.1.177 lineages, B.1.177, Alpha, and Delta for each region of England. The inferred $\tilde{N}_e(t)$ for England as a whole is shown for reference. Shaded regions show 95% confidence intervals (see Methods). (Bottom panels) The ratio between the inferred $\tilde{N}_e(t)$ of England and that of the region for each variant. A horizontal dashed line indicates a ratio of 1 (i.e. $\tilde{N}_e(t)$ is the same in that region of England and England as a whole). Shared regions show the minimum and maximum possible values of the ratio from the combined error intervals of the numerator and denominator (thus, not corresponding to a specific confidence interval range).

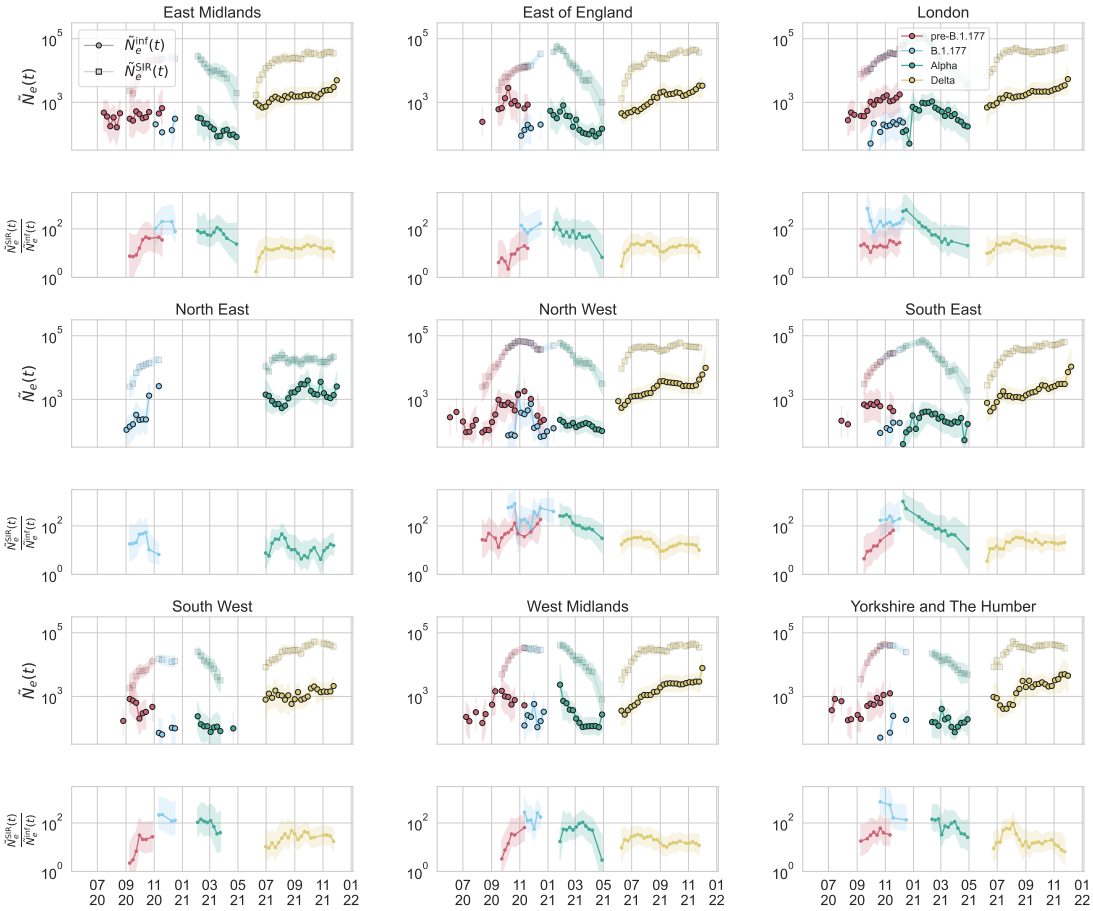

Figure S2: Inferred scaled effective population size by region in England, compared to that of an SIR model as calculated using the observed number of positives at the community level in that region reported by the COVID-19 Infection Survey [31] and the observed effective reproduction number in that region reported by the UK Health Security Agency [38].

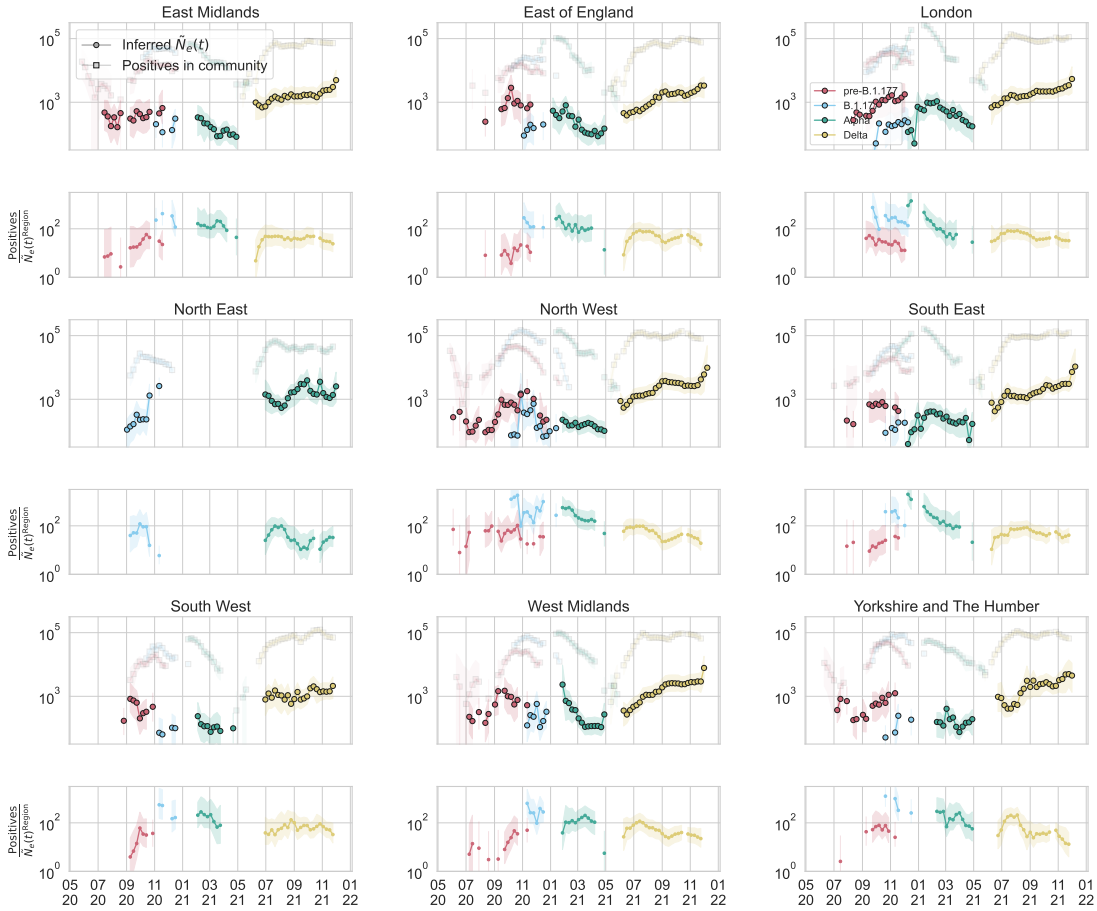

Figure S3: Inferred scaled effective population size by region in England, compared to number of positives at the community level in that region reported by the COVID-19 Infection Survey [31].

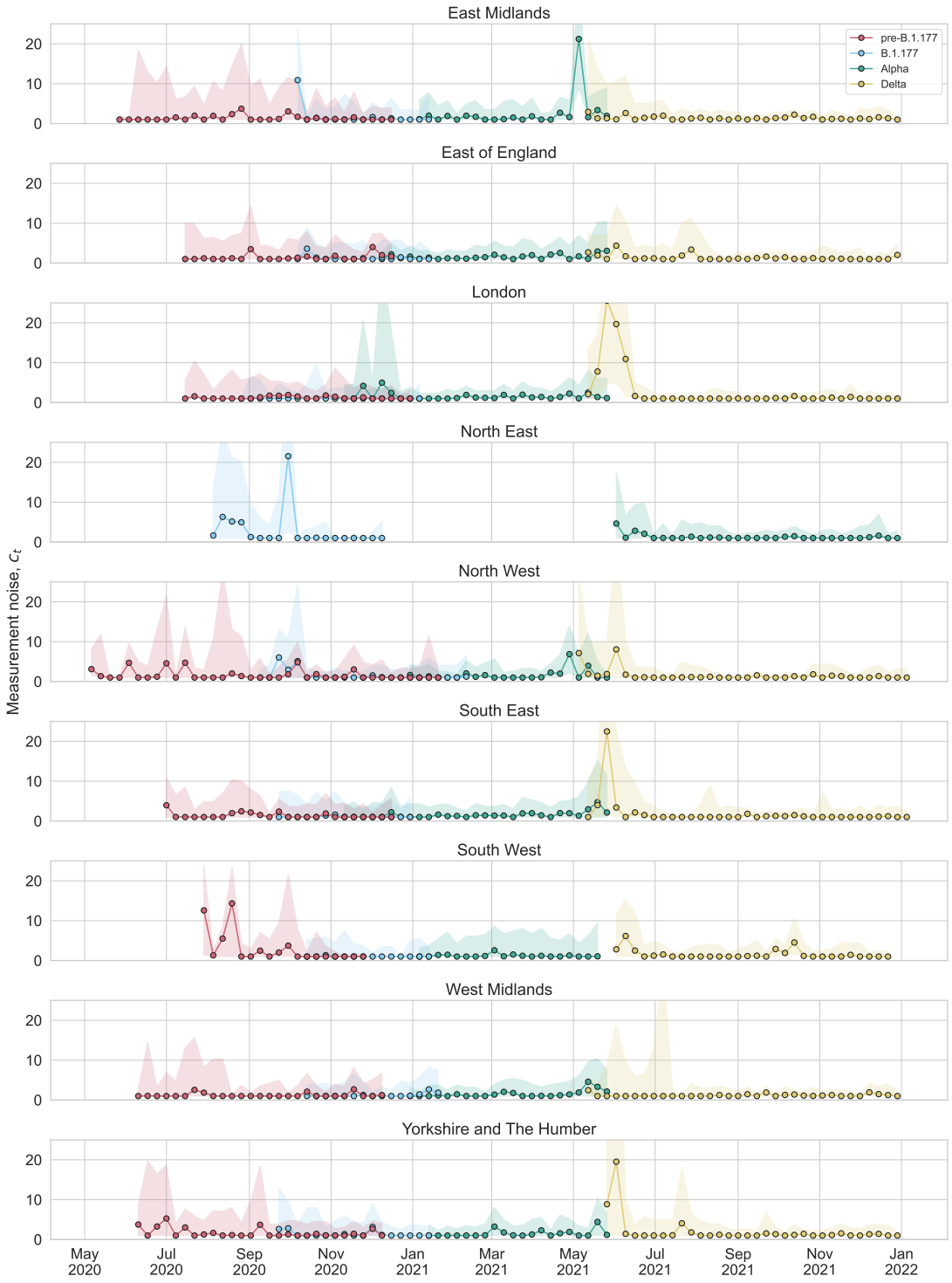

Figure S4: Inferred measurement noise by region in England.

Additional supplementary tables and figures

Date	Location	Method	$\langle Z \rangle$	Var(Z)	k	Reference
Beginning of pandemic to February 27 2020	Worldwide excluding China	Branching process model of number of imported and local cases	2.5 (,)	65 (33.75, 127.5)	0.1 (0.05, 0.2)	[42]
March 1 to May 3 2020	Georgia (USA)	Spatiotemporal transmission model fit to multiple data sources	2 (0.5, 3.5)	12.26 (0.88, 101.5)	0.39 (0.125, 0.65)	[73]
March 1 to November 1 2020	Denmark	Model fitting the case numbers across multiple regions	1.1 (0.8, 1.4)	12.1 (4.36, 25.9)	0.11 (0.08, 0.18)	[74]
Beginning of pandemic until January 18 2020	China (Wuhan)	Stochastic simulations fit to infected cases	2.2 (1.4, 3.8)	11.16 (1.68, 1035.2)	0.54 (0.014, 6.95)	[75]
August to September 2020	UK	Model using empirical viral load trajectories and contact numbers	1.21 (0.84, 2.51)	7.07 (2.65, 44.51)	0.25 (0.15, 0.39)	[43]
May 15 to August 1 2020	Tamil Nadu and Andhra Pradesh (India)	Contact tracing and incidence	1.25 (1.1, 1.4)	4.31 (3.43, 5.4)	0.51 (0.49, 0.52)	[76]
January to February 2021	UK	Model using empirical viral load trajectories and contact numbers	0.54 (0.4, 1.03)	1.42 (0.66, 9.19)	0.33 (0.13, 0.61)	[43]
January 23 to April 28 2020	Hong Kong	Contact tracing	0.58 (,)	1.36 (,)	0.43 (,)	[77]
January 16 to April 3 2020	Hunan (China)	Contact tracing	0.4 (0.35, 0.47)	0.93 (0.66, 1.43)	0.3 (0.23, 0.39)	[78]
January 14 to February 12 2020	Shenzhen (China)	Contact tracing	0.4 (0.3, 0.5)	0.68 (0.38, 1.21)	0.58 (0.35, 1.18)	[79]

Table S1: Overdispersion values from the literature ordered from highest to lowest variance in offspring number. Any error intervals that are reported are taken from the reference (sometimes defined differently). The estimate taken from Ref. [43] assumes no self-isolation upon symptom onset and no testing; lifting these assumptions leads to similar or lower overdispersion.

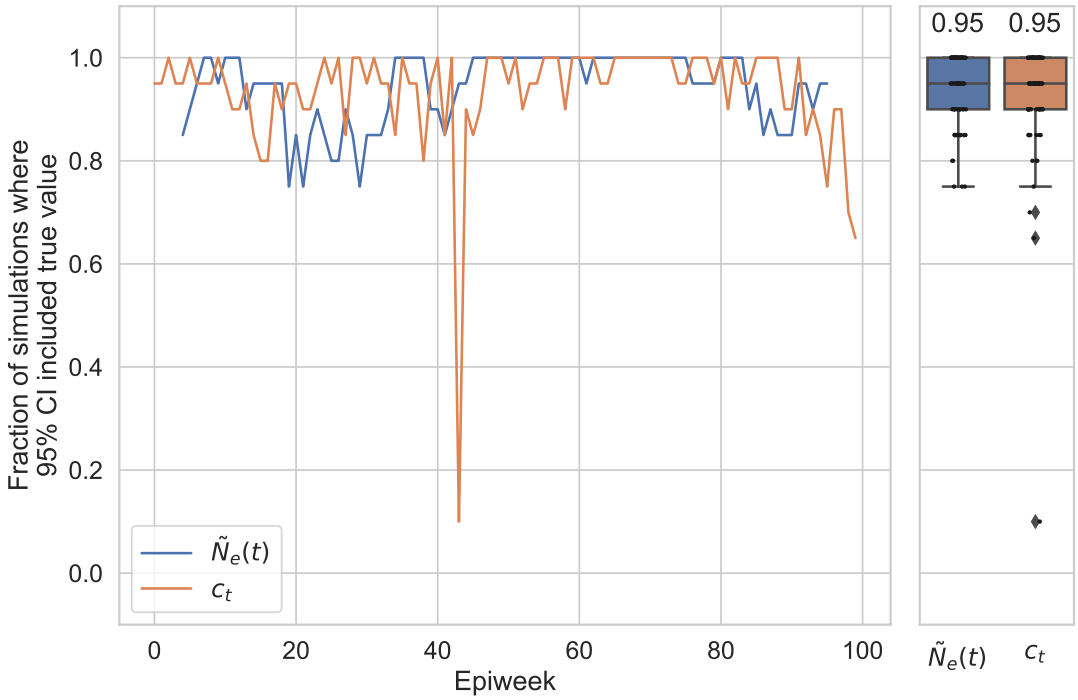

Figure S5: The fraction of simulations (20 total) where the inferred 95% confidence interval for $\tilde{N}_e(t)$ or c included the true value (left) by timepoint and (right) for all timepoints. (Right) Boxes indicate the quartiles and the line inside the box (and number above) indicates the median. Whiskers indicate the extreme values excluding outliers. Simulation parameters are specified in the Methods and Figure 1, which shows a single simulation instance. For the inference, we created superlineages randomly 20 times.

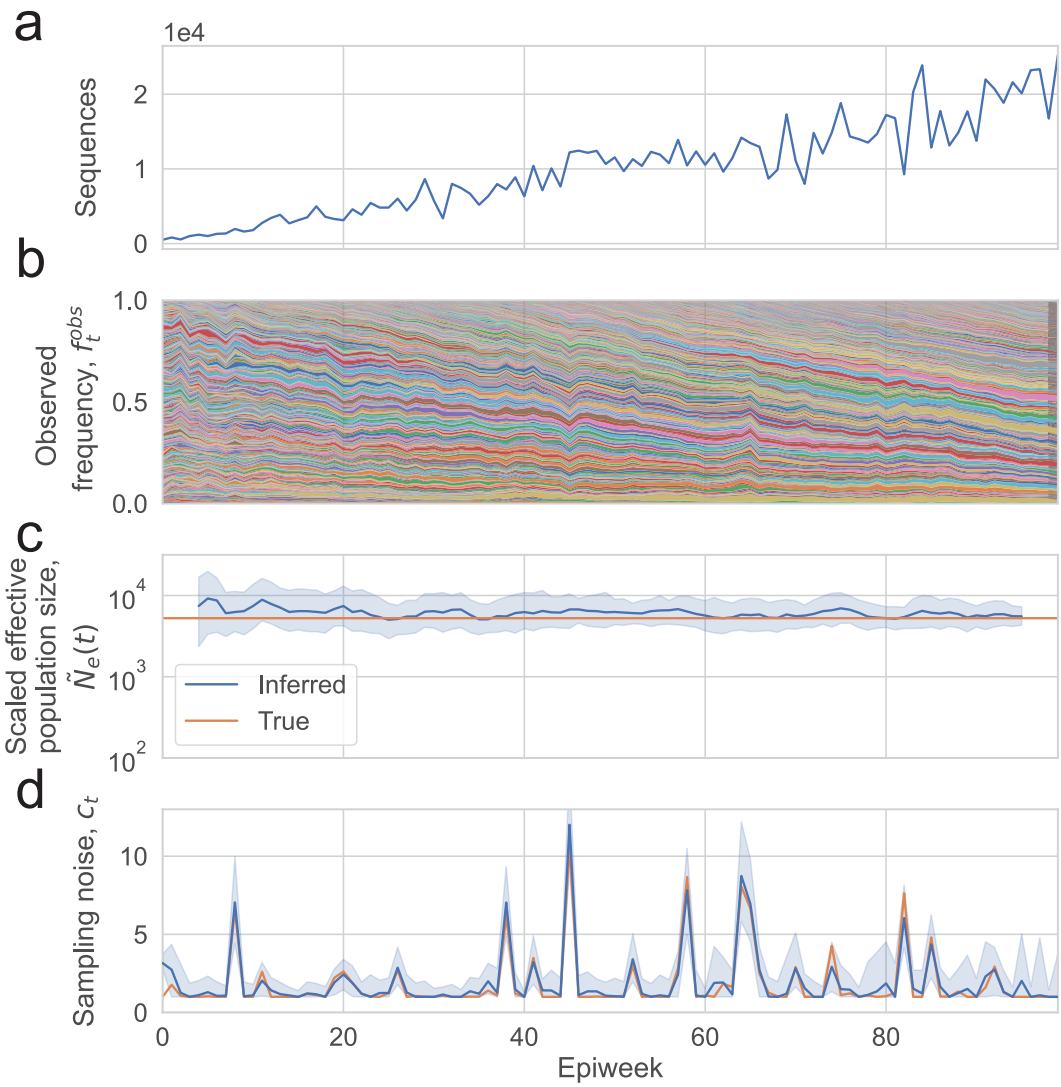

Figure S6: Wright-Fisher simulations where $\tilde{N}_e(t)$ is constant over time, and the inferred $\tilde{N}_e(t)$ and c_t. (a) Number of sequences sampled. (b) Simulated lineage frequency trajectories. (c) Inferred effective population size ($\tilde{N}_e(t)$) on simulated data compared to true values. (d) Inferred measurement noise (c_t) on simulated data compared to true values. In (c) the shaded region shows the 95% confidence interval calculated using the posterior, and in (d) the shaded region shows the 95% confidence interval calculated using bootstrapping (see Methods).

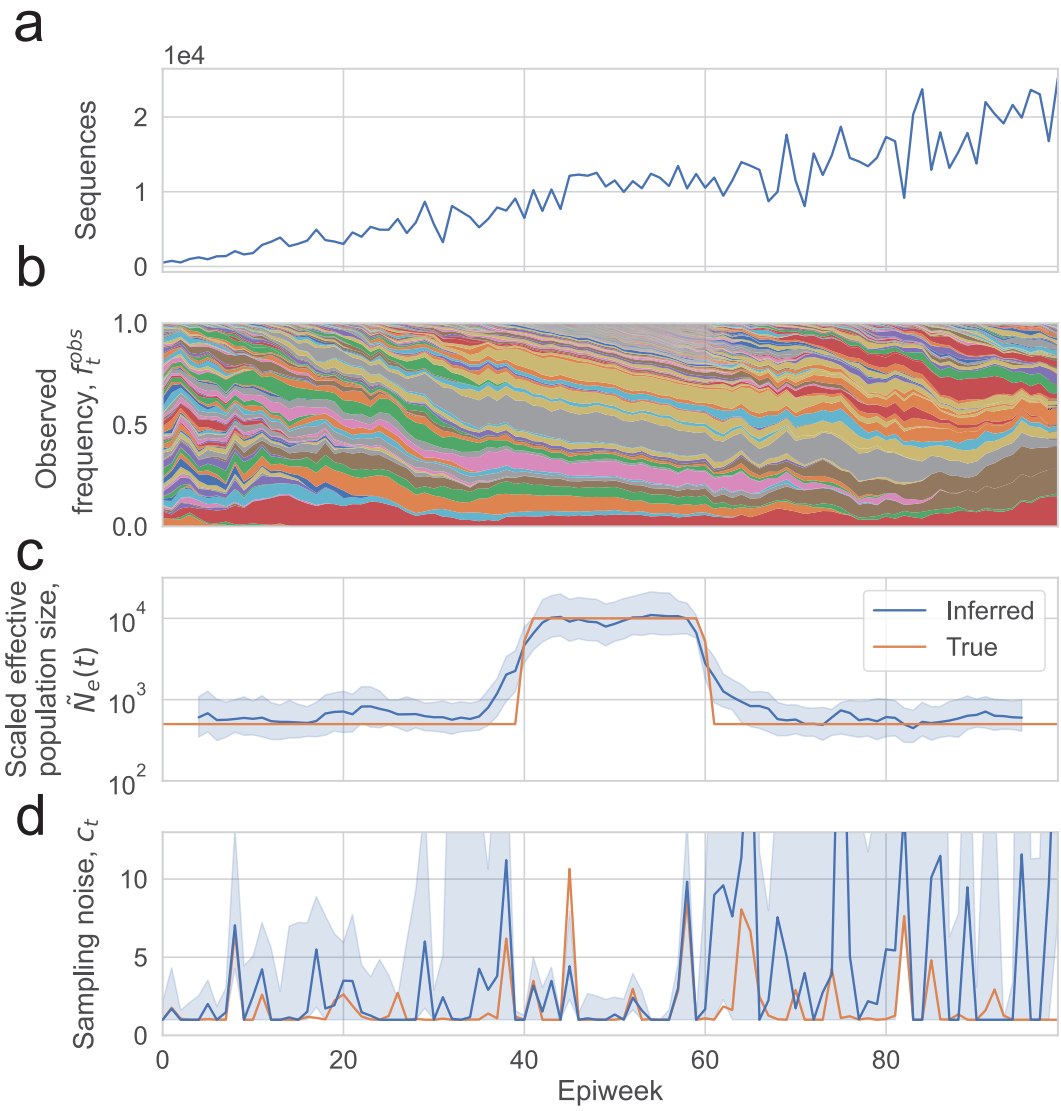

Figure S7: Wright-Fisher simulations where $\tilde{N}_e(t)$ changes over time according to a rectangular function, and the inferred $\tilde{N}_e(t)$ and c_t. (a) Number of sequences sampled. (b) Simulated lineage frequency trajectories. (c) Inferred effective population size ($\tilde{N}_e(t)$) on simulated data compared to true values when c_t is jointly inferred and when c_t is fixed at 1 (uniform sampling). (d) Inferred measurement noise (c_t) on simulated data compared to true values. In (c) the shaded region shows the 95% confidence interval calculated using the posterior, and in (d) the shaded region shows the 95% confidence interval calculated using bootstrapping (see Methods).

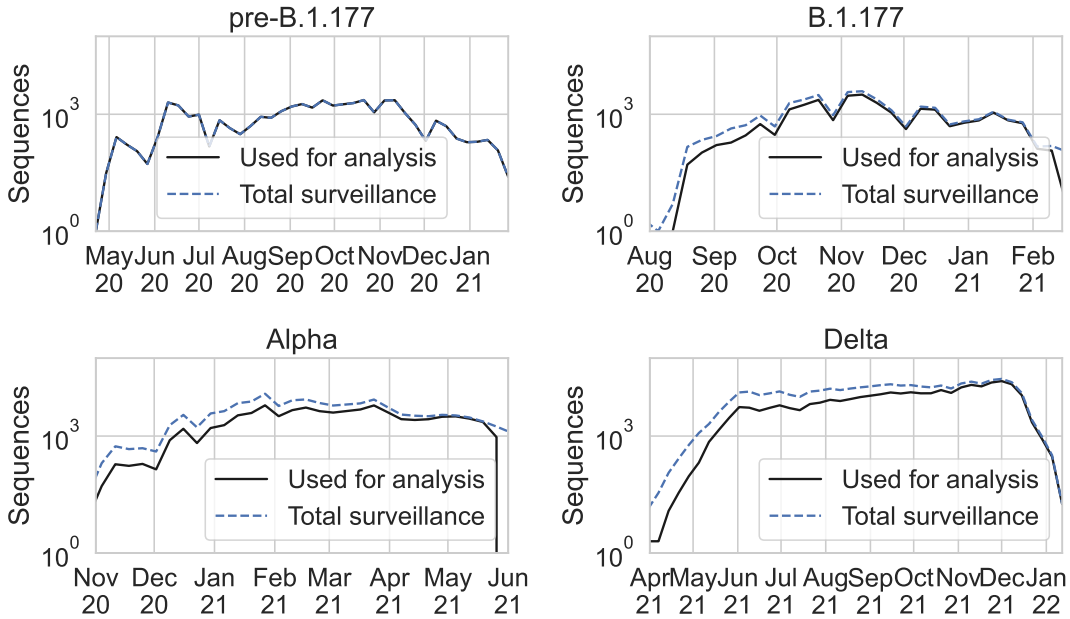

Figure S8: Total number of surveillance sequences of each variant in the metadata from COG-UK downloaded on January 16, 2022 and the number of sequences used in the analysis for each variant or group of lineages (determined by the number of sequences included in the tree, and the number of sequences which could be grouped into sublineages based on the procedure described in the Methods).

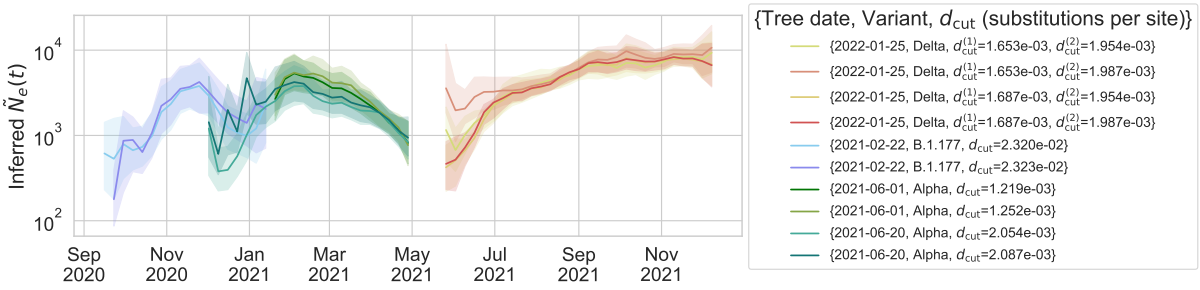

Figure S9: Varying the date of the tree downloaded from COG-UK and the depth at which the tree is cut for creating lineages (d_{cut}, see Methods) does not substantially change the inferred scaled effective population size. The the tree date and depth used in the main text are {2021-02-22, B.1.177, $d_{\text{cut}} = 2.323 \cdot 10^{-2}$}, {2021-06-20, Alpha, $d_{\text{cut}} = 2.054 \cdot 10^{-3}$}, {2022-01-25, Delta, $d_{\text{cut}}^{(1)} = 1.687 \cdot 10^{-3}$, $d_{\text{cut}}^{(2)} = 1.954 \cdot 10^{-3}$}. The color of the lines for the parameters that were used in the main text are the same as those shown in Figure 2.

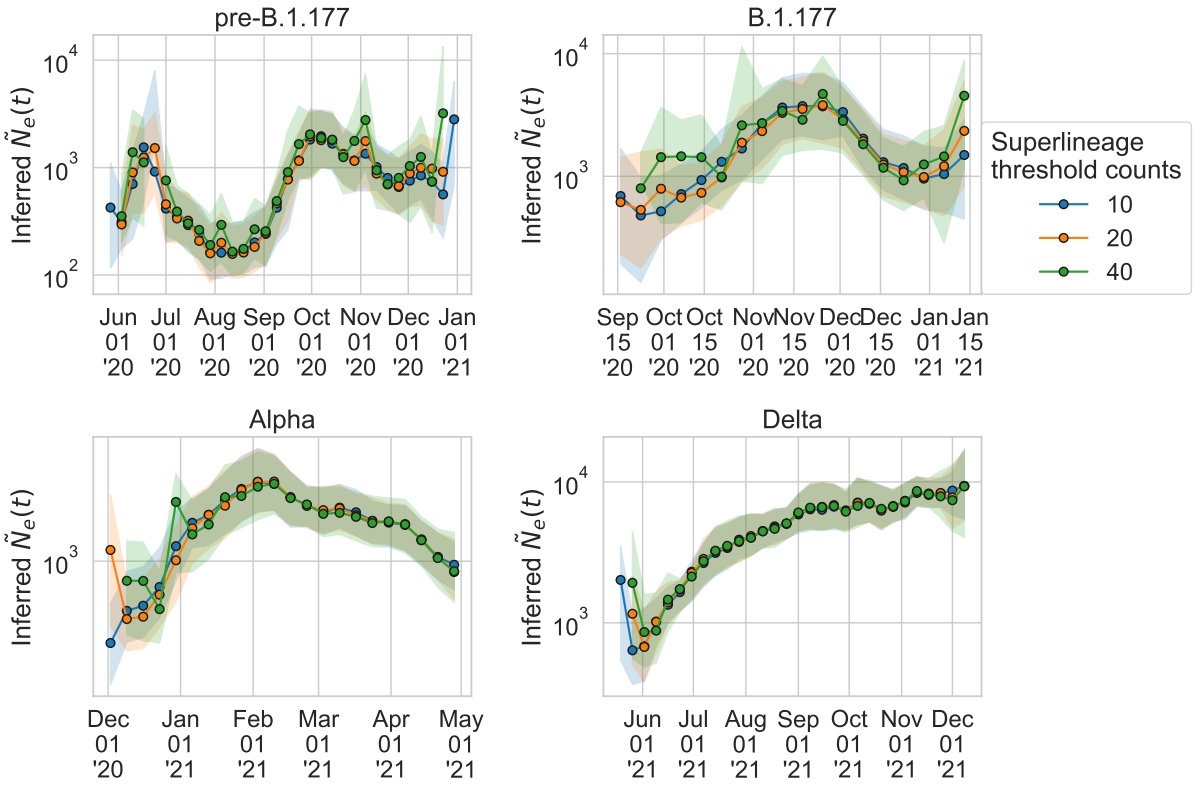

Figure S10: Varying the threshold counts for forming superlineages (see Methods) does not substantially change the inferred scaled effective population size. The superlineage threshold counts used in the main text is 20.

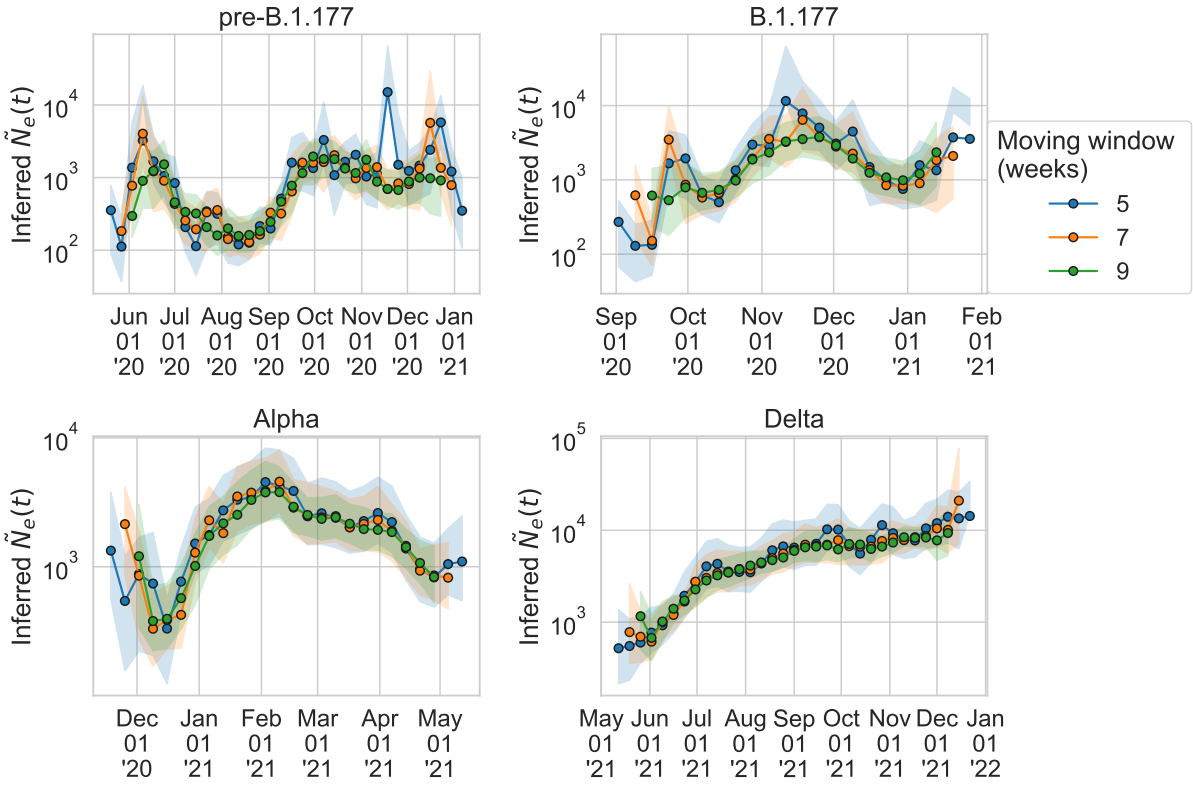

Figure S11: Varying the number of weeks in the moving window does not substantially change the inferred scaled effective population size. The size of the moving window used in the main text is 9 weeks.

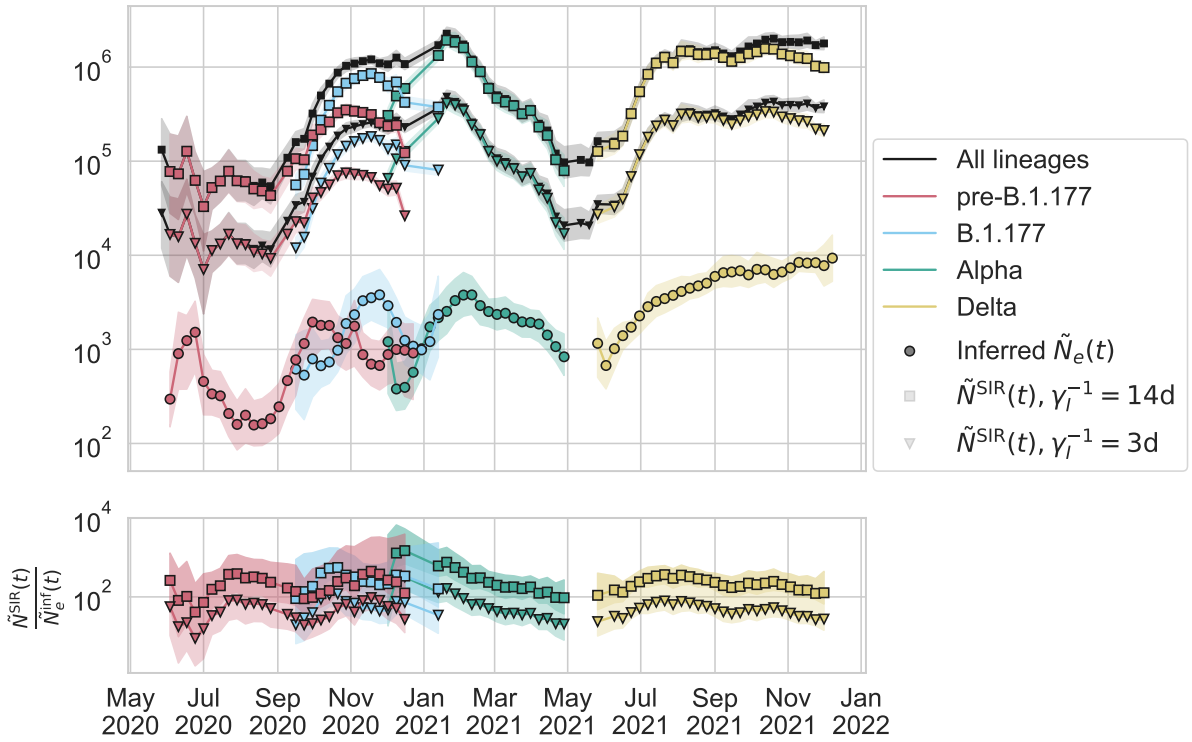

Figure S12: Varying the rate of transitioning from infected to recovered within literature ranges (γ_I =3 to 14 days) used for calculation of the SIR model $\tilde{N}_e(t)$ (Methods) does not substantially decrease the observed ratio $\tilde{N}_e^{\mathrm{SIR}}(t)/\tilde{N}_e^{\mathrm{inf}}(t)$.

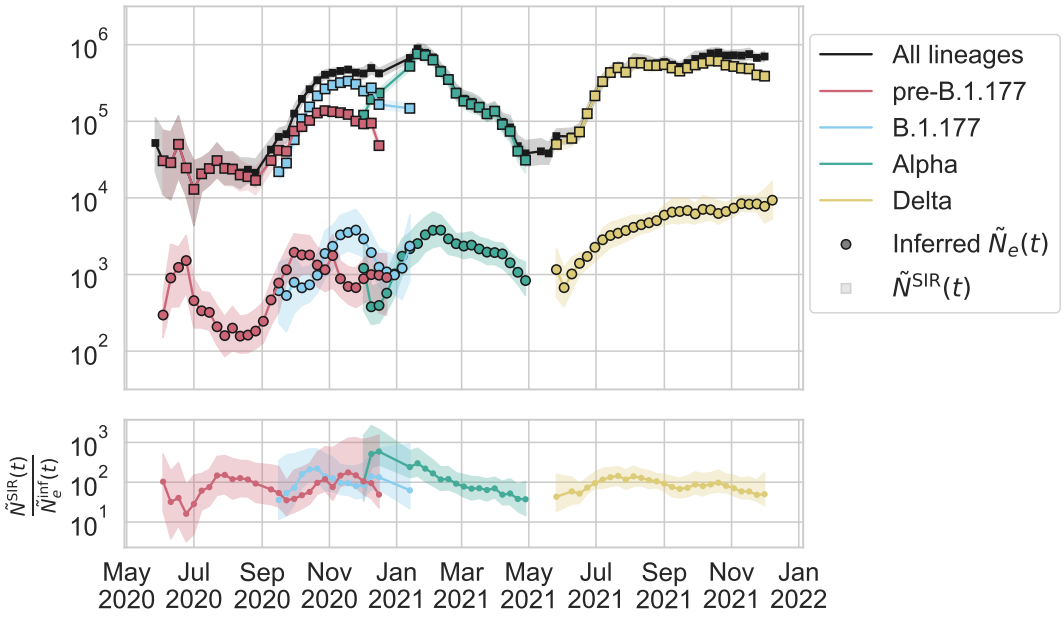

Figure S13: Inferred scaled effective population size compared to the SIR model scaled population size calculated using the observed number of positive individuals in England (see Methods).

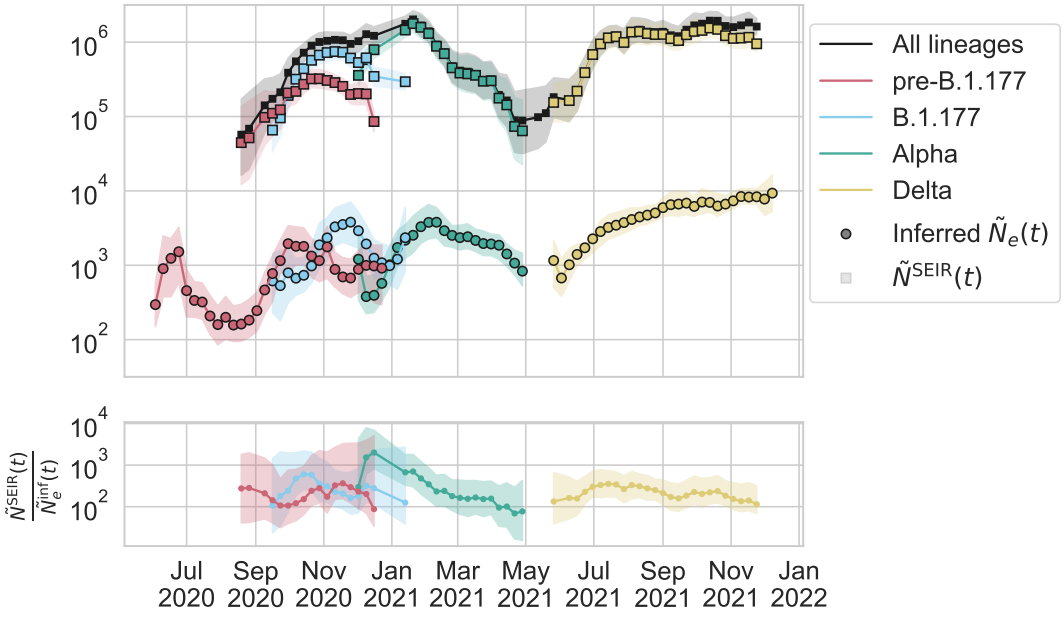

Figure S14: Inferred scaled effective population size compared to the SEIR model scaled population size calculated using the observed number of positive individuals in England (see Methods).

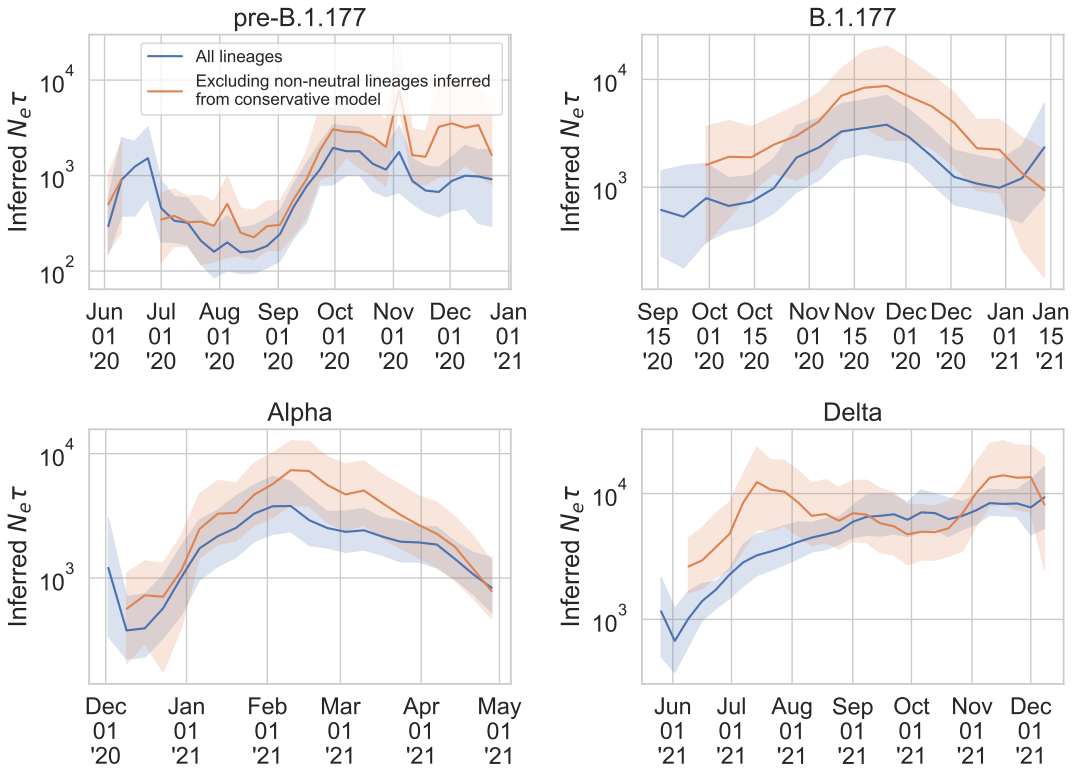

Figure S15: Inferred scaled effective population size of all lineages compared to that when excluding non-neutral lineages detected by conservative method (assumes no genetic drift).

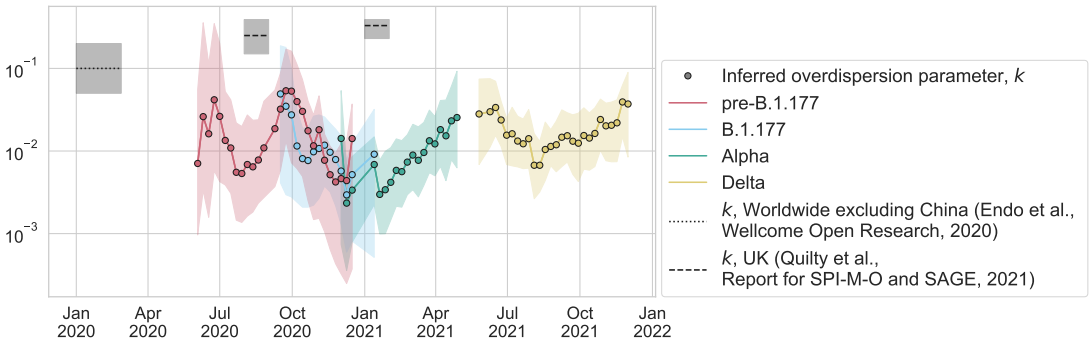

Figure S16: Same as Figure 3c, but plotting the overdispersion parameter, $k = \frac{R_t}{\frac{\sigma^2}{R_t}-1}$, where R_t is the effective reproduction number and σ^2 is the variance in offspring number. The circles show the inferred overdispersion parameter if we assume there is only superspreading and no deme structure. For the inferred overdispersion parameter, the estimated effective reproduction number in England by variant (see Methods) is used for R_t, and the ratio between the SIR model population size and the inferred effective population size is used for σ^2. The shaded area for the inferred overdispersion parameter k gives an estimate of the error and is calculated by combining minimum or maximum values of the individual parameters; note that this does not correspond to a particular confidence interval.

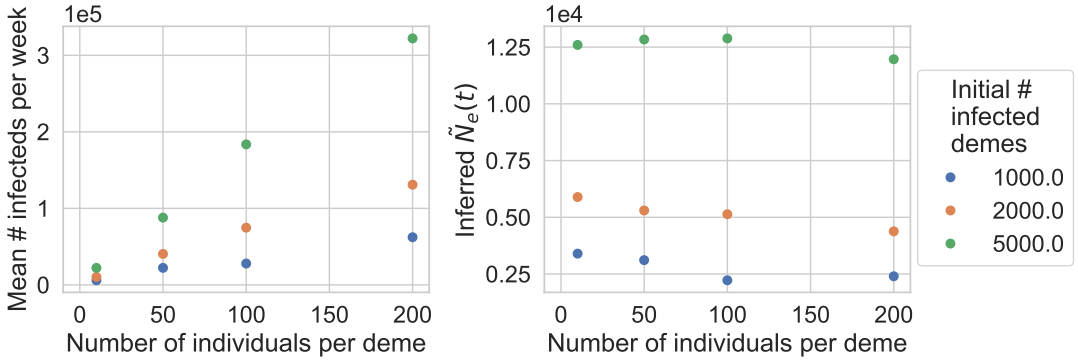

Figure S17: Simulations of deme structure (described in main text and Methods). (a) The mean number of infected individuals per week from Weeks 42 to 50. (b) The inferred $\tilde{N}_e(t)$ using lineage trajectories from Weeks 42 to 50.

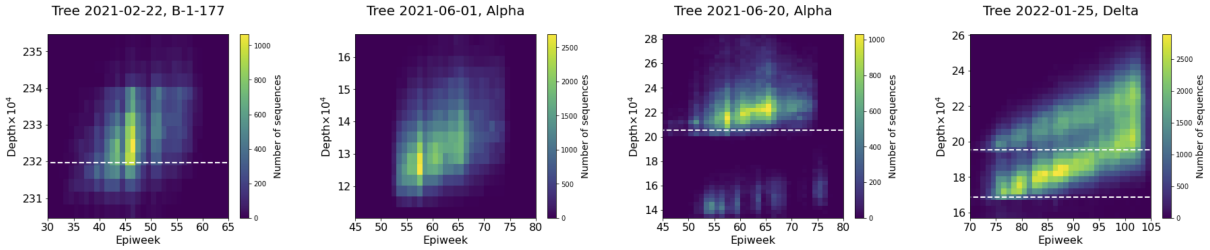

Figure S18: Sample epiweeks versus tree depths. In a phylogenetic tree, the number of sequences (leaf nodes) of a focal variant that fall within specific epiweek and tree depth ranges is counted and summarized as a two-dimensional histogram. The tree depth is the substitution rate measured in units of substitutions per site, with respect to the most recent common ancestor. From left to right, the phylogenetic tree (specified by date created by COG-UK, using the sequences available at the time) and focal variant are {2021-02-22, B-1-177}, {2021-06-01, Alpha}, {2021-06-20, Alpha}, and {2022-01-25, Delta}. Weeks are counted from 2019-12-29. The dashed horizontal lines indicate the values of d_{cut} ($d_{\text{cut}}^{(1)}$ and $d_{\text{cut}}^{(2)}$ for the Delta variant) used for the results presented in the main text.

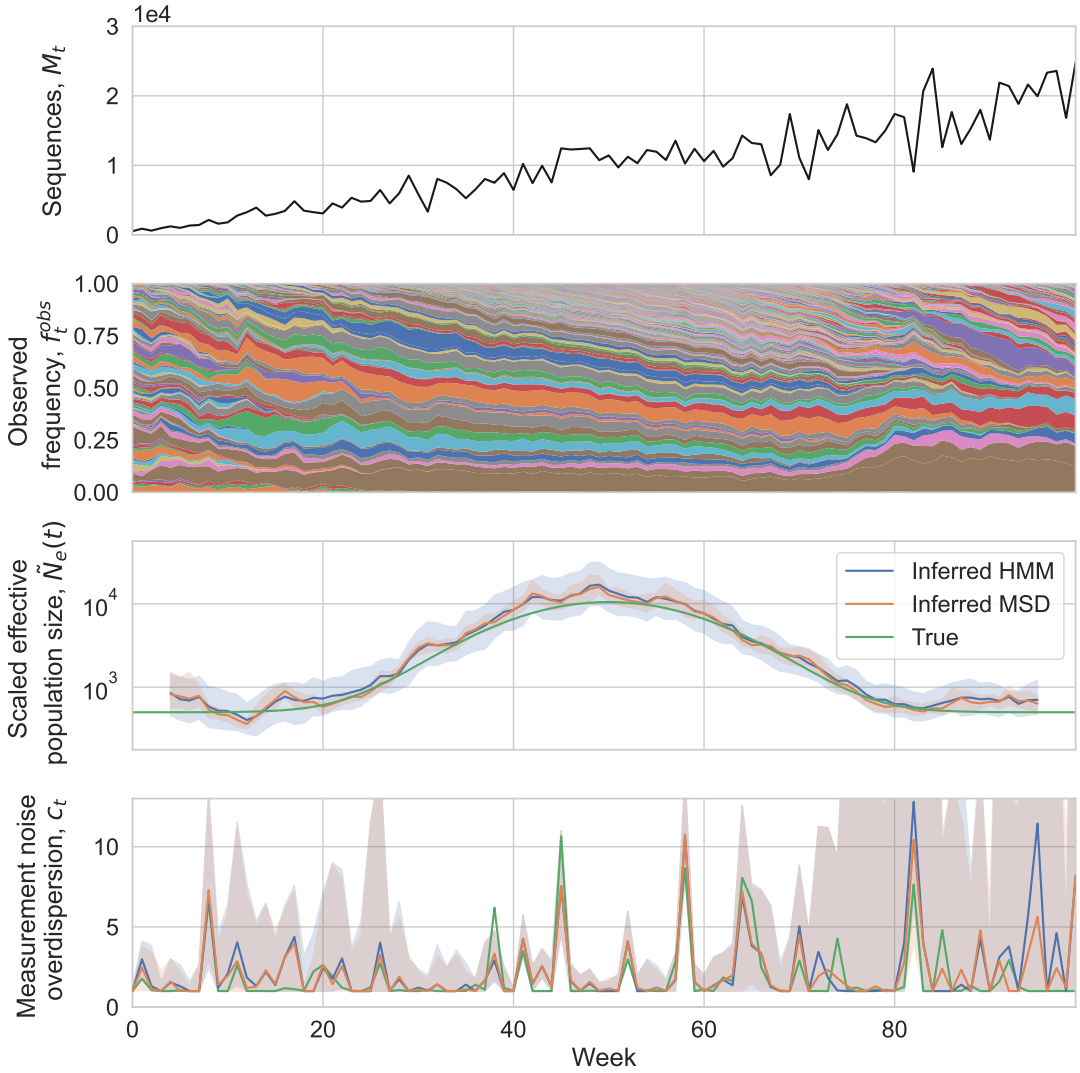

Figure S19: Comparing the inferred $\tilde{N}_e(t)$ and c_t in Wright-Fisher simulations using the method of moments and maximum likelihood estimation approaches (see Methods). (a) Number of sequences sampled. (b) Simulated lineage frequency trajectories. (c) Inferred effective population size ($\tilde{N}_e(t)$) on simulated data using the method of moments (MSD, for mean squared displacement) and maximum likelihood (HMM, for Hidden Markov Model) estimation approaches compared to true values. The shaded region shows the 95% confidence interval of the inferred values. The confidence interval using the method of moments approach was calculated by taking the middle 95% of values when bootstrapping over the superlineages. The confidence interval using the maximum likelihood estimation approach was determined using the posterior (see Methods) and takes into account joint errors in c_t and $\tilde{N}_e(t)$. (d) Inferred measurement noise (c_t) on simulated data using the method of moments and maximum likelihood estimation approaches compared to true values. The shaded region shows the 95% confidence interval calculated using bootstrapping (see Methods).

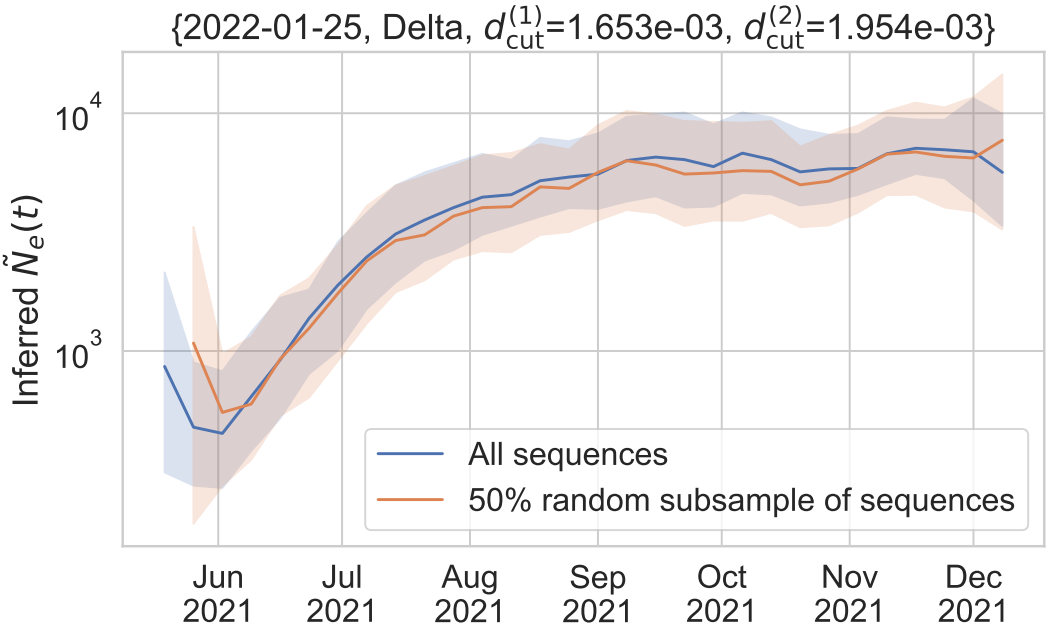

Figure S20: Randomly subsampling half of the Delta sequences used for the analysis does not substantially change the inferred scaled effective population size.

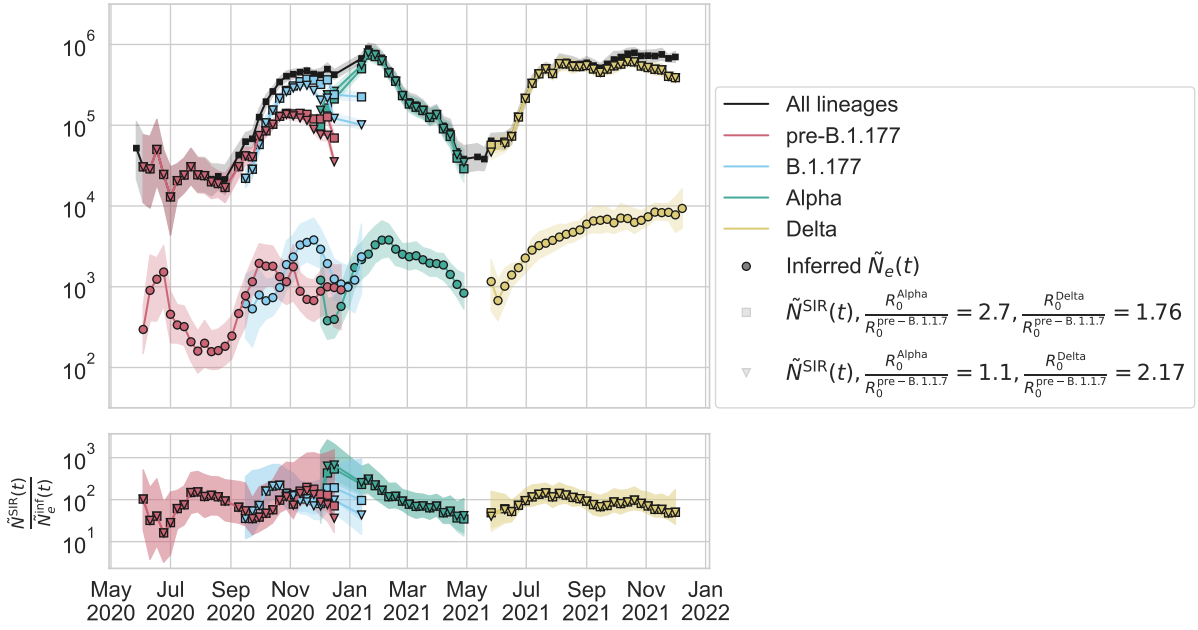

Figure S21: Varying the values of the basic reproduction number within literature ranges ($\frac{R_0^{\text{Alpha}}}{R_0^{\text{pre-B.1.1.7}}} = 1.1-2.7$ [17], $\frac{R_0^{\text{Delta}}}{R_0^{\text{pre-B.1.1.7}}} = 1.76-2.17$ [65]) used for calculation of the SIR model $\tilde{N}_e(t)$ by variant (Methods) does not substantially affect the calculated $\tilde{N}^{\text{SIR}}(t)$.

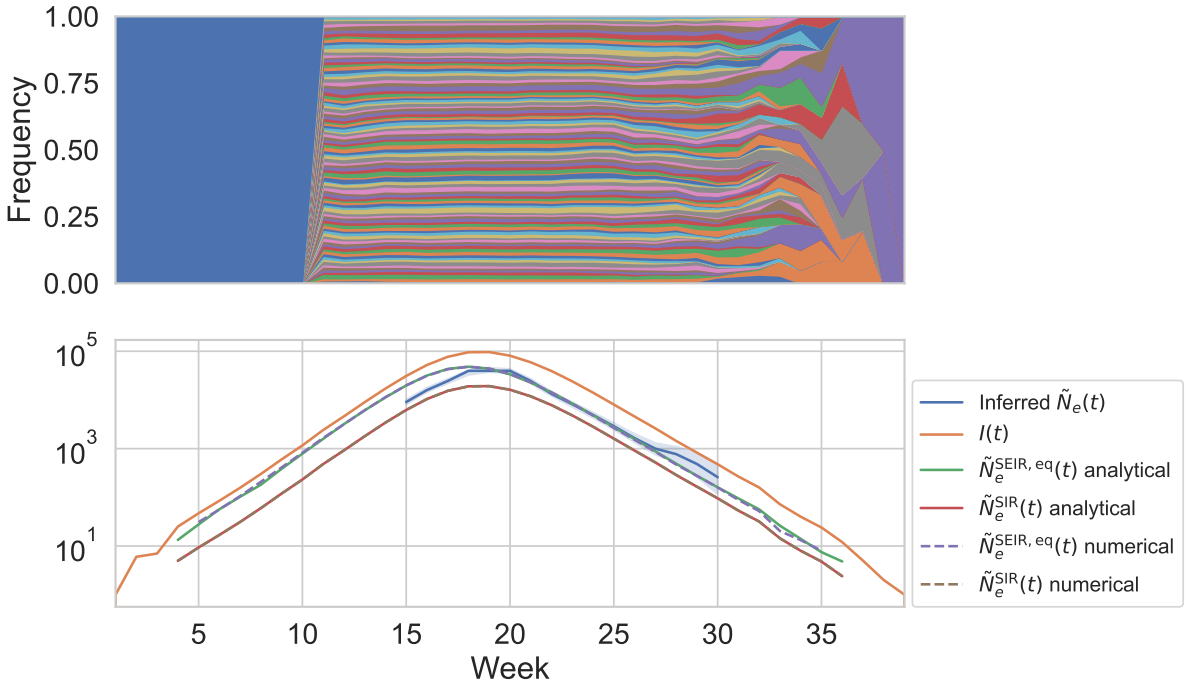

Figure S22: Simulations of stochastic SEIR dynamics without measurement noise, and comparison of the inferred $\tilde{N}_e(t)$ to Equations 1 and 51 when the reported positive individuals include only the infectious individuals. (Top) Muller plot of simulated infectious individuals' lineage trajectories (simulations described in Methods). Infectious individuals are randomly assigned a lineage in week 11, and individuals that they transmit to are infected with the same lineage. The blue lineage before week 11 indicates the infectious individuals that existed before lineages were assigned. (Bottom) Comparison of the inferred $\tilde{N}_e(t)$ using the lineage trajectories shown in the top panel to the number of infectious individuals $I(t)$, Equation 51 (SEIR model $\tilde{N}_e(t)$ at equilibrium), and Equation 1 (SIR model $\tilde{N}_e(t)$) calculated analytically or numerically as described in the Methods. The numerical solutions give the same results as the analytical solutions.

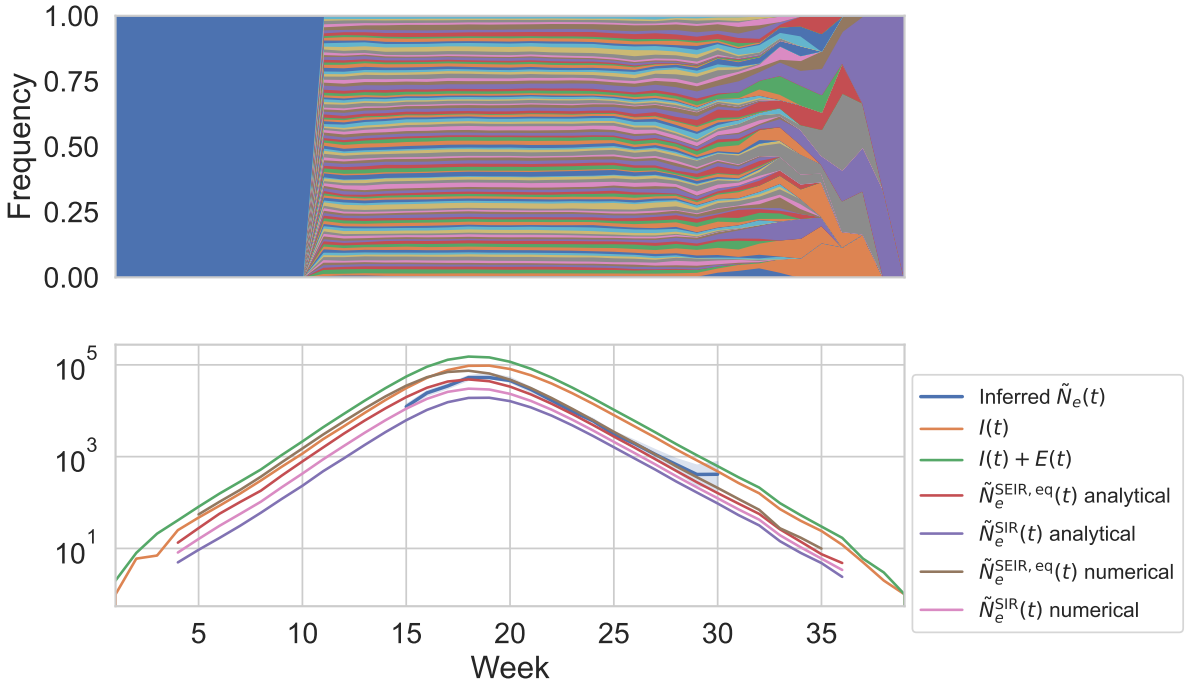

Figure S23: Simulations of stochastic SEIR dynamics without measurement noise, and comparison of the inferred $\tilde{N}_e(t)$ to Equations 1 and 51 when the reported positive individuals include both infectious and exposed individuals. (Top) Muller plot of simulated infectious and exposed individuals' lineage trajectories (simulations described in Methods). Infectious and exposed individuals are randomly assigned a lineage in week 11, and individuals that they transmit to are infected with the same lineage. The blue lineage before week 11 indicates the infectious and exposed individuals that existed before lineages were assigned. (Bottom) Comparison of the inferred $\tilde{N}_e(t)$ using the lineage trajectories shown in the top panel to the number of infectious individuals $I(t)$, the sum of the number of infectious and exposed individuals $I(t)+E(t)$, Equation 51 (SEIR model $\tilde{N}_e(t)$), and Equation 1 (SIR model $\tilde{N}_e(t)$) calculated analytically or numerically as described in the Methods. The numerical solutions give slightly higher $\tilde{N}_e(t)$ as compared with the analytical solutions; however, the numerical solutions to the SEIR and SIR models bound the inferred $\tilde{N}_e(t)$.