Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data

Brian Lee[1], Muhammad Saqib Sohail[2], Elizabeth Finney[1], Syed Faraz Ahmed[2], Ahmed Abdul Quadeer[2], Matthew R. McKay[2,3,4,5], and John P. Barton[1,†]

[1]Department of Physics and Astronomy, University of California, Riverside, USA. [2]Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China. [3]Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China. [4]Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Victoria, Australia. [5]Department of Microbiology and Immunology, University of Melbourne, at The Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia. †Address correspondence to: john.barton@ucr.edu

New and more transmissible variants of SARS-CoV-2 have arisen multiple times over the course of the pandemic. Rapidly identifying mutations that affect transmission could facilitate outbreak control efforts and highlight new variants that warrant further study. Here we develop an analytical epidemiological model that infers the transmission effects of mutations from genomic surveillance data. Applying our model to SARS-CoV-2 data across many regions, we find multiple mutations that strongly affect the transmission rate, both within and outside the Spike protein. We also quantify the effects of travel and competition between different lineages on the inferred transmission effects of mutations. Importantly, our model detects lineages with increased transmission as they arise. We infer significant transmission advantages for the Alpha and Delta variants within a week of their appearances in regional data, when their regional frequencies were only around 1%. Our model thus enables the rapid identification of variants and mutations that affect transmission from genomic surveillance data.

Viruses can acquire mutations that affect how efficiently they infect new hosts, for example by increasing viral load or escaping host immunity[1–4]. The ability to rapidly identify mutations that increase transmission could inform outbreak control efforts and identify potential immune escape variants[5–9]. However, estimating how individual mutations affect viral transmission is a challenging problem.

Current methods to estimate changes in viral transmission generally rely on phylogenetic analyses or fitting changes in variant frequencies to a simple growth model[5,10–12]. Phylogenetic analyses for viruses can be challenging due to a high degree of sequence similarity, which implies that the data can be explained equally well by a number of different trees[13]. Phylogenetic analyses also typically rely on extensive Markov chain Monte Carlo sampling that becomes intractable for very large data sets. Simple growth models can estimate the difference in transmissibility between one variant and others circulating in the same region. However, they typically do not systematically account for competition between multiple variants, and their estimates may be difficult to compare for variants that arose in other regions or with different genetic backgrounds. These approaches also do not consider travel of infected individuals, nor do they account for superspreading —where a small number of infected individuals cause the majority of secondary infections —which has been observed for viruses like SARS-CoV and SARS-CoV-2[14,15].

To overcome these challenges, we developed a method to infer the effects of single nucleotide variants (SNVs) on viral transmission from genomic surveillance data that accounts for competition between viral lineages, travel, superspreading, and outbreaks in different locations. For clarity, we refer to non-reference nucleotides (including deletions or insertions) as SNVs and viral lineages possessing common sets of SNVs as variants. Simulations show that our approach can reliably estimate transmission effects of SNVs even from limited data. We applied our method to more than 1.6 million SARS-CoV-2 sequences from 87 geographical regions to reveal the effects of mutations on viral transmission throughout the pandemic. While the vast majority of SARS-CoV-2 mutations have negligible effects, we readily observe increased transmission for sets of SNVs in Spike and other hotspots throughout the genome. We further quantified the influence of travel and competition between multiple variants, using Nextstrain lineage 20E (EU1) as an example case. We found that realistic rates of travel during the pandemic would only slightly affect estimated changes in viral transmission. However, competition between variants has significant effects that must be accounted for in order to accurately estimate changes in transmission.

Importantly, our approach is sensitive enough to identify variants with increased transmission before they reach high frequencies. We demonstrate our capacity for early detection by studying the rise of the Alpha and Delta variants using data from the UK. In both cases, we reliably infer increased transmission for these variants within a week of their emergence, when their frequency in the region was only around 1%. Collectively, these data show that our model can be applied for the surveillance of evolving pathogens to robustly identify variants with transmission advantages and to highlight key mutations that may be driving changes in transmission.

Results

Epidemiological Model

To quantify the effects of mutations on viral transmission, we developed a stochastic branching process model of infection based on the well-known susceptible-infected-recovered (SIR) model. Our model incorporates superspreading by drawing the number of secondary infections caused by an infected individual from a negative binomial distribution with

mean R, referred to as the effective reproduction number, and dispersion parameter k (refs.[14,15]). Multiple variants with different transmission rates are included by assigning a variant a an effective reproduction number $R_a = R(1+w_a)$. Under an additive model, the net increase or decrease in transmission for a variant is the sum of the individual transmission effects s_i for each SNV i that the variant contains. In analogy with population genetics, we refer to the w_a and s_i as selection coefficients. In addition to superspreading and multiple variants, our model also incorporates travel of infected individuals into or out of a localized outbreak (Methods).

We can then apply Bayesian inference to estimate the transmission effects of SNVs that best explain the observed evolutionary history of an outbreak. To simplify our analysis, we use a path integral technique from statistical physics, recently applied in the context of population genetics[16], to efficiently quantify the probability of the model parameters given the data (for details, see Supplementary Information). This allows us to derive an analytical estimate for the maximum *a posteriori* selection coefficients $\hat{\boldsymbol{s}}$ for a given set of viral genomic surveillance data,

$$\hat{\boldsymbol{s}} = \left[\gamma I + C_{\text{int}}\right]^{-1} \left[\Delta \boldsymbol{x} + \boldsymbol{\tau}_{\text{int}}\right] . \quad \textbf{(1)}$$

Here $\Delta \boldsymbol{x}$ is the change in the SNV frequency vector over time, γ specifies the width of a Gaussian prior probability distribution for the selection coefficients s_i, and I is the identity matrix. C_{int} is the covariance matrix of SNV frequencies integrated over time, and accounts for competition between variants as well as the speed of growth for different viral lineages (Supplementary Information). $\boldsymbol{\tau}_{\text{int}}$ accounts for travel and is given explicitly in equation (S2). Data from multiple outbreaks can be combined by summing contributions to the integrated covariance, frequency change, and travel terms from each individual trajectory.

Validation in simulations

To test our ability to reliably infer selection, we analyzed simulation data using a wide range of parameters. We found that inference is accurate even without abundant data, especially when we combine information from multiple outbreaks (**Fig. 1**, **Supplementary Fig. 1**). Because we model the evolution of relative frequencies of different variants, accurate inference of selection does not require the knowledge of difficult-to-estimate parameters such as the current number of infected individuals or the effective reproduction number (Methods). Selection can be accurately inferred not only when the population evolves according to the superspreading model, but also if it evolves according to a classical multi-variant SIR model (**Supplementary Figs. 2-3**). This indicates that selection can be accurately estimated for a broad range of epidemiological dynamics.

Global patterns of selection in SARS-CoV-2

We studied the evolutionary history of SARS-CoV-2 using genomic data from GISAID[17] as of August 6th, 2021. We separated data by region and estimated selection coefficients jointly over all regions (Methods). After filtering regions

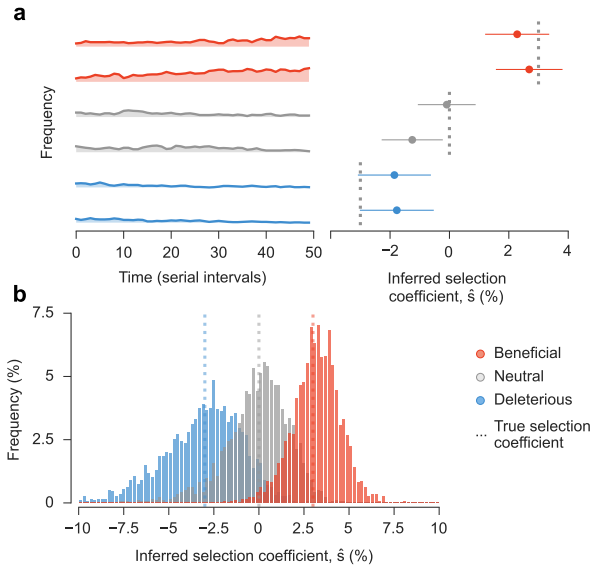


Fig. 1. Our approach accurately estimates transmission effects of mutations in simulations. Simulated epidemiological dynamics beginning with a mixed population containing variants with beneficial, neutral, and deleterious mutations. **a**, Selection coefficients for individual SNVs, shown as mean values $\pm$ one theoretical s.d., can be accurately inferred from stochastic dynamics in a typical simulation (Methods). **b**, Extensive tests on 1,000 replicate simulations with identical parameters show that inferred selection coefficients are centered around their true values. Deleterious coefficients are slightly more challenging to accurately infer due to their low frequencies in data. *Simulation parameters.* The initial population is a mixture of two variants with beneficial SNVs ($s = 0.03$), two with neutral SNVs ($s = 0$), and two with deleterious SNVs ($s = -0.03$). The number of newly infected individuals per serial interval rises rapidly from 6,000 to around 10,000 and stays nearly constant thereafter. Dispersion parameter k is fixed at 0.1.

with low or infrequent coverage, our analysis included more than 1.6 million SARS-CoV-2 sequences from 87 different regions, containing 13,189 nonsynonymous SNVs observed at nontrivial frequencies.

Our analysis revealed that, while the great majority of SNVs were nearly neutral, a few dramatically increased viral transmission (**Fig. 2a**, **Table 1**). We observe clusters of SNVs with strong effects on transmission along the SARS-CoV-2 genome (**Fig. 2b**). The highest density of SNVs that increase transmission is in Spike, especially in the S1 subunit. Of the top 20 mutations that we infer to be most strongly selected, 10 are in Spike (**Table 1**). However, clusters of SNVs with a strong selective advantage are also found in other proteins, especially in N, M, ORF3a, and NSP6.

Mutations inferred to substantially increase transmission

The top 50 mutations inferred to increase SARS-CoV-2 transmission the most are given in **Table 1**. Experimental evidence exists to directly or indirectly support many of these inferences. For clarity, we will reference mutations at the amino acid level rather than the underlying SNVs, which are also given in **Table 1**. Spike mutations L452R and P681R/H comprise three of the top four mutations, and all have demonstrated functional effects that could increase transmission[4,18–21]. Similarly, Spike mutations such as E484K ($\hat{s} = 5.2\%$, ranked 9th) and S477N ($\hat{s} = 4.4\%$, ranked 15th) appear prominently in our analysis. E484K has been shown to increase resistance to antibodies[22], and both E484K and

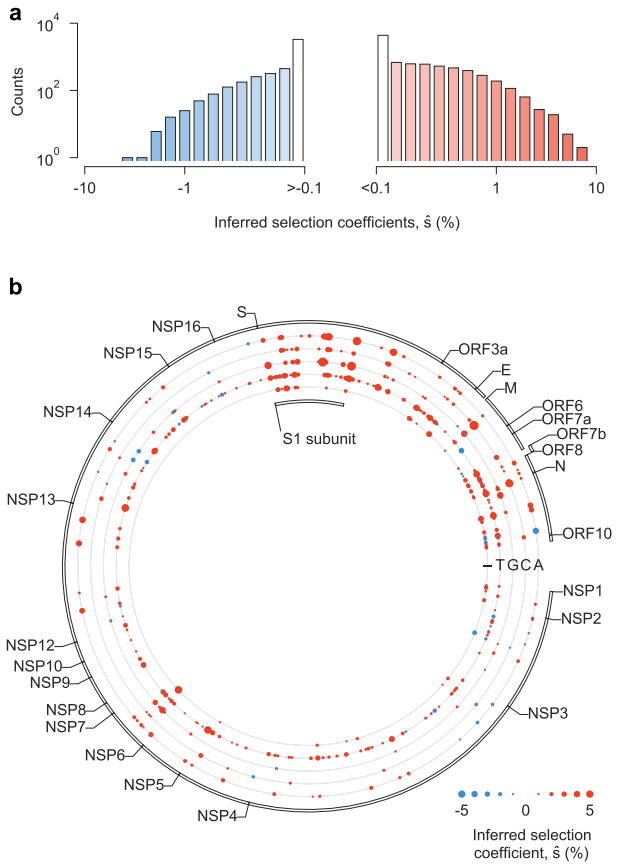


Fig. 2. Inferred transmission effects of SARS-CoV-2 mutations. **a**, The vast majority of the 13,189 nonsynonymous SNVs included in our study are inferred to have negligible effects on transmission (that is, $\hat{s}$ close to zero). However, a few SNVs have strong effects, as evidenced by a large value of $\hat{s}$. **b**, Patterns of selection across the SARS-CoV-2 genome. Beneficial SNVs often cluster together in the genome. Clustering is especially apparent for the S1 subunit of Spike, where many SNVs that are inferred to have the largest effects on transmission are located.

S477N were rapidly selected for increased ACE2 receptor binding during *in vitro* evolution[23]. Four Spike N-terminal domain (NTD) mutations/deletions (L18F, Δ141-142, and D253G) are also strongly selected. These lie in the antigenic supersite where mutations have been shown to decrease the neutralization potency of NTD-specific monoclonal antibodies[24]. Spike mutations D614G ($\hat{s} = 2.9\%$, ranked 52nd) and N501Y ($\hat{s} = 2.9\%$, ranked 58th) fall just outside the top 50 mutations in **Table 1**. D614G has been shown to increase binding affinity to the ACE2 receptor, thus increasing viral load and likely contributing to increased transmission[7,25]. Similarly, there is evidence that N501Y increases ACE2 binding affinity as well as transmission of infection[26].

Research on viral transmission has naturally focused on Spike because of its role in viral entry and as a target of neutralizing antibodies. However, our analysis also reveals strongly selected mutations outside of Spike. These include the Nucleocapsid mutations D3L and R203M. D3L ($\hat{s} = 6\%$, ranked 5th) has been reported to increase production of a non-canonical subgenomic RNA that encodes for ORF9b (ref.[27]), an interferon suppressing gene that can aid innate immune evasion and thereby increase transmission[28]. R203M ($\hat{s} = 4.6\%$, ranked 13th), present in the linker region of Nucleocapsid, has been shown to enhance viral RNA replication, delivery, and packaging, which may increase transmission[29]. The Nucleocapsid mutation T205I ($\hat{s} = 3.3\%$, ranked 40th) improves RNA delivery and expression to a lesser degree[29]. A Nucleocapsid mutation S202N ($\hat{s} = 3.3$, ranked 39th) is also inferred to be strongly selected. While the effect of this specific mutation is unknown, a mutation to Arginine at the same residue (S202R, which we infer to be more moderately beneficial, $\hat{s} = 1.7\%$) has been reported to increase replication, RNA delivery, and packaging[29]. Other strongly selected mutations outside of Spike include the Membrane I82T mutation ($\hat{s} = 7.5\%$, ranked 2nd) and NSP6 deletions Δ106-108 ($\hat{s} = 5.4\%$, 4.2%, and 3.1%, ranked 7th, 23rd, and 43rd). Similar examples may provide good targets for future studies of the functional effects of non-Spike mutations.

Estimates of selection for major SARS-CoV-2 variants

We estimated the net increase in viral transmission relative to the Wuhan-Hu-1 reference sequence for well-known SARS-CoV-2 variants by adding contributions from the individual variant-defining SNVs (**Fig. 3**, see Methods). Because our model uses global data and infers the transmission effects of individual SNVs, variants can be compared to one another directly even if they arose on different genetic backgrounds, or if they appeared in different regions or at different times. This also allows us to infer substantially increased transmission for variants such as Gamma, Beta, Lambda, and Epsilon, which never achieved the level of global dominance exhibited by Alpha and Delta (**Fig. 3**). For reference, we show the selection coefficient inferred for the cluster of mutations including the Spike mutation D614G that fixed early in the pandemic.

Our findings are consistent with past estimates that have shown a substantial transmission advantage for Alpha and Delta relative to other lineages[30–32]. However, past estimates have varied substantially depending on the data source and method of inference. For example, in different analyses Delta has been inferred to have an advantage of between 34% and 97% relative to other lineages[30–32]. Similarly, Alpha has been estimated to increase transmission by 29% to 90% relative to pre-existing lineages in different regions[5,11,12,31,33]. One advantage of our approach is that it can infer selection coefficients that best explain the growth or decline of variants across many regions, allowing selection for different variants to be compared on even footing.

In November 2021, a new variant, Omicron, was detected in South Africa. The data that we consider only extends to August 6th, 2021, and thus Omicron is not present in our data set. However, because this variant bears SNVs observed in other lineages, we can provide a preliminary estimate of its transmission advantage. Even without considering the effects of 17 Omicron SNVs (out of 96 total) that were not previously observed in data, the total selection coefficient for Omicron is $\hat{w} = 55.2\%$, which surpasses Alpha. While more data will be necessary to fully assess the transmission advantage of this variant, our model suggests that Omicron is highly transmissible, supporting its designation as a variant

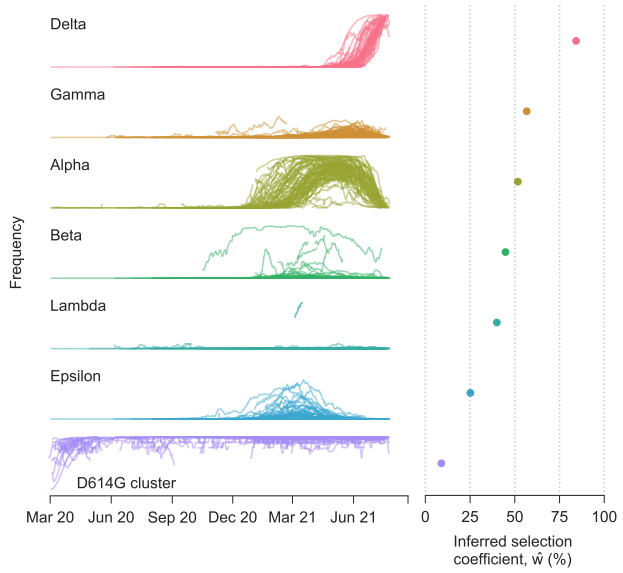


Fig. 3. Multiple SARS-CoV-2 variants strongly increase transmission rate. Frequencies of major variants and their total inferred selection coefficients, defined relative to the Wuhan-Hu-1 reference sequence. Selection coefficients for variants with multiple SNVs are obtained by summing the effects of all variant-defining SNVs (Methods). Because our method uses global data and accounts for competition between variants, we infer large transmission advantages even for variants such as Gamma, Beta, Lambda, and Epsilon, which never achieved the same level of global dominance as variants such as Alpha and Delta.

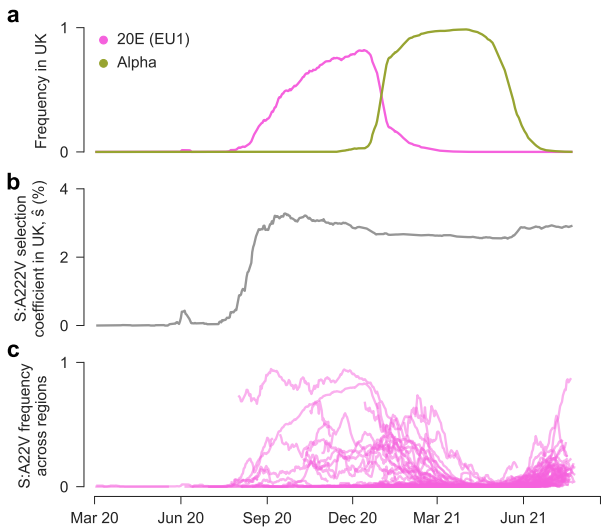


Fig. 4. Inferred transmission benefit for the Spike mutation A222V. a, The 20E (EU1) lineage, bearing the Spike mutation A222V, is outcompeted by Alpha in the UK. **b**, Because our model accounts for competition between variants, we nonetheless infer a transmission benefit for A222V, as well as other 20E (EU1) mutations, that persists even as 20E (EU1) dies out in the UK. **c**, Globally, S:A222V has arisen on different sequence backgrounds and increased in frequency since June 2021, consistent with a transmission advantage for this mutation.

of concern (VOC) by WHO[34].

Effects of travel and competition on inferred coefficients

In addition to differences in transmissibility, variant frequencies are affected by the travel of infected individuals between regions and competition between variants. Multiple introductions of a new variant into a region can increase its local frequency even if the variant has no transmission advantage[35,36]. This could, for example, make a neutral variant appear beneficial. Conversely, variants that transmit more effectively than ancestral SARS-CoV-2 can still be outcompeted by other, more transmissible variants, causing them to decline. In population genetics, this is referred to as clonal interference[37,38].

We studied the history of Nextstrain lineage 20E (EU1) as an example to investigate the influence of travel and inter-lineage competition on inferred changes in transmission. A detailed analysis showed how 20E (EU1) spread from Spain to other regions in Europe[35]. There it was estimated that 20E (EU1) was introduced into the UK roughly 380 times during the summer of 2020. Assuming no travel between the UK and other regions, we infer a total selection coefficient for novel 20E (EU1) SNVs of 15.6% using UK data gathered from the beginning of the pandemic through May 1st, 2021, when 20E (EU1) had died out locally. Including 380 importations into the UK during the summer of 2020, our inferred selection coefficient is only slightly reduced to 15.4% (Methods). Around 37,000 importations would be necessary during this time for 20E (EU1) to be inferred to be completely neutral. Thus, while travel was crucial for the early spread of 20E (EU1) in the UK and across Europe, its subsequent growth ultimately dominates estimates of its effects on transmission.

Competition between Alpha and 20E (EU1), which was rising in the UK before the emergence of Alpha, provides a clear example of clonal interference in SARS-CoV-2. Interference can be readily observed in their frequency trajectories in the UK (**Fig. 4a**). Changes in frequency for the two variants have a Pearson correlation of -0.73.

To measure the influence of competition on estimates of transmission, we inferred selection coefficients for variants circulating in the UK with all nucleotide changes present in the Alpha variant reverted to the Wuhan-Hu-1 reference sequence. This provides an estimate for the 20E (EU1) selection coefficient that ignores competition with Alpha. Using all of the data, 20E (EU1) is inferred to have a significant transmission advantage relative to the previously dominant variant B.1 ($\hat{w}_{20E(EU1)} - \hat{w}_{B.1} = 10.2\%$). However, when competition with Alpha is ignored, 20E (EU1) is inferred to be mildly deleterious ($\hat{w}_{20E(EU1)} - \hat{w}_{B.1} = -1.7\%$). Thus, ignoring competition between variants would lead to a dramatic and likely incorrect change in inferred selection.

The re-emergence of the Spike mutation A222V, which appeared within the 20E (EU1) variant, further supports a transmission advantage. Using global data, we infer A222V to significantly increase transmission ($\hat{s} = 3.8\%$, ranked 29th, see **Table 1**). In our analysis, A222V contributes far more than any other mutation to the increased transmission of 20E (EU1) relative to contemporary variants. Focusing on data from the UK alone, the inferred benefit of A222V remains roughly constant even as 20E (EU1) is outcompeted by Alpha (**Fig. 4b**). Later, A222V arises on other sequence backgrounds and steadily increases in frequency (**Fig. 4c**), consistent with the increase in transmission inferred by our model.

Our analysis therefore suggests that 20E (EU1) possessed

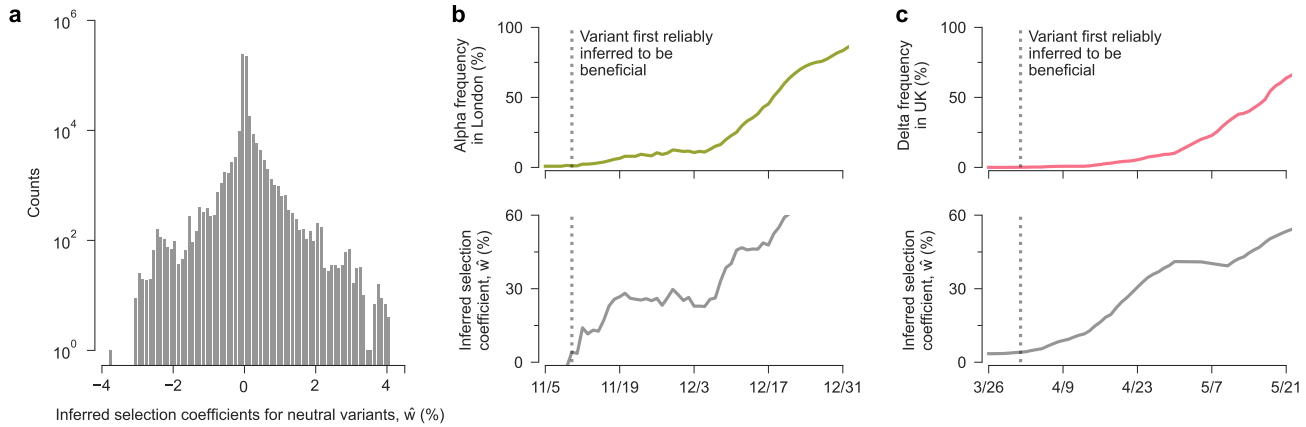


Fig. 5. Our model rapidly infers increased transmission for Alpha and Delta. **a**, The null distribution of inferred selection coefficients for neutral variants over all intermediate times and across all regions. Neutral variants are defined as (collections of) SNVs that are inferred to have total selection coefficients $|w| < 1\%$ using global data. **b**, Frequency of Alpha in London and its inferred selection coefficient over time. The inferred coefficient exceeds the largest one in the null distribution on November 9th, 2020. **c**, Frequency of Delta in the UK and the inferred selection coefficient for novel Delta SNVs over time. The inferred coefficient exceeds the null on March 31st, 2021.

a modest but real transmission advantage relative to contemporary SARS-CoV-2 sequences. This finding is consistent with analysis in ref.[35], where a model of 20E (EU1) spread due to travel alone underestimated the observed frequency of the variant by 1- to 12-fold in all regions. More generally, our analysis suggests that very large inferred selection coefficients for variants or SNVs in our model are unlikely to be explained by travel alone.

Rapid detection of variants with enhanced transmission

Rapidly identifying variants with increased transmission is important to inform public health efforts to limit viral spread. However, the inherent stochasticity of infection and of genomic surveillance data collection makes accurate inferences difficult. For example, neutral or modestly deleterious variants may initially appear to be beneficial due to a transient rise in frequency despite having no selective advantage.

To quantify how fluctuations affect estimates of selection for neutral variants, we first identified all variants (including both SNVs and collections of SNVs that are strongly linked to one another) that are inferred to have selection coefficients with magnitude less than 1% using all of the data. We then calculated the selection coefficients that would have been inferred for the SNVs or variants at all earlier time points and in all regions after they were first observed in the data. This "null" distribution (**Fig. 5a**) quantifies fluctuations in inferred selection coefficients for nearly-neutral variants due to stochasticity in viral spread and sampling. Variants with selection coefficients larger than any in the null distribution could then be expected with high confidence to have some transmission advantage.

To assess our ability to rapidly detect variants with a transmission advantage, we studied the rise of the Alpha variant in the London area. Using the above criterion, we identify Alpha as very likely to increase transmission using sequence data that was collected on or before November 9th, 2020 (**Fig. 5b**). This is roughly three weeks before Public Health England labeled Alpha as a variant of interest (VOI)[39], and more than a month before it was classified as a VOC[40]. At this time the frequency of Alpha in London was around 1%.

A similar analysis also shows that our model rapidly infers increased transmission for the Delta variant. Using data from the UK, we reliably infer Delta to increase transmission by March 31st, 2021 (**Fig. 5c**). Delta was classified as a VOI on April 4th, 2021 and as a VOC more than one month later on May 6th, 2021[41]. At the time that we detected increased transmission for Delta, its frequency was still low ($< 1\%$) in the UK. Collectively, these results demonstrate our ability to rapidly identify variants with higher transmission, even when they represent a small fraction of all infections in a region.

Discussion

Quantifying the effects of mutations on viral transmission is an important but challenging problem. To overcome limitations of current methods, we developed a flexible, SIR-based epidemiological model that provides analytical estimates for the transmission effects of SNVs from genomic surveillance data. Applying our model to SARS-CoV-2 data, we identified SNVs that substantially increase viral transmission, including both experimentally-validated Spike mutations and other, less-studied mutations that may be promising targets for future investigation. We further explored the effects of travel and competition between variants on inferred changes in transmission, using the history of 20E (EU1) as an example. Importantly, we found that our model is sensitive enough to detect substantial transmission advantages for variants such as Alpha and Delta even when they comprised only a small fraction of the total number of infections in a region, thus providing an "early warning" for more transmissible variants.

Further monitoring will be important to identify and characterize new variants as they arise. The Omicron variant that was recently detected in South Africa provides one such example. While the data in our study only extends to August 6th, 2021, we would estimate a selection coefficient of 55.2% for Omicron based on the mutations that it shares with previous variants alone.

While our study has focused on SARS-CoV-2, the epidemiological model that we have developed is very general. The same methodology could be applied to study the transmission of other pathogens such as influenza. Combined with thorough genomic surveillance data, our model provides a powerful method for rapidly identifying more transmissible viral lineages and quantifying the contributions of individual mutations to changes in transmission.

ACKNOWLEDGEMENTS

We gratefully acknowledge the numerous laboratories worldwide that have provided sequence data and metadata to GISAID. A full list of originating and submitting laboratories for the sequences used in our analysis can be found at https://www.gisaid.org using the EPI-SET-ID: EPI_SET_20211201ze. The work of B.L., E.F., and J.P.B. reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM138233. The work of S.F.A., M.S.S., A.A.Q., and M.R.M. was supported by the General Research Fund of the Hong Kong Research Grants Council under Grant Number 16213121.

AUTHOR CONTRIBUTIONS

All authors contributed to methods development, data analysis, interpretation of results, and writing the paper. B.L. and J.P.B. led theoretical analyses. M.S.S. led SIR model simulations. B.L. led branching process simulations. J.P.B. conceptualized and supervised the project.

References

1. Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. *Nature Reviews Microbiology* **16**, 47–60 (2018).
2. Revill, P. A. *et al.* The evolution and clinical impact of hepatitis B virus genome diversity. *Nature Reviews Gastroenterology and Hepatology* **17**, 618–634 (2020).
3. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310 (2020).
4. Li, Q. *et al.* The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* **182**, 1284–1294 (2020).
5. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
6. Wibmer, C. K. *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nature Medicine* **27**, 622–625 (2021).
7. Korber, B. *et al.* Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827 (2020).
8. Diehl, W. E. *et al.* Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. *Cell* **167**, 1088–1098.e6 (2016).
9. Imai, M. *et al.* Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* **486**, 420–428 (2012).
10. Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics* **10**, 540–550 (2009).
11. Emergence and rapid transmission of SARS-CoV-2 B. 1.1. 7 in the United States .
12. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1.7 in England. *Science* **372** (2021).
13. Morel, B. *et al.* Phylogenetic analysis of SARS-CoV-2 data is difficult. *Molecular biology and evolution* **38**, 1777–1791 (2021).
14. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
15. Althouse, B. M. *et al.* Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLOS Biology* **18**, 1–13 (2020).
16. Sohail, M. S., Louie, R. H. Y., McKay, M. R. & Barton, J. P. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature Biotechnology* **39**, 472–479 (2021).
17. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global challenges* **1**, 33–46 (2017).
18. Deng, X. *et al.* Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell* **184**, 3426–3437.e8 (2021).
19. Saito, A. *et al.* Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature* 1–10 (2021).
20. Mohammad, A., Abubaker, J. & Al-Mulla, F. Structural modelling of SARS-CoV-2 Alpha variant (B.1.1.7) suggests enhanced furin binding and infectivity. *Virus Research* **303**, 198522 (2021).
21. Lista, M. J. *et al.* The P681H mutation in the spike glycoprotein confers type I interferon resistance in the SARS-CoV-2 alpha (B.1.1.7) variant. *bioRxiv* 2021.11.09.467693.
22. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host & Microbe* **29**, 463–476 (2021).
23. Zahradník, J. *et al.* SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nature microbiology* **6**, 1188–1198 (2021).
24. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347.e16 (2021).
25. Yurkovetskiy, L. *et al.* Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**, 739–751 (2020).
26. Liu, Y. *et al.* The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature* (2021).
27. Parker, M. D. *et al.* Altered subgenomic RNA expression in SARS-CoV-2 B.1.1.7 infections. *bioRxiv* 2021.03.02.433156 (2021).
28. Thorne, L. G. *et al.* Evolution of enhanced innate immune evasion by SARS-CoV-2. *Nature* (2021).
29. Syed, A. M. *et al.* Rapid assessment of SARS-CoV-2 evolved variants using virus-like particles. *Science* **374**, 1626–1632 (2021).
30. Allen, H. *et al.* Household transmission of COVID-19 cases associated with SARS-CoV-2 delta variant (B.1.617.2): National case-control study. *The Lancet Regional Health - Europe* **12**, 100252 (2021).
31. Campbell, F. *et al.* Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Eurosurveillance* **26**, 2100509 (2021).
32. Alizon, S. *et al.* Rapid spread of the SARS-CoV-2 Delta variant in some French regions, June 2021. *Eurosurveillance* **26**, 2100573 (2021).
33. Chen, C. *et al.* Quantification of the spread of SARS-CoV-2 variant B.1.1.7 in Switzerland. *Epidemics* **37**, 100480 (2021).
34. World Health Organization. Classification of Omicron (b. 1.1. 529): SARS-CoV-2 variant of concern. (2021). URL https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern.
35. Hodcroft, E. B. *et al.* Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
36. Grubaugh, N. D., Hanage, W. P. & Rasmussen, A. L. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. *Cell* **182**, 794–795 (2020).
37. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102**, 127–144 (1998).
38. Strelkowa, N. & Lässig, M. Clonal interference in the evolution of influenza. *Genetics* **192**, 671–682 (2012).
39. Chand, M. *et al.* Investigation of novel SARS-CoV-2 variant, variant of concern 202012/01, technical briefing. Tech. Rep. (2020). URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959438/Technical_Briefing_VOC_SH_NJL2_SH2.pdf.
40. Chand, P. M. *et al.* Investigation of novel SARS-CoV-2 variant, variant of concern 202012/01, technical briefing 2. Tech. Rep. (2020). URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959361/Technical_Briefing_VOC202012-2_Briefing_2.pdf.
41. PHE Genomics Cell, P. E. C. P. C. T. D. T., PHE Outbreak Surveillance Team. SARS-CoV-2 variants of concern and variants under investigation in england. Tech. Rep. (2021). URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984274/Variants_of_Concern_VOC_Technical_Briefing_10_England.pdf.

Methods

Epidemiological model

We use a discrete time branching process to model the spread of infection. Individuals can be infected by any one of M viral variants, which are represented by genetic sequences $\boldsymbol{g} = \{g_1, g_2, \ldots, g_L\}$ of length L. For simplicity, we will first assume that alleles at each site i in the genetic sequence for variant a are either equal to the "wild-type" or reference ($g_i^a = 0$) or mutants ($g_i^a = 1$). Later we will relax this assumption to consider genetic sequences with 5 possible states at each site (4 nucleotides or a gap). We call $n_a(t)$ the number of individuals infected by variant a at time t. To account for super-spreading, the number of newly infected individuals at time $t+1$ follows a negative binomial distribution[42], $P(n_a(t+1)|n_a(t), k, R_a) = P_{NB}(r,p)$, where $r = n_a k$, $p = k/(k+R_a)$, and $R_a = R(1+w_a)$. Here r and p are the negative binomial distribution parameters, k is the dispersion, R is the average effective reproductive number, and w_a encodes the variant dependence of the infectivity. Here R is also an implicit function of time, representing change in the number of susceptible and recovered individuals as well as the effects of public health interventions or changes in behavior that modify viral transmission.

To incorporate travel, n_a is the sum of the locally infected population and the net flux of infected individuals into the region, $n_a = n_{a,\text{local}} + n_{a,\text{inflow}} - n_{a,\text{outflow}} \equiv n_{a,\text{local}} + \delta n_a$. Defining the frequency of variant a as $y_a = n_a / \sum_b n_b$, the probability that the frequency vector is $\boldsymbol{y}(t+1) = \{y_1(t+1), y_2(t+1), \ldots\}$ given the initial frequency vector $\boldsymbol{y}(0)$, is

$$P((\boldsymbol{y}(t))_{t=1}^{T} | \boldsymbol{y}(0)) = \prod_{t=0}^{T-1} P(\boldsymbol{y}(t+1)|\boldsymbol{y}(t)) . \tag{2}$$

Derivation of the estimator

Because (2) is difficult to work with directly, we introduce a "diffusion approximation" where we assume that the total number of infected individuals is large and the effects of mutations on transmission are small. Similar approximations have been widely used in population genetics[43–45]. Under these assumptions, the probability distribution for the variant frequencies satisfies a Fokker-Planck equation with terms derived from the first and second moments of the frequency changes $y_a(t+1) - y_a(t)$ under the negative binomial distributions above.

However, the genotype space is high-dimensional (dimension 2^L, with either a mutant or wild-type allele at each site) and undersampled, making inference of selection for genotypes extremely challenging. To simplify the inference problem, we assume that selection is additive, so the total selection coefficient w_a for a variant a is the sum of selection coefficients s_i for mutant alleles at each site i:

$$w_a = \sum_{i=1}^{L} g_i^a s_i .$$

We can then derive a Fokker-Planck expression for the dynamics of mutant allele frequencies

$$x_i = \sum_{a=1}^{M} g_i^a y_a .$$

At the allele level, the Fokker-Planck equation has a drift vector given by

$$d_i(\boldsymbol{x}) = x_i(1-x_i)s_i + \sum_{j=1}^{L}(x_{ij} - x_i x_j)s_j + \frac{1}{R}\left[\delta x_i - x_i \sum_{b=1}^{M} \frac{\delta n_b}{n}\right] , \tag{3}$$

and a diffusion matrix

$$C_{ij} = \left(\frac{1}{k} + \frac{1}{R}\right) \times \begin{cases} x_{ij} - x_i x_j & i \neq j \\ x_i(1-x_i) & i = j \end{cases} ,$$

where x_{ij} is the frequency of infected individuals that have mutant alleles at both site i and site j at time t, and δx_i is the change in the frequency due to individuals traveling. Here $n = \sum_{a=1}^{M} n_a$ is the total number of individuals infected by all variants. If $l_i = l_{i,\text{inflow}} - l_{i,\text{outflow}}$ is the number of people traveling into a region minus that traveling out of it who are infected with a variant that has a mutant allele at site i, then $\delta x_i = l_i/n$.

The Fokker-Planck equation can then be used to derive a path integral, which expresses the probability of an entire evolutionary history or "path" (i.e., frequencies of genetic variants over time, $(\boldsymbol{x}(t_k))_{k=1}^{T-1}$). The path integral is

$$P\left((\boldsymbol{x}(t_k))_{k=1}^{T-1} | \boldsymbol{x}(t_0), \boldsymbol{s}, n, \delta\boldsymbol{n}\right) \approx \left(\prod_{k=0}^{T-1} \frac{1}{\sqrt{\det C}} \left(\frac{n}{2\pi}\right)^{L/2} \prod_{i=1}^{L} dx_i(t_{k+1})\right) \exp\left(-\frac{n}{2} S((\boldsymbol{x}(t_k)_{k=0}^{T})\right) \tag{4}$$

$$S((\boldsymbol{x}(t_k)_{k=0}^{T}) = \sum_{k=1}^{T-1} \left[\frac{\boldsymbol{x}(t_{k+1}) - \boldsymbol{x}(t_k)}{\Delta t_k} - \boldsymbol{d}(\boldsymbol{x}(t_k))\right] C^{-1}(\boldsymbol{x}(t_k)) \left[\frac{\boldsymbol{x}(t_{k+1}) - \boldsymbol{x}(t_k)}{\Delta t_k} - \boldsymbol{d}(\boldsymbol{x}(t_k))\right] .$$

The path integral quantifies the probability density for paths of mutant allele frequencies in the evolutionary history of the pathogen. We can then use Bayesian inference to find the maximum *a posteriori* estimate for the selection coefficients given the frequencies, the infected population size, the parameters R and k, and the composition of the population traveling into or out of the regional outbreak. The posterior probability of the selection coefficients is

$$P\left(\boldsymbol{s}|(\boldsymbol{x}(t_k))_{t=0}^{T}\right) \propto P\left((\boldsymbol{x}(t_k))_{t=1}^{T}|\boldsymbol{x}(t_0)\right) P_{\text{Prior}}(\boldsymbol{s}), \quad \textbf{(5)}$$

where $P\left((\boldsymbol{x}(t_k))_{t=1}^{T}|\boldsymbol{x}(t_0)\right)$ is the probability of a path given by (4) and the $P_{\text{Prior}}(\boldsymbol{s})$ is a Gaussian prior probability for the selection coefficients with zero mean and covariance matrix $\sigma^2 I$. Here, I is the identity matrix and σ^2 is the variance of the prior. The selection coefficients that maximize (5) are

$$\hat{\boldsymbol{s}} = \left[\gamma I + \sum_{t=0}^{T} \frac{nk^2R^2}{(k+R)^2} C(t)\right]^{-1} \left[\sum_{t=0}^{T-1} \frac{nkR}{k+R}\left(\boldsymbol{x}(t+1) - \boldsymbol{x}(t) - \frac{1}{R}\left(\delta\boldsymbol{x}(t) - \boldsymbol{x}(t)\sum_{b=1}^{M}\frac{\delta n_b(t)}{n}\right)\right)\right] \quad \textbf{(6)}$$

where the parameters k, R, and n are implicitly functions of t.

There are two interesting limiting forms of the estimator. First, we define the new matrix $\bar{C}$ whose entries are

$$\bar{C}_{ij} = \begin{cases} x_{ij}(t) - x_i(t)x_j(t) & i \neq j \\ x_i(t)(1 - x_i(t)) & i = j \end{cases}. \quad \textbf{(7)}$$

In the limit that $k \to \infty$, the negative binomial distribution for new infections becomes a Poisson distribution with rate $\lambda = R$. In this special case, the model is equivalent to the Wright-Fisher model from population genetics. The estimator reduces to

$$\hat{\boldsymbol{s}} = \left[\gamma I + \sum_{t=0}^{T} nR\,\bar{C}\right]^{-1} \left[\sum_{t=0}^{T-1} nR\left(\boldsymbol{x}(t+1) - \boldsymbol{x}(t)\right)\right], \quad \textbf{(8)}$$

where we have dropped the migration term for simplicity.

The opposite limit $k \to 0$ corresponds to a distribution for new infections with extremely heavy tails, i.e., one where super-spreading is dominant. In this case the drift in (3), which quantifies expected frequency changes due to selection and travel, is unchanged. However, the diffusion matrix, which encodes linkage as well as the changes in frequency that are due to the stochastic nature of infection transmission, diverges. In this case, diffusion dominates the process entirely.

Simplifying the estimator and robustness to incomplete knowledge of time-varying parameters

In practice, parameters appearing in (6), such as the infected population size n, the dispersion k, and the mean reproductive number R, are likely time-varying. While such time dependencies are accommodated by our model, they can be challenging to reliably estimate from data. However, we generally do not require full knowledge of these time-dependent parameters to accurately estimate selection.

In fact, due to finite sampling noise, estimates of selection produced by assuming constant (and incorrect) parameters are more accurate than estimates that use the true time-varying parameters (**Supplementary Fig. 4**). The naive estimator in (6) implies that time points or regions with larger R, n, or k should be weighted more heavily in the estimate. However, frequency information is always inaccurate due to noise from finite sampling, so weighing some time points or regions significantly more than others based upon the parameters alone means that undue weight is given to the uncertain information available from these times and regions.

For this reason, we assume parameters that are spatially and temporally constant in all of the following analysis, except when considering travel, as discussed below. This allows the estimator to be simplified substantially. If we assume constant parameters and scale the regularization γ by the prefactor in the numerator in (6), the parameter dependence in the numerator and the denominator is almost the same and largely cancels out. With the same definition of the matrix $\bar{C}$ as above, and additionally defining $\bar{C}_{\text{int}} = \sum_{t=0}^{T} \bar{C}$, the simplified estimator is given by

$$\hat{\boldsymbol{s}} = \left[\gamma I + \bar{C}_{\text{int}}\right]^{-1} \left[\boldsymbol{x}(T) - \boldsymbol{x}(0) + \tau_{\text{int}}\right], \quad \textbf{(9)}$$

$$\text{where } \tau_{\text{int}} = -\sum_{t=0}^{T} \frac{1}{R}\left(\delta\boldsymbol{x}(t) - \boldsymbol{x}(t)\sum_{b=1}^{M}\frac{\delta n_b}{n}\right).$$

This form of the estimator has significant advantages over (6). The most important is that, if the travel term is dropped, the difficult-to-estimate parameters R, k, and n are no longer required. For methods of inferring these parameters as well as discussions about the difficulty of inferring them, see refs.[46–55].

Extension to multiple regions and multiple SNVs at each site

The model can easily account for outbreaks in multiple regions or outbreaks at different times. If the probability of the evolutionary path in each region is independent, which is the case if there is no travel between regions, then the probability of all of the evolutionary paths in all of the regions is simply the product of the probabilities of the paths in each region, given by (4). Bayesian inference can be applied in the same

way as before, resulting in the estimator

$$\hat{\boldsymbol{s}} = \left[\gamma I + \sum_{r=1}^{Q} \bar{C}_{r,\text{int}}\right]^{-1} \left[\sum_{r=1}^{Q} \boldsymbol{x}_r(T_r) - \boldsymbol{x}_r(t_{r,0})\right], \quad \textbf{(10)}$$

where Q is the number of regions, t_r is the time in region r, T_r is the final time in region r, $t_{r,0}$ is the initial time in region r, $\boldsymbol{x}_r$ is the frequency in region r, and $\bar{C}_{r,\text{int}}$ is the scaled integrated covariance matrix in region r given by integrating (7) over time. The estimator can further be extended to allow for multiple different nucleotides at each site by simply letting each different nucleotide have its own entry in the frequency vector x_i. If there are J mutations at each site this results in a frequency vector of length LJ, and a covariance matrix of size $LJ \times LJ$. By convention, reference sequence alleles have selection coefficients of zero, so the mutant allele selection coefficients at each site are normalized by subtracting the inferred coefficient for the reference allele.

Branching process simulations

We implemented the superspreading branching process for the number of infected individuals in Python. We used a negative binomial distribution for the number of secondary infections caused by a group of individuals infected with the same pathogen variant. To test how finite sampling affects model estimates, we sampled n_s genomes per time point to use for analysis. We computed the single and double mutant frequencies, x_i and x_{ij}, respectively, from the sampled sequences and estimated the selection coefficients from these using (1), possibly extended to account for multiple outbreaks or multiple alleles at each locus as described above.

Susceptible-infected-recovered simulations

We simulated a multi-variant Susceptible-Infected-Recovered (SIR) model with M variants of a pathogen circulating in a population of N individuals, assuming total cross-immunity between variants. Mathematically, the SIR dynamics are expressed as

$$\begin{aligned}
\frac{dS}{dt} &= -\sum_{b=1}^{M} \beta_b \frac{SI_b}{N} \\
\frac{dI_a}{dt} &= \beta_a \frac{SI_a}{N} - r_a I_a \\
\frac{dR}{dt} &= \sum_{b=1}^{M} r_b I_b \,,
\end{aligned}$$

for $a = 1, \ldots, M$, where I_a is the number of individuals infected by variant a, β_a and r_a are the transmission and recovery rate associated with the ath variant, and S and R are the total number of susceptible and recovered individuals, respectively. Each variant is represented by a binary sequence of length L, with 0 representing the wild-type (WT) allele and 1 representing the mutant allele. Considering the all-zero sequence as the reference with transmission rate β_{ref}, the transmission rate of variant a can be expressed as $\beta_a = \beta_{\text{ref}} \left(1 + \sum_{i=1}^{L} g_i^a \, s_i\right)$, with g_i^a and s_i defined as above.

We further incorporate the effect of public health interventions and changing human behaviour on transmission by making the transmission rate a function of time, i.e., $\beta_{\text{ref}}(t)$. As the number of susceptible individuals decreases, the effective transmission rate will decrease. The effective reproductive number of the ath variant at time t is

$$R_{t,a} = \frac{\beta_a(t)}{r} \frac{S(t)}{N}. \quad \textbf{(11)}$$

We used MATLAB to simulate the SIR model under a scenario where the number of newly-infected individuals continues to increase and then remains fixed (**Supplementary Fig. 2**), and a scenario where we fix $r_a = 1$ and adapt the transmission rate over time such that the system follows the typical SIR dynamics (**Supplementary Fig. 3**). In the SIR model there is no superspreading, which corresponds to the limit $k \to \infty$ in the branching process model described above. The estimator for the selection coefficients then reduces to a scaled version of (8),

$$\hat{\boldsymbol{s}} = \left[\gamma I + \sum_t nR\bar{C}\right]^{-1} \left[\sum_t \frac{nR}{\beta_{\text{ref}}} \left(\boldsymbol{x}(t+1) - \boldsymbol{x}(t)\right)\right],$$

where $n = n(t)$ is the number of newly-infected individuals at time t (and is different therefore from I), $\boldsymbol{x}$ is the frequency vector of newly infected individuals, and γ is the regularization. In both cases selection coefficients are accurately recovered.

Regions and time-series for SARS-CoV-2 analysis

We used sequence alignments and metadata downloaded from GISAID (ref.[56]) on August 14th, 2021, which includes more than 3 million sequences. Ideally, we would like to divide this data into the smallest separate areas that have outbreaks that are largely independent of those in the surrounding regions, so as to avoid biases due to travel between regions or unequal sampling in different locations. However, this needs to be balanced with the limitations of the data, since regions with poor sampling could contribute more noise than signal. We therefore divided data into the smallest regions available in the metadata that are still large enough such that infections resulting from travel outside of the region are likely to be far less frequent than transmission within the region. This results in the inclusion of mostly separate countries in Europe and Asia and states in North America. Two exceptions to this are that we separate northern and southern California due to the geographical separation of population centers, and we separate Northern Ireland from the rest of the United Kingdom due to its geographical isolation.

To minimize the effects of sampling noise, we chose regions and time-series within these regions based on the following criteria:

1. In any period of 5 days within the time-series there are at least 20 total samples.

2. The number of days in the time-series is greater than 20.

3. The number of new infections per day is at least around 100.

The last criterion ensures that there are enough infected individuals that transmission is not driven overwhelmingly by stochasticity.

Our results are robust to reasonable variation in these parameters. Comparing the number of locations used and the sample sizes shown in **Supplementary Fig. 5** in the data to those used in the simulations shown in **Supplementary Fig. 1**, we expect our inference to accurately distinguish beneficial, deleterious, and neutral SNVs from one another.

Data processing

We perform a number of preprocessing steps to ensure data quality. We first eliminated incomplete sequences with gaps at more than one third of the genome. We then removed sites from our analysis where gaps are observed at $> 95\%$ frequency, since these sites may represent very rare insertions or sequencing errors. We also removed sites in noncoding regions of the SARS-CoV-2 genome and ones where all observed SNVs are synonymous. We imputed ambiguous nucleotides with the nucleotide at the same site that occurs most frequently in other sequences from the same region.

For the remaining sites, we excluded rare SNVs whose frequency is never larger than 1% in any region and ones that are not observed at least 5 times. These sites, if included, are almost always inferred to have extremely small selection coefficients. Furthermore, since their frequencies are so small, their covariance with other sites is also small and is therefore unlikely to have a large effect on inference. We verified that different reasonable values for these cutoffs result in essentially identical selection coefficients (**Supplementary Fig. 6**).

Calculating frequency changes and covariances

To increase robustness to finite sampling, we integrated terms in (6) over time, assuming that frequencies and covariances are piecewise linear, rather than summing contributions from each time point[57]. To obtain better estimates of changes in SNV frequencies (the term $x(T) - x(0)$ in (9)), we averaged $\boldsymbol{x}(T)$ as the frequencies in the window of the final 10 days and $\boldsymbol{x}(0)$ as the frequencies in the window of the first 10 days for each time-series and region. This smoothing is necessary especially in regions where sampling is sparse, where the number of genomes sampled on a particular day may be as small as 1 or 2. We confirmed that our results are robust to reasonable changes of this window size of 10 days (**Supplementary Fig. 6**).

We also normalized time in units of serial intervals or "generations" by dividing the integrated covariance matrix by 5, following results that the serial interval for SARS-CoV-2 is roughly 5 days[58–60]. This allows us to convert from units of time in days to generations, as in (9).

Calculating selection coefficients

After the above preprocessing (before eliminating synonymous SNVs) there remain $21,050$ SNVs observed at a frequency above 1% in at least one region and observed at least 5 times. We assume constant values for R, n, and k in all regions, and use (10) to estimate selection. When R, n, and k are constant, these terms can be effectively absorbed into the regularization γ.

We normalize selection coefficients such that the nucleotide for the Wuhan-Hu-1 reference sequence at each site has a selection coefficient of 0. To do this, we subtract the selection coefficient for the reference nucleotide from the inferred coefficient for each other allele at that site after all selection coefficients have been computed.

We used these estimates for the selection coefficients for nonsynonymous SNVs to estimate the corresponding selection coefficients for amino acid substitutions (**Table 1**). If there were multiple SNVs in a codon that result in the same amino acid variant, but are not strongly linked to one another, then the selection coefficient for the amino acid was calculated as the largest (in absolute value) of the SNVs. If there were multiple SNVs in the same codon that yield the same amino acid and these SNVs are strongly linked to one another, then the selection coefficient for the mutant amino acid was calculated as the sum of the selection coefficients for the SNVs.

We calculated selection coefficients for major variants by summing the individual nucleotide SNVs that define the variant, which follows from our assumption of additive fitness. SNVs for major variants were obtained by first finding groups of strongly linked SNVs that correspond to a variant, and then adding any other mutations given on https://covariants.org that were not identified by our linkage analysis.

We also computed selection coefficients for collections of strongly linked SNVs that may not be officially-designated variants. To determine sets of strongly linked SNVs, we considered the following statistics. If the number of genomes with a SNV at site i is called h_i and the number of genomes with SNVs at both site i and site j is h_{ij}, then we say that two sites i and j are strongly linked if h_{ij}/h_i and h_{ij}/h_j are both greater than 80%. As for the major variants, we computed selection coefficients for sets of strongly linked SNVs by summing the contributions from individual SNVs. Selection coefficients for strongly linked SNVs were used to compute the "null" distribution that we use as a metric for early detection of variants with increased transmission.

Choice of regularization

In principle, the regularization strength γ is related to the width of the prior distribution for SNV selection coefficients. The regularization strength also plays a role in reducing noise in selection coefficient estimates due to finite sampling of viral sequences. This is especially important for SNVs that are observed only briefly in data, as they will have small integrated variances in the "denominator" of (6). Larger values of the regularization more strongly suppress noise, but they also shrink inferred selection coefficients towards zero.

We use a regularization strength of $\gamma = 40$ after absorbing factors of n, k, and R into γ (see (S4) in Supplementary Information). For much smaller values of γ, selection coeffi-

cient estimates are unstable due to sampling noise. However, inferred selection coefficients stabilize and become insensitive to the precise value of γ for $\gamma \gtrsim 10$ (**Supplementary Fig. 6**). Larger values of γ will result in selection coefficients with smaller absolute values, but for large enough γ the rank ordering of inferred selection coefficients is highly reliable. In summary, the coefficients that appear to be the most beneficial or deleterious remain this way regardless of reasonable choices for γ, though their precise values scales with the regularization strength.

Rapid detection of variants with increased transmission

To estimate how quickly we can detect a transmission advantage for a new SNV or variant, selection coefficients are calculated only in the specific region where the variant arose. Since inference is only done in a single region, SNVs that appear only briefly at low frequencies —and which therefore are unlikely to change transmission rate —only appear once, whereas in the global analysis such SNVs may appear at low frequencies in multiple regions. For this reason we use a lower regularization of 10 for regional analysis. The null distribution is calculated by first finding all variants (including one or more SNVs) that are inferred to have a selection coefficient of absolute value less than 1% using the joint inference over all regions. We then calculated the selection coefficients that would have been inferred for these variants at all earlier time points in each region after they were first observed in that region. We can then say with high confidence that a variant increases transmission once the inferred coefficient for that variant in a specific region surpasses any of the inferred coefficients in the null distribution.

Effects of travel on inferred selection

Travel of infected individuals can bias inferred selection coefficients by changing the frequency of variants in a region for reasons that are not due to increased or decreased transmission. To analyze the effect that travel has on the inferred selection coefficients, we focused on the United Kingdom, and especially on the variant 20E (EU1), because sampling the UK is excellent and because there exists a high-quality estimate for the number of importations of this variant[61]. In order to quantify the effect of travel, it is important to have an estimate for the number of newly infected individuals on each day since the effect due to travel depends on the number of cases that are imported relative to the number of local cases (see 6). We used statistics from the Institute for Health Metrics and Evaluation[62] to estimate the number of newly infected individuals in the UK. For simplicity, we assumed a constant number of importations of 20E (EU1) per day starting on July 7th, 2020 (the first day a 20E (EU1) sequence was sampled in the UK) and continuing for 100 days. We then inferred the selection coefficients for many different numbers of importations. The results are shown in **Supplementary Fig. 7**, where we find that a very large number of importations is necessary for 20E (EU1) to be inferred to be neutral ($\hat{w} = 0$).

In our full data set, selection coefficients that are inferred to be close to zero may in fact be slightly beneficial or deleterious and are inferred incorrectly due to travel. However, given the degree of travel needed to substantially bias inferred selection demonstrated in **Supplementary Fig. 7**, travel is unlikely fully explain large inferred selection coefficients. This is especially true for variants observed in regions where travel restrictions reduced the number of infected individuals entering or leaving a region. In addition, the effects of travel are also muted when the number of local infections is large.

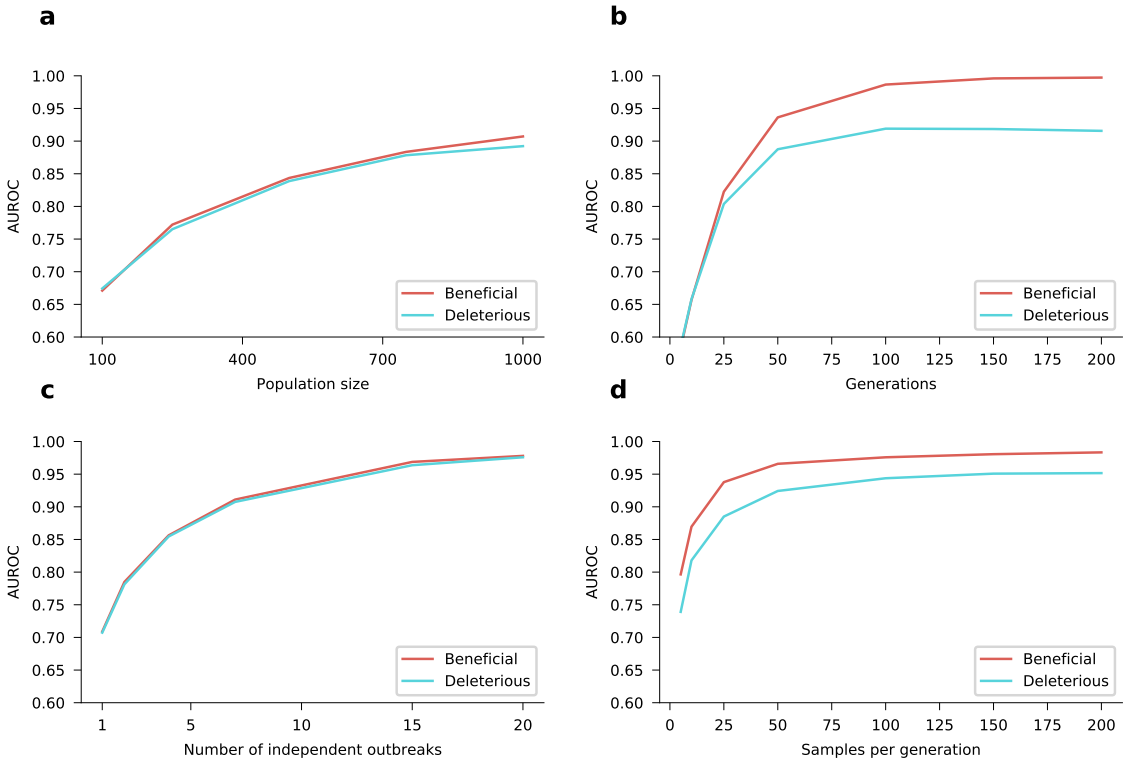
Data and code

Sets of processed data, computer code, and scripts that we have used in our analysis are available in the GitHub repository located at https://github.com/bartonlab/paper-SARS-CoV-2-inference. This repository also contains Jupyter notebooks that can be run to reproduce the results presented here, using sequence data and metadata from GISAID.

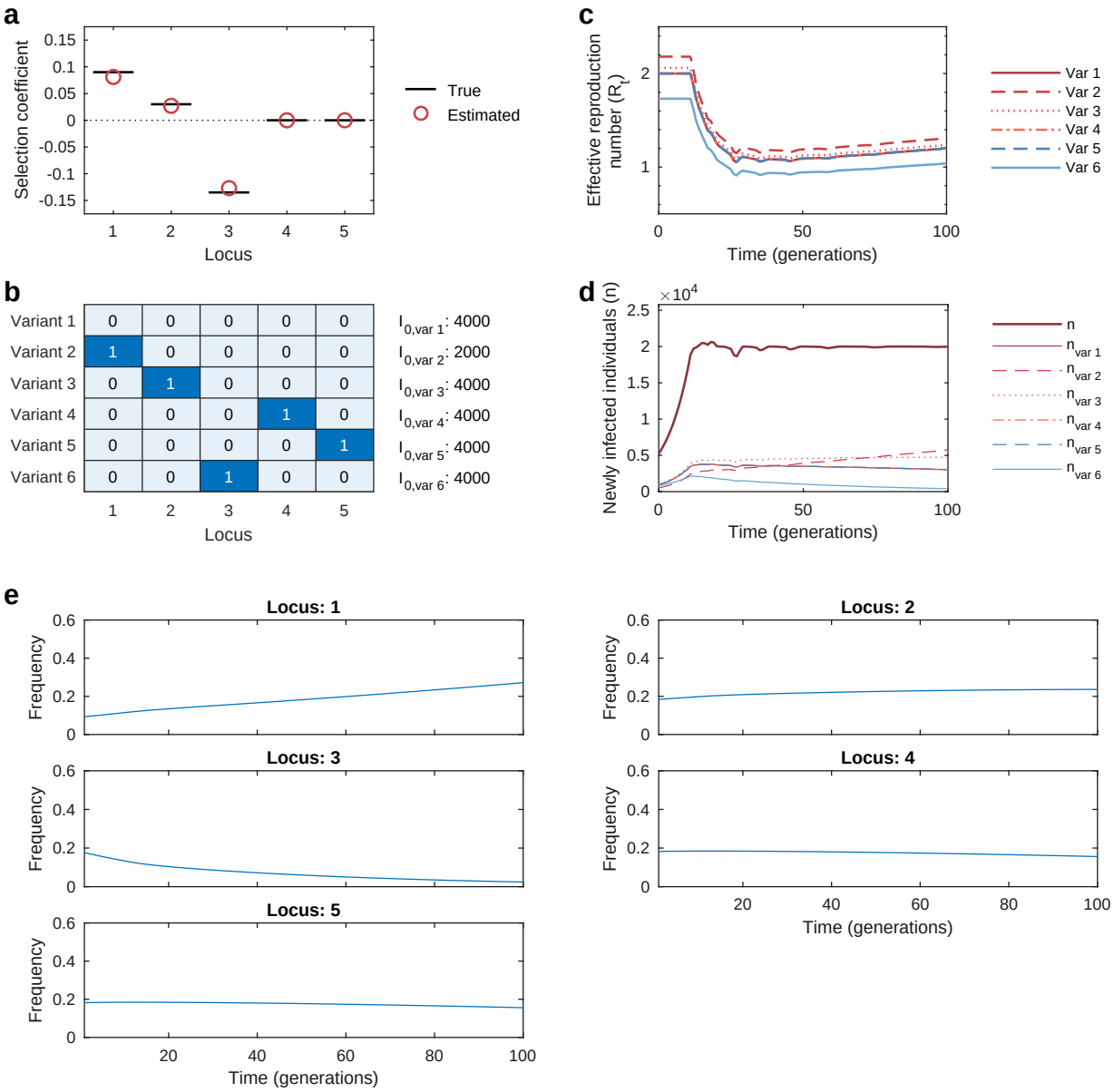
References

42. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
43. Kimura, M. Diffusion models in population genetics. *Journal of Applied Probability* **1**, 177–232 (1964).
44. Ewens, W. J. *Mathematical Population Genetics 1: Theoretical Introduction* (Springer Science & Business Media, 2012).
45. Malaspinas, A.-S., Malaspinas, O., Evans, S. N. & Slatkin, M. Estimating allele age and selection coefficient from time-serial data. *Genetics* **192**, 599–607 (2012).
46. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International journal of infectious diseases* **92**, 214–217 (2020).
47. Systrom, K., Vladek, T. & Krieger, M. Model powering rt.live. https://github.com/rtcovidlive/covid-model (2020).
48. Dietz, K. The estimation of the basic reproduction number for infectious diseases. *Statistical methods in medical research* **2**, 23–41 (1993).
49. D'Arienzo, M. & Coniglio, A. Assessment of the SARS-CoV-2 basic reproduction number, R0, based on the early phase of COVID-19 outbreak in Italy. *Biosafety and Health* **2**, 57–59 (2020).
50. Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T. & Jacobsen, K. H. Complexity of the basic reproduction number (R0). *Emerging Infectious Diseases* **25**, 1–4 (2019).
51. Clark, S. J. & Perry, J. N. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics* 309–316 (1989).
52. Saha, K. & Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* **61**, 179–185 (2005).
53. Hilbe, J. M. *Negative binomial regression* (Cambridge University Press, 2011).
54. Miller, A. C. *et al.* Statistical deconvolution for inference of infection time series. *medRxiv* 2020.10.16.20212753 (2020).
55. Manski, C. F. & Molinari, F. Estimating the COVID-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics* **220**, 181–192 (2021).
56. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global challenges* **1**, 33–46 (2017).
57. Sohail, M. S., Louie, R. H. Y., McKay, M. R. & Barton, J. P. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature Biotechnology* **39**, 472–479 (2021).
58. Pung, R. *et al.* Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *Lancet* **395**, 1039–1046 (2020).
59. Du, Z. *et al.* Serial interval of COVID-19 among publicly reported confirmed cases. *Emerging infectious diseases* **26**, 1341 (2020).
60. Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (COVID-19) infections. *International journal of infectious diseases* **93**, 284–286 (2020).
61. Hodcroft, E. B. *et al.* Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
62. Institute of health metrics and evaluation, SARS-CoV-2 estimates of newly infected per day. URL http://www.healthdata.org/covid/data-downloads.
63. Li, Q. *et al.* The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* **182**, 1284–1294.e9 (2020).
64. Deng, X. *et al.* Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell* **184**, 3426–3437.e8 (2021).

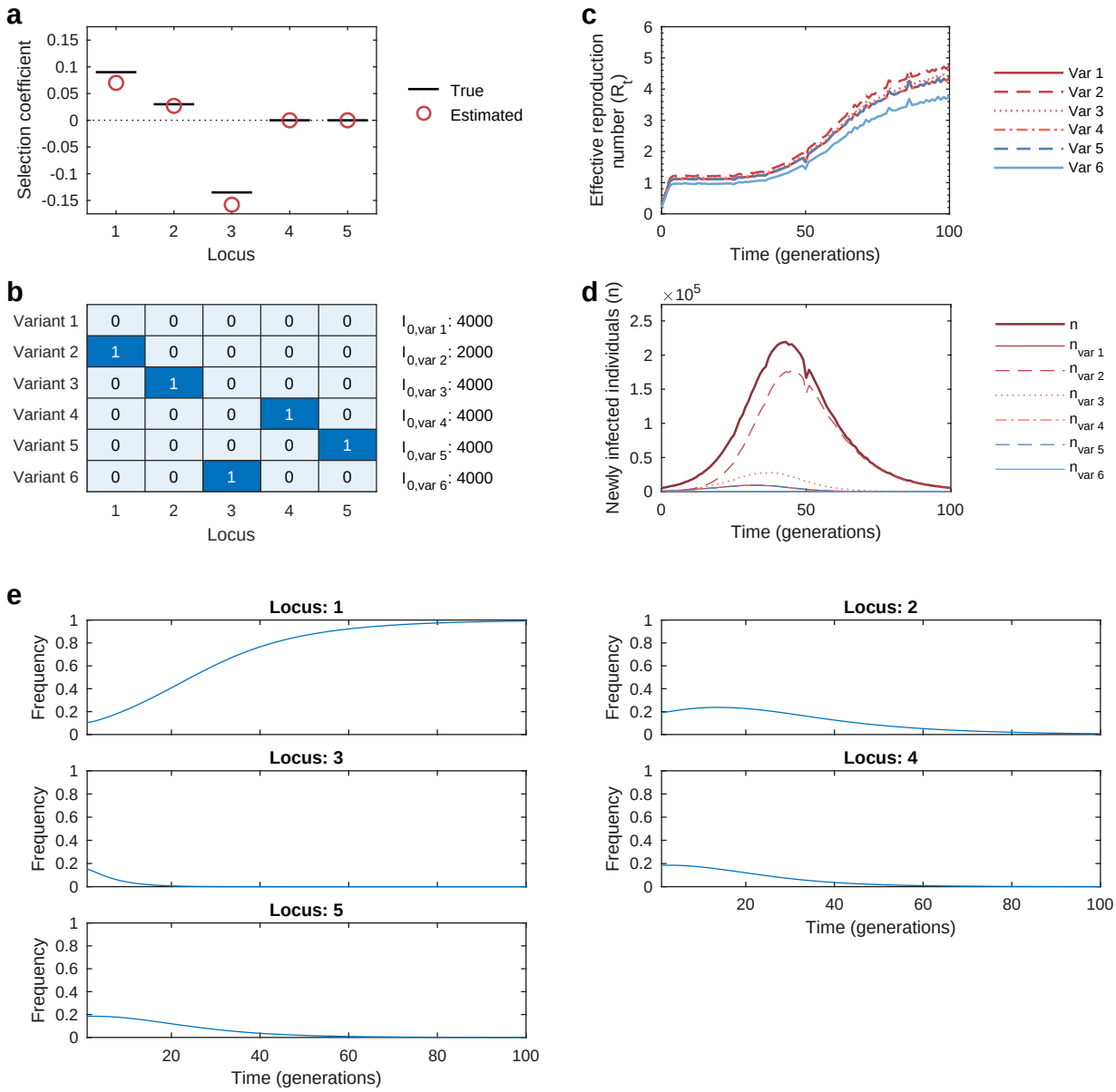
65. Saito, A. *et al.* Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature* 1–10 (2021).
66. Mohammad, A., Abubaker, J. & Al-Mulla, F. Structural modelling of SARS-CoV-2 Alpha variant (B.1.1.7) suggests enhanced furin binding and infectivity. *Virus Research* **303**, 198522 (2021).
67. Lista, M. J. *et al.* The P681H mutation in the spike glycoprotein confers type I interferon resistance in the SARS-CoV-2 alpha (B.1.1.7) variant. *bioRxiv* 2021.11.09.467693.
68. Parker, M. D. *et al.* Altered subgenomic RNA expression in SARS-CoV-2 B.1.1.7 infections. *bioRxiv* 2021.03.02.433156 (2021).
69. Xia, H. *et al.* Evasion of type I interferon by SARS-CoV-2. *Cell Reports* **33**, 108234 (2020).
70. Dejnirattisai, W. *et al.* Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184**, 2939–2954.e9 (2021).
71. Liu, C. *et al.* Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184**, 4220–4236.e13 (2021).
72. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347.e16 (2021).
73. Syed, A. M. *et al.* Rapid assessment of SARS-CoV-2 evolved variants using virus-like particles. *Science* **374**, 1626–1632 (2021).
74. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310 (2020).
75. Cao, Z. *et al.* Ubiquitination of SARS-CoV-2 ORF7a promotes antagonism of interferon response. *Cellular & molecular immunology* **18**, 746–748 (2021).



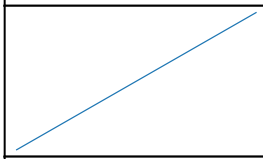
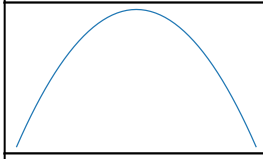
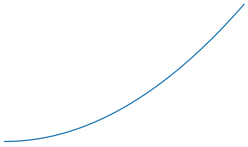
Supplementary Fig. 1. Accuracy of inference for different parameters. How the AUROC scores for both beneficial SNVs (in red) and deleterious SNVs (in blue) depends upon the different model parameters. **a**, Inference accuracy for different values of newly-infected population size. The parameters used are 10 simulations each with 50 sampled genomes per generation for 25 generations. **b**, Inference accuracy for different numbers of generations (serial intervals). Data is from a single simulation with 25 samples per generation and a newly-infected population size of 10,000. **c**, Inference accuracy for different numbers of independent outbreaks (simulations). The parameters used are 50 samples per generation for 10 generations and a newly-infected population size of 10,000. **d**, Inference accuracy for different values of samples per generations. Data is from a single simulation with 50 generations with a newly-infected population size of 10,000. The initial population is a mixture of two variants with beneficial SNVs ($s = 0.03$), two with neutral SNVs ($s = 0$), and two with deleterious SNVs ($s = -0.03$). Dispersion parameter k is fixed at 0.1. This is the same initial population composition as described in **Fig. 1**. All AUROC scores are calculated by averaging over 1,000 replicate simulations.



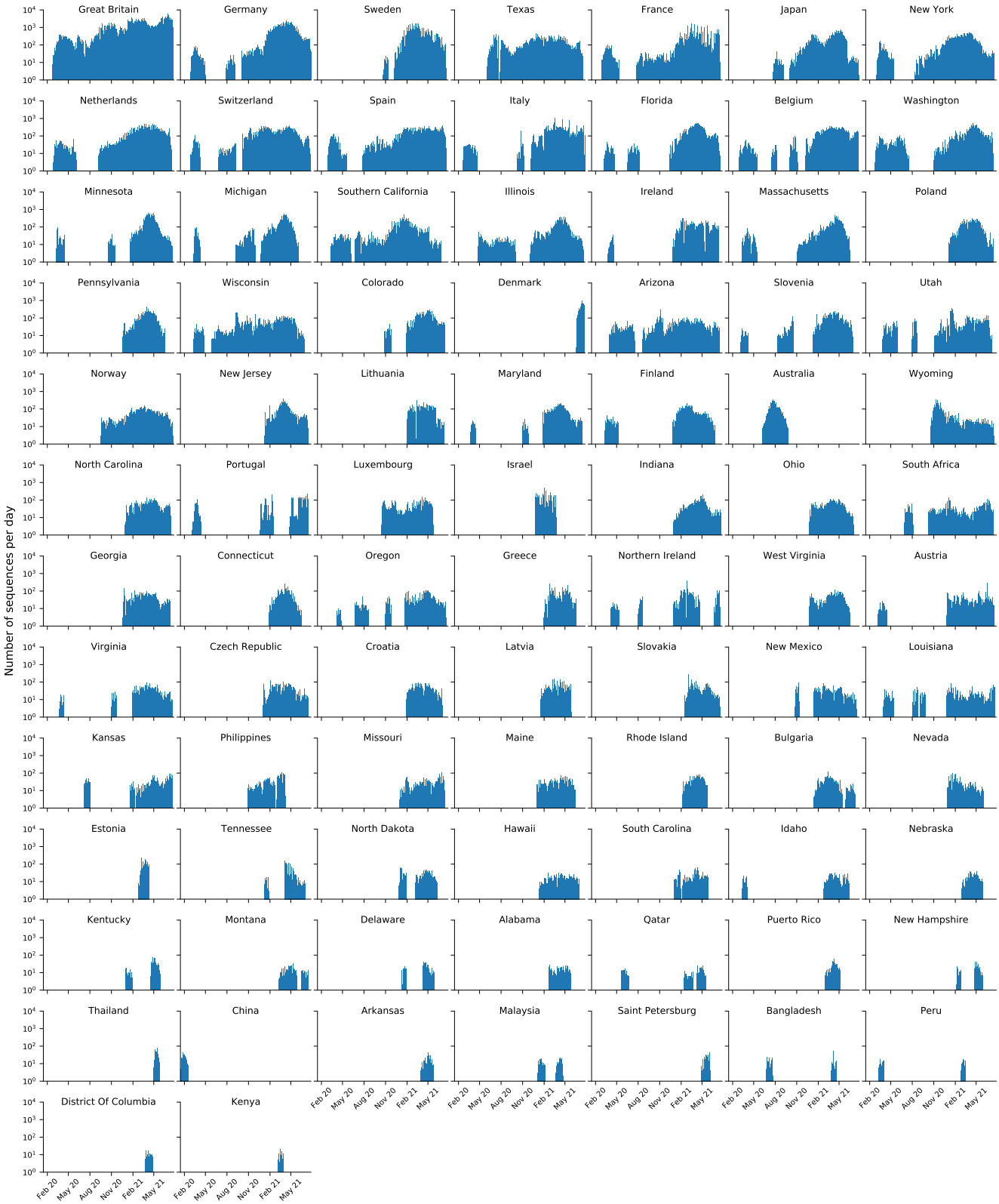
Supplementary Fig. 2. Estimated selection coefficients are accurate even under a deterministic SIR model of disease spread. Here the simulation parameters are such that the number of newly infected individuals increases over time and then remains almost constant. **a**, True and estimated selection coefficients. We observe that the estimates are quite accurate. **b**, Variants present in the population along with the number of individuals infected by each variant at the initial time. **c**, Plot of effective reproduction number (R_t) of each variant over time. In this simulation we adjust the value of R_t by adaptively updating transmission rate β of each variant such that the total number of newly-infected individuals at each time remains almost constant. **d**, Plot of individuals newly-infected by each variant along with the total number of newly-infected individuals. **e**, Mutant frequency trajectories observed at each locus. *Simulation parameters*: Recovery rate for all variants is assumed to be the same, i.e., $r_a = r = 0.12$, population size $N = 10,000,000$, the effective reproduction number R_t of the WT variant is initialized to a value of 2 while R_t for other variants is calculated from (11).



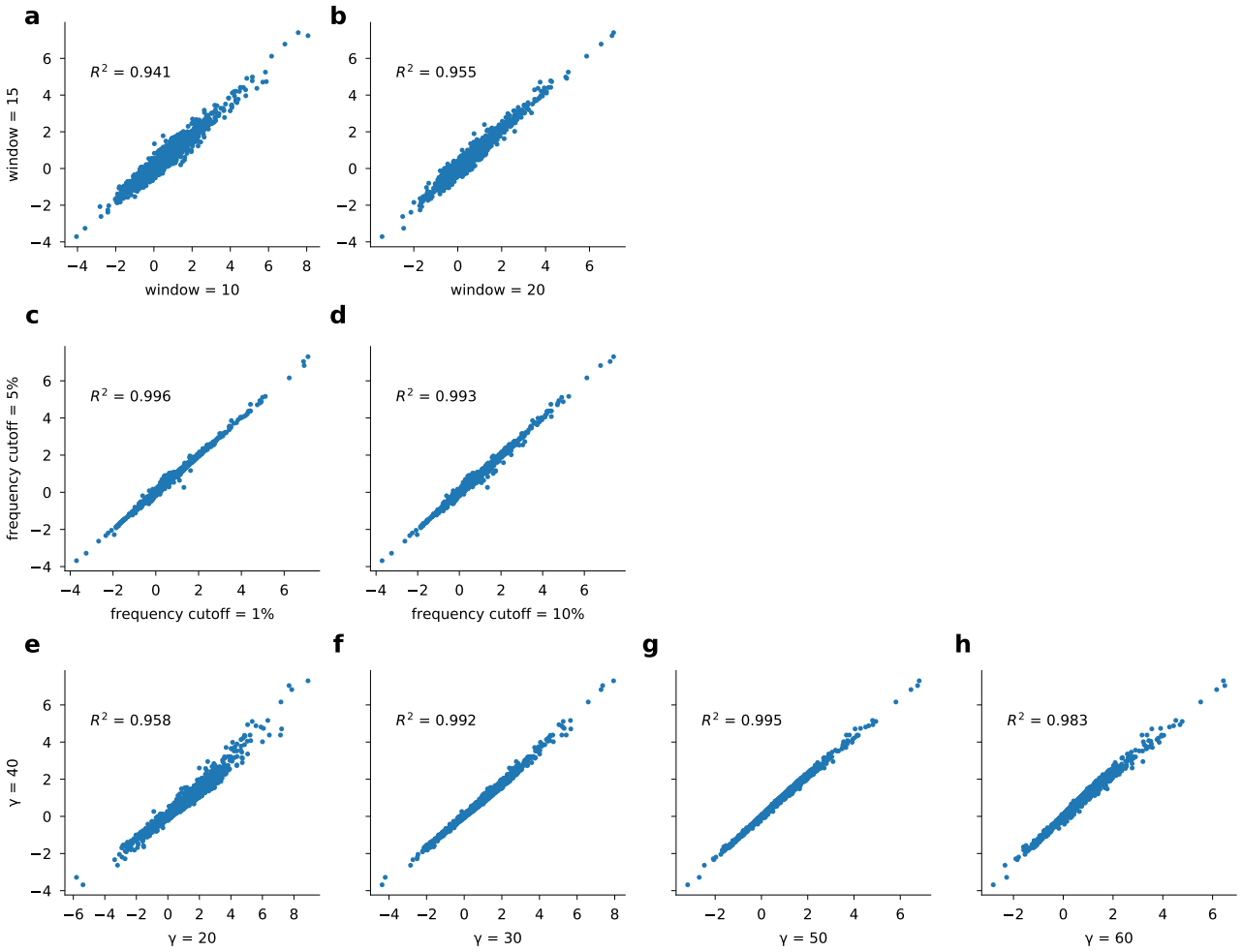
Supplementary Fig. 3. Estimated selection coefficients are accurate even under a deterministic SIR model of disease spread. Here the simulation parameters are such that the number of newly-infected individuals follow the typical SIR dynamics. **a**, True and estimated selection coefficients. We observe that the estimates are quite accurate. **b**, Variants present in the population along with the number of individuals infected by each variant at the initial time. **c**, Plot of effective reproduction number (R_t) of each variant over time. In this simulation we adjust the value of R_t of each variant such that the total number of newly-infected individuals at each time follows typical SIR dynamics. **d**, Plot of individuals newly-infected by each variant along with the total number of newly-infected individuals. **e**, Mutant frequency trajectories observed at each locus. *Simulation parameters*: Recovery rate for all variants is assumed to be the same, i.e., $r_a = r = 1$, population size $N = 10,000,000$, the transmission rate is adapted such that system follows the typical SIR dynamics.

Population Size	Sampling	Inference Parameter (N)	AUROC Beneficial	AUROC Deleterious
	Finite	Time-Varying	0.832	0.779
		Constant	**0.937**	**0.881**
	Perfect	**Time-Varying**	**0.999**	**0.992**
		Constant	0.973	0.940
	Finite	Time-Varying	0.873	0.821
		Constant	**0.944**	**0.882**
	Perfect	**Time-Varying**	**1.0**	**0.999**
		Constant	0.986	0.950
	Finite	Time-Varying	0.798	0.736
		Constant	**0.873**	**0.824**
	Perfect	**Time-Varying**	**0.981**	**0.935**
		Constant	0.905	0.863

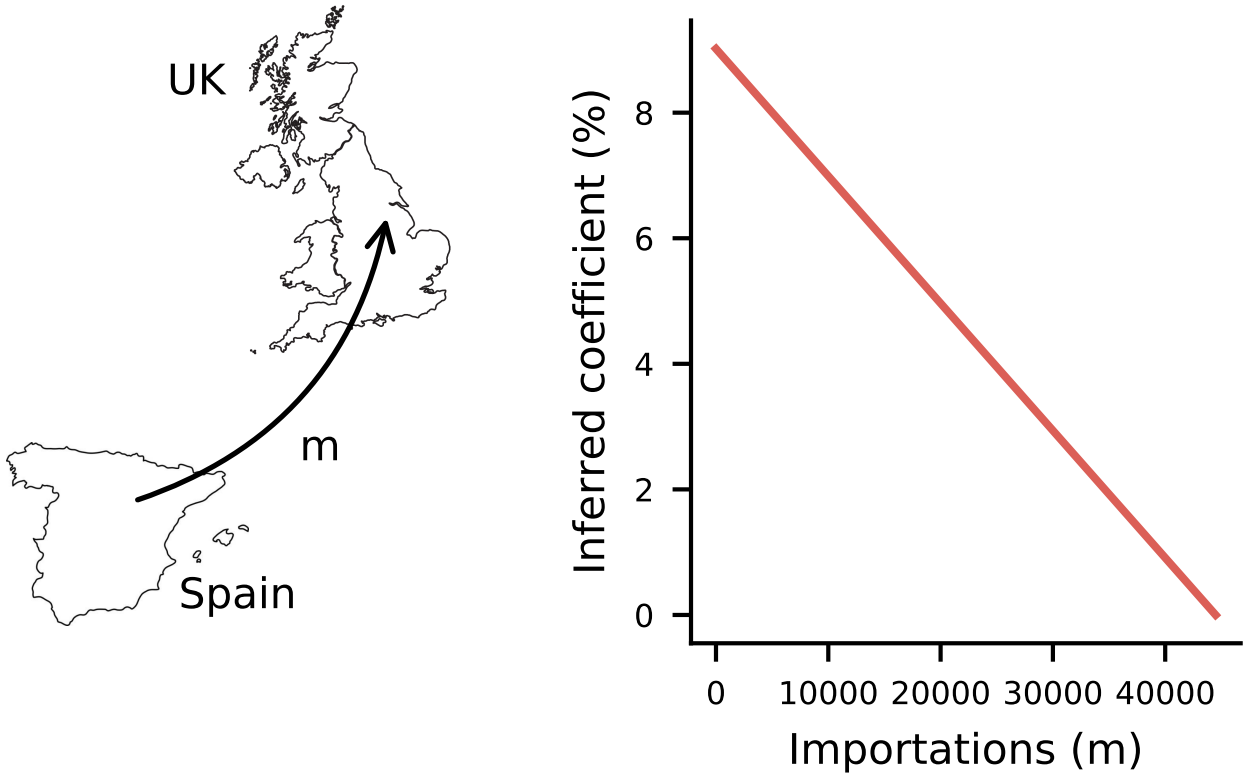
Supplementary Fig. 4. Effects of finite sampling on inference using constant and time-varying parameters. The ability of the model to distinguish beneficial and deleterious SNVs, as measured by the AUROC score, depending on whether the sampling is perfect or finite and whether constant arbitrary parameters or the true time-varying parameters are used for the population size n in the inference. Both simulations use constant values of $k = 0.01$ and $R = 1$. The results are similar but less dramatic if the correct time-varying values are used for k or R as well. Results are shown for different population trajectories and are consistent regardless of the trajectory. Rows that yield better inference are marked by bold text. If the sampling is finite, then it is better to use constant parameters; if the sampling is perfect, then it is better to use the real time-varying parameters. The initial population is a mixture of two variants with beneficial SNVs ($s = 0.03$), two with neutral SNVs ($s = 0$), and two with deleterious SNVs ($s = -0.03$), which is the same as that used in **Fig. 1**. Simulations are run for 50 simulations with 25 samples in each generation, and AUROC scores are averaged over 1,000 replicate simulations.



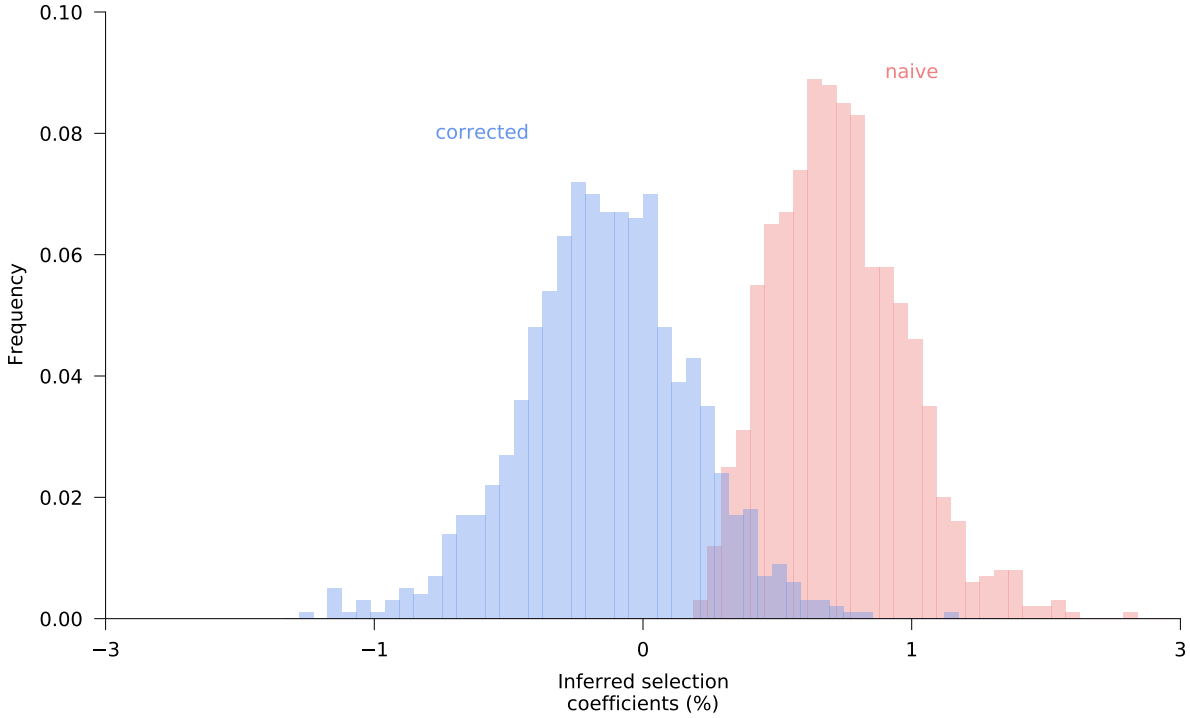
Supplementary Fig. 5. Sampling Distributions. The number of genomes per day in the regions that are used for inference.



Supplementary Fig. 6. Inferred selection coefficients are robust to different values of the regularization γ, different frequency cutoffs, and different numbers of days used to calculate the frequency changes. a-b, Comparison of inferred coefficients when the number of days at the beginning and end of the time-series are used in order to calculate the frequency changes. Inferred coefficients are largely robust to these changes **c-d**, Comparison of inferred coefficients for different frequency cutoffs. Including more or less sites does not alter the order of inferred coefficients. **e-h**, Comparison of inferred coefficients for different values of the regularization. Altering the regularization value has little effect upon the distribution of inferred selection coefficients, and selection coefficients for different values of the regularization are highly correlated.



Supplementary Fig. 7. Travel into the United Kingdom in unlikely to explain the apparent increase in transmission rate of the 20E (EU1) variant. A plot of the inferred coefficient for the 20E (EU1) variant versus the total number of importations of this variant into the United Kingdom. The number of importations must be very large for the inferred coefficient to be small, indicating that travel of individuals infected with the variant 20E (EU1) is unlikely to fully explain the apparent fitness benefit provided by the group of SNVs.



Supplementary Fig. 8. Correction to inference due to travel. The distribution of inferred selection coefficients for a neutral SNV when individuals who are infected with the variant travel into a regional outbreak. *Simulation parameters*: $n = 10,000$ and $k = 0.1$. The number of newly infected individuals per serial interval rises rapidly from 6,000 to around 10,000 and stays nearly constant thereafter. The initial population is a mixture of two variants with beneficial SNVs ($s = 0.03$), two with neutral SNVs ($s = 0$), and two with deleterious SNVs ($s = -0.03$). This is the same initial population composition as in **Supplementary Figure 1**. 25 individuals infected with a variant containing a single neutral mutation travel into the population per generation for 100 generations. The distribution of inferred coefficients for the neutral SNV including travel (blue) and not including travel (red) is shown over 1,000 replicate simulations.

Rank	Protein	Mutation(s) (nt)	Mutation (aa)	Selection (%)	Location	Phenotypic effect
1	S	T22917G	L452R	7.5	RBM	Increased resistance to nAbs [63] and increased cell entry [64]
2	M	T26767C	I82T	7.5		
3	S	C23604G	P681R	7.2	FCS	Enhanced cleavage, fusogenicity, and pathogenicity [65]
4	S	C23604A	P681H	6.5	FCS	Enhanced cleavage [66] and increased resistance to interferon-induced immunity [67], leading to increased replication and/or transmission
5	N	G28280C, A28281T, T28282A#	D3L	6.0		Increased transmissibility by introducing a transcription regulatory sequence upstream of ORF9b [68]
6	NSP13	C16466T	P77L	5.5		
7	NSP6	Δ11288-11290	S106-	5.4		*Increased transmission by interferon antagonism [69]
8	NSP4	C10029T	T492I	5.3		
9	S	G23012A	E484K	5.2	RBM	Increased resistance to nAbs [63] and increased ACE2 binding [70]
10	S	G24410A	D950N	5.1	HR1	
11	S	C21618G	T19R	5.1	NTD	*Increased resistance to NTD-specific nAbs [71,72]
12	N	A28299T	Q9L	4.6		
13	N	G28881T	R203M	4.6		Enhanced replication, RNA delivery and packaging [73]
14	S	G24368T	D936Y	4.4	HR1	
15	S	G22992A	S477N	4.4	RBM	Increased ACE2 binding [74]
16	ORF7a	C27752T	T120I	4.4		*Mutation at residue 119 results in complete loss of ubiquitination and partial loss of interferon pathway inhibition [75]
17	S	C23604T	P681L	4.4	FCS	*Enhanced cleavage [66]
18	N	A28461G	D63G	4.3		
19	NSP12	C15952A	L838I	4.3		
20	S	T22917A	L452Q	4.2	RBM	*Increased resistance to nAbs [63] and increased cell entry [64]
21	NSP6	A11201G	T77A	4.2		
22	M	T26767G	I82S	4.2		
23	NSP6	Δ11291-11293	G107-	4.2		*Increased transmission by interferon antagonism [69]
24	S	C22995A	T478K	4.1	RBM	Increased resistance to nAbs [63]
25	NSP14	C18086T	T16I	4.0		
26	ORF3a	C25469T	S26L	3.9		
27	N	G29402T	D377Y	3.8		
28	NSP12	G15451A	G671S	3.8		
29	S	C22227T	A222V	3.8	NTD	*Slightly increased cell entry [61]
30	S	Δ21986-21988	G142-	3.6		
31	NSP3	C5184T	P822L	3.5		
32	N	Δ28877-28879	S202-	3.5		Enhanced replication, RNA delivery and packaging [73]
33	ORF3a	Δ26158-26160	V256-	3.5		
34	ORF3a	C25904T	S171L	3.4		
35	NSP12	G14030A	R197Q	3.4		
36	S	G21974T	D138Y	3.4	NTD	
37	S	A22320G	D253G	3.4	NTD	Increased resistance to NTD-specific nAbs [71,72]
38	NSP8	C12357T	T89I	3.4		
39	N	G28878A	S202N	3.3		Enhanced replication, RNA delivery and packaging [73]
40	N	C28887T	T205I	3.3		Improved RNA delivery and packaging [73]
41	M	G26730C	V70L	3.2		
42	S	G23012C	E484Q	3.2	RBM	*Increased resistance to nAbs [63] and increased ACE2 binding [70]
43	NSP6	Δ11294-11296	F108-	3.1		*Increased transmission by interferon antagonism [69]
44	S	C24642T	T1027I	3.1		
45	S	C21614T	L18F	3.1	NTD	Increased resistance to NTD-specific nAbs [72]
46	N	G28975A, G28975T, G28975C†	M234I	3.1		
47	S	Δ21983-21985	L141-	3.1	NTD	*Increased resistance to NTD-specific nAbs [72]
48	NSP6	A11451G	Q160R	3		
49	ORF8	T28251C	F120L	3		
50	NSP14	C19161T	S374F	3		

Table 1. Table of most highly selected amino acid substitutions across the SARS-CoV-2 genome. * represents the cases where phenotypic effect of an amino acid variant has not been reported explicitly in the literature. Instead, it is either based on the function of the encompassing gene, for a mutation to a different amino acid or deletion at the same position, or for a mutation at a neighboring position. # all three mutations appear together; † each individual mutation leads to the same amino acid mutation; RBM = receptor binding motif; NTD= N-terminal domain; FCS = S1/S2 furin cleavage site; HR1 = heptad repeat 1; nAbs = neutralizing antibodies.

Variant	Pango Lineage	Selection Coefficient (%)	Mutations
B.1	B.1	9	S-D614G, NSP12-P323L
20E-EU1	B.1.177	19.3	S-A222V, S-D614G, NSP12-P323L, N-A220V, ORF10-V30L
Epsilon	B.1.427/B.1.429	25.2	S-L452R, N-T205I, S-D614G, NSP12-P323L, NSP13-D260Y, S-S13I, S-W152C, NSP9-I65V, ORF3a-Q57H
Lambda	C.37	39.9	NSP4-T492I, S-L452Q, NSP3-P1469S, NSP4-L438P, S-D614G, NSP6-S106-, NSP12-P323L, NSP5-G15S, S-F490S, NSP6-G107-, NSP6-F108-, S-T859N, NSP3-T428I, S-G75V, S-T76I, S-P251-, S-T250-, N-G214C, S-G252-, S-D253-, S-S247-, S-Y248-, S-L249-, NSP3-F1569V, S-R246-
Beta	B.1.351	44.8	S-E484K, N-T205I, S-N501Y, S-D215G, S-A701V, S-D614G, NSP6-S106-, NSP12-P323L, NSP6-G107-, NSP5-K90R, NSP6-F108-, NSP3-K837N, S-D80A, S-A243-, E-P71L, S-L242-, S-K417N, S-L241-, S-T240-, NSP2-T85I, ORF3a-Q57H
Alpha	B.1.1.7	51.7	S-P681H, S-N501Y, S-D614G, N-R203K, S-H69-, N-D3L, NSP6-S106-, NSP12-P323L, S-T716I, NSP6-G107-, NSP3-A890D, NSP6-F108-, S-D1118H, ORF8-Q27*, S-S982A, S-Y144-, ORF8-R52I, N-S235F, NSP3-T183I, S-V143-, ORF8-Y73C, S-A570D, N-G204R, S-V70-, S-I68-, NSP3-I1412T
Gamma	P.1	56.6	S-L18F, S-T1027I, S-N501Y, S-H655Y, S-D614G, N-R203K, S-V1176F, NSP6-S106-, NSP12-P323L, NSP6-G107-, ORF8-E92K, N-P80R, S-K417T, NSP6-F108-, S-R190S, NSP3-K977Q, S-T20N, ORF3a-S253P, S-P26S, NSP3-S370L, NSP13-E341D, N-G204R
Delta	B.1.617.2	84.2	M-I82T, S-P681R, S-L452R, NSP13-P77L, S-T19R, S-D950N, N-R203M, ORF7a-T120I, N-D63G, S-T478K, N-D377Y, ORF3a-S26L, NSP12-G671S, ORF7a-V82A, S-D614G, NSP12-P323L, ORF8-F120-, S-F157-, ORF8-D119-, S-R158-, S-E156-

Table 2. Table of selection coefficients for groups of amino acid mutations. Mutations that contribute most strongly to selection are listed first. The selection coefficient for a variant is calculated as the sum of the selection coefficients for the individual mutations that the variant contains.

Supplementary Information

1. Summary

Here we discuss two main topics. First, we give a detailed introduction of our epidemiological model as well as a derivation of the estimator (1) and an important simplification of it. Second, describe simulations of an outbreak and show that selection coefficients can be accurately recovered from simulation data even with relatively poor sampling.

2. Epidemiological model

2.1. Introduction

In epidemiology, the spread of infection can be modeled as a branching process where each infected individual (also referred to as a case) infects n additional individuals[76]. The distribution of n is often taken to be Poisson, but differences in the number of contacts with susceptible individuals, disease course within an individual, and other factors mean that the Poisson rate λ is not generally the same for all cases[77]. Below, we first follow ref.[77] to explore families of distributions for the number of new cases per infected individual. Next, we extend these models to consider multiple variants of the pathogen that differ in their spreading efficiency. We seek to characterize how the distribution of pathogen variant frequencies is expected to change over time, and how such data can be used to estimate the relative spreading efficiency of different variants.

2.2. Distributions for the number of infected individuals

As noted above, the basic distribution of the number of new cases n caused by one case in a susceptible population is Poisson,

$$P_{\mathrm{P}}(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}\,.$$

Typically we might take the Poisson rate λ to be R, the effective reproduction number, which is the expected number of cases directly caused by one case. In that case, the average number of cases following the Poisson distribution is

$$\langle n \rangle_{P_{\mathrm{P}}(n|R)} = \sum_{n=0}^{\infty} n P_{\mathrm{P}}(n|R) = R\,.$$

To account for variability in transmission dynamics, the basic Poisson distribution with a single rate R can be replaced with a continuous mixture of Poisson distributions, where the rate parameter λ follows a gamma distribution,

$$P_{\Gamma}(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\,,$$

with shape parameter α and rate parameter β. The average value of λ is

$$\langle \lambda \rangle_{P_{\Gamma}(\lambda|\alpha,\beta)} = \frac{\alpha}{\beta}\,,$$

and its variance is

$$\left\langle \left(\lambda - \frac{\alpha}{\beta}\right)^2 \right\rangle_{P_{\Gamma}(\lambda|\alpha,\beta)} = \frac{\alpha}{\beta^2}\,.$$

In this context, it is natural to take $\alpha = k$ and $\beta = k/R$. With these choices, the gamma distribution reads

$$P_{\Gamma}(\lambda|k,R) = \frac{1}{\Gamma(k)}\left(\frac{k}{R}\right)^k \lambda^{k-1}e^{-k\lambda/R}\,. \tag{S1}$$

The parameter k is a dispersion parameter that determines how long-tailed the distribution is. The mean value of λ is always R, but when k is smaller its variance increases. In the limit that $k \to \infty$, we recover the pure Poisson distribution with rate $\lambda = R$. When $k = 1$, the distribution of the number of cases n is geometric,

$$\int_0^\infty d\lambda\, P_{\Gamma}(\lambda|k=1,R)\, P_{\mathrm{P}}(n|\lambda) = P_g(n|p) = (1-p)^n\, p\,,$$

where $p = 1/(1+R)$. For arbitrary values of $k > 0$, the number of cases follows a negative binomial distribution,

$$P_{\mathrm{NB}}(n|k,R) = \frac{\Gamma(k+n)}{n!\,\Gamma(k)}\left(\frac{k}{k+R}\right)^k \left(\frac{R}{k+R}\right)^n\,.$$

The standard parameters of the negative binomial distribution are r and p, which are set to k and $k/(k+R)$ in our parameterization above.

2.3. Dynamics for the pathogen variant frequencies

Let us assume that there exist multiple variants of a pathogen, which are distinguished by an index a. The number of cases infected with variant a is n_a. We assume that different variants have slightly different transmission probabilities, so that $R_a = R(1+w_a)$, with $|w_a| \ll 1$. The term w_a is analogous to a selection coefficient in population genetics.

2.3.1. Dynamics of multiple cases infected by a single variant

First, let us assume that n individuals, each labeled by an index i, are all infected by the same variant of a pathogen. For now we will assume that there is no travel into or out of the population, though we will include it later. How many cases will be generated from these individuals? The number of new cases for all individuals is

$$n' = \sum_{i=1}^{n} n'_i \,,$$

where the numbers of cases n'_i generated by individual i follows a negative binomial distribution. Because all individuals are infected by the same variant, the negative binomial parameter $p = k/(k+R)$ is the same for each of them. Then, assuming that all of the infection events are independent, it can be shown that the probability distribution for the total number of new cases n' also follows a negative binomial distribution with the same value of p, and with $r = nk$ (that is, the new r parameter value is the sum of the individual r parameter values). Thus, the distribution of n' is

$$P_{\mathrm{NB+}}(n'|k,R,n) = \frac{\Gamma(nk+n')}{n'!\,\Gamma(nk)} \left(\frac{k}{k+R}\right)^{nk} \left(\frac{R}{k+R}\right)^{n'} .$$

2.3.2. Dynamics for multiple cases infected by multiple variants

Let us extend the previous example to consider m variants of a pathogen. At the starting point, the number of individuals infected by a given variant a is n_a, with $a \in \{1,\ldots,m\}$. The fraction of cases infected by variant a is

$$y_a = \frac{n_a}{\sum_{b=1}^{m} n_b} \,.$$

Now, we would like to know how the fraction of individuals infected by each variant is expected to change with each round of infections. In other words, for variant a, we would like to compute

$$\langle y'_a \rangle = \left\langle \frac{n'_a}{\sum_{b=1}^{m} n'_b} \right\rangle = \sum_{\boldsymbol{n}'} \left(\prod_{b=1}^{m} P_{\mathrm{NB+}}(n'_b|k, R(1+w_b), n_b) \right) \frac{n'_a}{\sum_{c=1}^{m} n'_c}$$

where the outer sum is over all vectors $\boldsymbol{n}'$ with entries $\{n'_1, n'_2, \ldots\}$, and with $n'_b \geq 0$ for all b. Here, we have assumed that the n'_b's are independent across b.

To proceed, it is convenient to write the negative binomial distributions as mixtures of Poisson distributions (as indicated above), giving

$$\begin{aligned}
\langle y'_a \rangle &= \sum_{\boldsymbol{n}'} \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_b)\right) P_{\mathrm{P}}(n'_b|\lambda_b) \right) \frac{n'_a}{\sum_{c=1}^{m} n'_c} \\
&= \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_b)\right) \right) \sum_{\boldsymbol{n}'} \left(\prod_{b=1}^{m} P_{\mathrm{P}}(n'_b|\lambda_b) \right) \frac{n'_a}{\sum_{c=1}^{m} n'_c} \,.
\end{aligned}$$

Next, we use the fact that the sum of independent Poisson-distributed random variables is also Poisson with rate parameter equal to the sum of the individual rates, and that the distribution of independent Poisson random variables conditioned on their sum is multinomial, to write

$$\begin{aligned}
\langle y'_a \rangle &= \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_b)\right) \right) \sum_{n'=0}^{\infty} P_{\mathrm{P}}\left(n'|\lambda\right) \sum_{\boldsymbol{n}' : \sum_{c=1}^{m} n'_c = n'} P_{\mathrm{M}}\left(\boldsymbol{n}' \middle| n', \frac{\boldsymbol{\lambda}}{\lambda}\right) \frac{n'_a}{n'} \\
&= \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_b)\right) \right) \sum_{n'=0}^{\infty} P_{\mathrm{P}}\left(n'|\lambda\right) \frac{\lambda_a}{\lambda} \\
&= \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_b)\right) \right) \frac{\lambda_a}{\lambda} \,.
\end{aligned}$$

Here $\boldsymbol{\lambda}$ is a vector with entries $\{\lambda_1, \lambda_2, \ldots\}$, and we have also introduced $\sum_a \lambda_a = \lambda$. Note also that the outer sum on the first line is over all vectors $\boldsymbol{n}'$ whose (non-negative) entries sum to n'.

Computing the remaining integrals exactly is challenging, largely because the Gamma distributions have different rate parameters. To address this, next we will expand our expression to first order in the w_a, since these are assumed to be small parameters. Referring back to Eq. (S1), the expansion gives

$$\begin{aligned}\langle y_a' \rangle &= \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma(\lambda_b | n_b k, R) \left[1 - k w_b \left(n_b - \frac{\lambda_b}{R} \right) \right] \right) \frac{\lambda_a}{\lambda} + \mathcal{O}(w^2) \\ &= \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma(\lambda_b | n_b k, R) \right) \left[1 - \sum_{c=1}^{m} k w_c \left(n_c - \frac{\lambda_c}{R} \right) \right] \frac{\lambda_a}{\lambda} + \mathcal{O}(w^2) \, .\end{aligned}$$

Next we change variables to $\{\lambda, q_1 = \lambda_1/\lambda, q_2 = \lambda_2/\lambda, \ldots, q_{m-1} = \lambda_{m-1}/\lambda\}$, because the distribution of the sum of gamma-distributed random variables, λ, with the same rate parameter and the ratios of the individual variables to the total (λ_a/λ) follow independent gamma and Dirichlet distributions[78]. The mth ratio $q_m = 1 - \sum_{a=1}^{m-1} q_a$ by conservation. By convention we will also set $w_m = 0$, which can be thought of as normalizing the value of R relative to a reference genotype. The transformation then gives

$$\begin{aligned}\langle y_a' \rangle &= \int_0^\infty d\lambda \, P_\Gamma(\lambda | nk, R) \left(\prod_{b=1}^{m-1} \int dq_b \right) P_{\mathrm{D}}(\boldsymbol{q} | \boldsymbol{n}k) \left[1 - \sum_{c=1}^{m} k w_c \left(n_c - \frac{\lambda q_c}{R} \right) \right] q_a \\ &= \left(\prod_{b=1}^{m-1} \int dq_b \right) P_{\mathrm{D}}(\boldsymbol{q} | \boldsymbol{n}k) \left[1 - \sum_{c=1}^{m} k w_c (n_c - n q_c) \right] q_a \\ &= \left(1 - k \sum_{c=1}^{m} n_c w_c \right) y_a + \left(\prod_{b=1}^{m-1} \int dq_b \right) P_{\mathrm{D}}(\boldsymbol{q} | \boldsymbol{n}k) \, nk \left(\sum_{c \neq a} w_c q_c q_a + w_a q_a^2 \right) \\ &= \left(1 - nk \sum_{b=1}^{m} w_b y_b \right) x_a + \frac{nk}{nk+1} \left[nk \sum_{b \neq a} w_b y_a y_b + w_a \left(nk y_a^2 + y_a \right) \right] \\ &= y_a + \frac{nk}{nk+1} y_a \left(w_a - \sum_{b=1}^{m} w_b y_b \right) .\end{aligned}$$

In the expressions above $P_{\mathrm{D}}(\boldsymbol{q}|\boldsymbol{\alpha})$ is the Dirichlet distribution, with concentration parameters $\boldsymbol{\alpha}$ given by $\boldsymbol{n}k$ in our case. Note that if $w_m \neq 0$, the last line should instead read

$$\langle y_a' \rangle = y_a + \frac{nk}{nk+1} y_a \left(w_a - w_m - \sum_{b=1}^{m} w_b y_b \right) .$$

Thus, we obtain (with $w_m = 0$)

$$\langle y_a' - y_a \rangle = \langle \Delta y_a \rangle = \frac{nk}{nk+1} y_a \left(w_a - \sum_{b=1}^{m} w_b y_b \right) .$$

Following a similar approach, we can compute the second moments. First, we consider

$$\begin{aligned}\left\langle (y_a')^2 \right\rangle &= \left\langle \left(\frac{n_a'}{\sum_{b=1}^{m} n_b'} \right)^2 \right\rangle \\ &= \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma(\lambda_b | n_b k, R(1+w_b)) \right) \sum_{n'=0}^{\infty} P_{\mathrm{P}}(n' | \lambda) \sum_{\boldsymbol{n}' : \sum_{c=1}^{m} n_c' = n'} P_{\mathrm{M}}\left(\boldsymbol{n}' \middle| n', \frac{\boldsymbol{\lambda}}{\lambda} \right) \left(\frac{n_a'}{n'} \right)^2 \\ &= \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma(\lambda_b | n_b k, R(1+w_b)) \right) \sum_{n'=0}^{\infty} P_{\mathrm{P}}(n' | \lambda) \left[\left(\frac{\lambda_a}{\lambda} \right)^2 + \frac{1}{n'} \frac{\lambda_a}{\lambda} \left(1 - \frac{\lambda_a}{\lambda} \right) \right] \\ &\approx \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma(\lambda_b | n_b k, R) \right) \left[1 - \sum_{c=1}^{m} k w_c \left(n_c - \frac{\lambda_c}{R} \right) \right] \left[\left(\frac{\lambda_a}{\lambda} \right)^2 + \frac{1}{\lambda} \frac{\lambda_a}{\lambda} \left(1 - \frac{\lambda_a}{\lambda} \right) \right] \\ &= \int_0^\infty d\lambda \, P_\Gamma(\lambda | nk, R) \left(\prod_{b=1}^{m-1} \int dq_b \right) P_{\mathrm{D}}(\boldsymbol{q} | \boldsymbol{n}k) \left[1 - \sum_{c=1}^{m} k w_c \left(n_c - \frac{\lambda q_c}{R} \right) \right] \left[q_a^2 + \frac{q_a(1-q_a)}{\lambda} \right] .\end{aligned}$$

In going from the third to the fourth line above, we have made the approximation that

$$\left\langle \frac{1}{n'} \right\rangle_{P_{\mathrm{P}}(n'|\lambda)} \approx \frac{1}{\lambda},$$

which is valid for $\lambda \gtrsim 1$. Similarly,

$$\begin{aligned}
\langle y'_a y'_b \rangle &= \left\langle \frac{n'_a n'_b}{\left(\sum_{c=1}^m n'_c\right)^2} \right\rangle \\
&= \int_0^\infty \left(\prod_{c=1}^m d\lambda_c \, P_\Gamma\left(\lambda_c | n_c k, R(1+w_c)\right) \right) \sum_{n'=0}^\infty P_{\mathrm{P}}\left(n'|\lambda\right) \left(1 - \frac{1}{n'}\right) \frac{\lambda_a \lambda_b}{\lambda^2} \\
&\approx \int_0^\infty \left(\prod_{c=1}^m d\lambda_c \, P_\Gamma\left(\lambda_c | n_c k, R\right) \right) \left[1 - \sum_{d=1}^m k w_d \left(n_d - \frac{\lambda_d}{R} \right) \right] \left(1 - \frac{1}{\lambda}\right) \frac{\lambda_a \lambda_b}{\lambda^2} \\
&= \int_0^\infty d\lambda \, P_\Gamma\left(\lambda | nk, R\right) \left(\prod_{c=1}^{m-1} \int dq_c \right) P_{\mathrm{D}}\left(\boldsymbol{q}|\boldsymbol{n}k\right) \left[1 - \sum_{d=1}^m k w_d \left(n_d - \frac{\lambda q_d}{R} \right) \right] \left(1 - \frac{1}{\lambda}\right) q_a q_b \,.
\end{aligned}$$

Simplifying the expressions above is tedious but straightforward. The following results are helpful:

$$\begin{aligned}
\int_0^\infty d\lambda \, P_\Gamma\left(\lambda | nk, R\right) \lambda &= nR, \\
\int_0^\infty d\lambda \, P_\Gamma\left(\lambda | nk, R\right) \frac{1}{\lambda} &= \frac{k/R}{nk-1}, \\
\left(\prod_{c=1}^{m-1} \int dq_c \right) P_{\mathrm{D}}\left(\boldsymbol{q}|\boldsymbol{n}k\right) q_a q_b &= \frac{nk}{nk+1} y_a y_b, \\
\left(\prod_{b=1}^{m-1} \int dq_b \right) P_{\mathrm{D}}\left(\boldsymbol{q}|\boldsymbol{n}k\right) q_a^2 &= y_a^2 + \frac{y_a(1-y_a)}{nk+1} = \frac{nk}{nk+1} y_a^2 + \frac{1}{nk+1} y_a, \\
\left(\prod_{c=1}^{m-1} \int dq_c \right) P_{\mathrm{D}}\left(\boldsymbol{q}|\boldsymbol{n}k\right) q_a^2 q_b &= \left(y_a^2 + \frac{y_a(1-y_a)}{nk+1} \right) \frac{nk}{nk+2} y_b, \\
\left(\prod_{b=1}^{m-1} \int dq_b \right) P_{\mathrm{D}}\left(\boldsymbol{q}|\boldsymbol{n}k\right) q_a^3 &= \left(y_a^2 + \frac{y_a(1-y_a)}{nk+1} \right) \frac{nky_a + 2}{nk+2}.
\end{aligned}$$

Here we have frequently used $n_a = ny_a$ to simplify expressions.

With the above results, simplifying expressions for the second moments, we finally find

$$\left\langle (\Delta y_a)^2 \right\rangle = \left[\frac{1}{nk+1} + \frac{nk}{nk+1} \frac{k/R}{nk-1} \right] y_a \left(1 - y_a\right) + \mathcal{O}\left(1/n^2\right),$$

and

$$\langle \Delta y_a \Delta y_b \rangle = -\left[\frac{1}{nk+1} + \frac{nk}{nk+1} \frac{k/R}{nk-1} \right] y_a y_b + \mathcal{O}\left(1/n^2\right),$$

where we have assumed that the w_a are $\mathcal{O}(1/n)$, as in the Wright-Fisher model with weak selection. We have thus found that the first and second moments of frequency changes in our multi-variant epidemiological model have the same frequency dependence as those in the multispecies Wright-Fisher model, but with different scaling. The first moment ('drift') is multiplied by a factor of $nk/(nk+1)$, and the second moment ('diffusion') by

$$\frac{1}{nk+1} + \frac{nk}{nk+1} \frac{k/R}{nk-1}.$$

These prefactors match with the Wright-Fisher model exactly when $k \to \infty$ (i.e., a pure Poisson distribution for the number of new cases per infected individual) and $R = 1$.

2.4. Correction to the first moment due to travel

Travel of infected individuals can change the total number of infected individuals in a region and will thus lead to a correction in the first and second moments that have been calculated above. Call $n_{a,\text{in}} - n_{a,\text{out}} = \delta n_a$. There will be a first order correction to the first moment due to fact that the number of individuals of variant a in the next generation is now $n'_a + \delta n_a$, where now n'_a represents the number of individuals of variant a in the next generation that don't come from travel. In addition, the total population in the next generation will be $n' + \sum_b \delta n_b$, where $n' = \sum_b n'_b$. Therefore, in calculating the first moment we will now have

$$\langle y'_a \rangle = \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_b)\right) \right) \sum_{n'=0}^{\infty} P_\text{P}\left(n'|\lambda\right) \sum_{\boldsymbol{n}' : \sum_{c=1}^{m} n'_c = n'} P_\text{M}\left(\boldsymbol{n}' \middle| n', \frac{\boldsymbol{\lambda}}{\lambda}\right) \frac{n'_a + \delta n_a}{n' + \sum_d \delta n_d} \, .$$

If we allow that the total inflow and outflow of a specific variant, δn_a, and of all variants together, $\sum_d \delta n_d$, are both much smaller than the number of infected individuals, then the term on the far right can be expanded:

$$\frac{n'_a + \delta n_a}{n' + \sum_d \delta n_d} \approx \left(\frac{n'_a}{n'} - \frac{n'_a \sum_d \delta n_d}{(n')^2} + \frac{\delta n_a}{n'} - \frac{\delta n_a \sum_d \delta n_d}{(n')^2} \right) .$$

The first term simply reproduces the first order moment without travel. The last two terms lead to the correction

$$\frac{\delta y_a}{R} \frac{nk}{nk-1} \left[1 - \sum_d w_d x_d - \frac{\sum_d \delta y_d}{R} \frac{nk}{nk-2} \left(1 - 2 \sum_e w_e y_e \right) \right] ,$$

where $\delta y_a = \delta n_a / n$. Finally, for the second term, we have

$$\left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_b)\right) \right) \sum_{n'=0}^{\infty} P_\text{P}\left(n'|\lambda\right) \frac{\lambda_a}{\lambda} \left(-\frac{\sum_c \delta n_c}{n'} \right) .$$

Ultimately, this produces the additional correction

$$-\frac{y_a \sum_b \delta y_b}{R} \left[\frac{nk}{nk-1} + \frac{nk}{nk+1} \left(w_a - 2 \frac{nk}{nk-1} \sum_d w_d y_d \right) \right] .$$

Remembering that both δy_d and w_d are small, the overall first order correction will be

$$\frac{nk}{nk-1} \frac{1}{R} \left(\delta y_a - y_a \sum_b \delta y_b \right) . \tag{S2}$$

2.5. Correction to the second moment due to travel

There will also be a correction to the second moment due to travel, and for the diagonal terms, we will have,

$$\langle (y'_a)^2 \rangle = \left(\prod_{b=1}^{m} \int_0^\infty d\lambda_b \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_b)\right) \right) \sum_{n'=0}^{\infty} P_\text{P}\left(n'|\lambda\right) \sum_{\boldsymbol{n}' : \sum_{c=1}^{m} n'_c = n'} P_\text{M}\left(\boldsymbol{n}' \middle| n', \frac{\boldsymbol{\lambda}}{\lambda}\right) \left(\frac{n'_a + \delta n_a}{n' + \sum_d \delta n_d} \right)^2 .$$

Again assuming that the number of infected individuals traveling is small compared to the total number of infected individuals in a region, we can expand the term on the far right:

$$\left(\frac{n'_a + \delta n_a}{n' + \sum_d \delta n_d} \right)^2 \approx \left(\frac{n'_a}{n'} \right)^2 + \frac{2 n'_a \delta n_a}{(n')^2} - \frac{2 n'_a \sum_d \delta n_d}{(n')^3} .$$

Similarly, we have

$$\langle y'_a y'_b \rangle = \left(\prod_{c=1}^{m} \int_0^\infty d\lambda_c \, P_\Gamma\left(\lambda_b | n_b k, R(1+w_c)\right) \right) \sum_{n'=0}^{\infty} P_\text{P}\left(n'|\lambda\right) \sum_{\boldsymbol{n}' : \sum_{d=1}^{m} n'_d = n'} P_\text{M}\left(\boldsymbol{n}' \middle| n', \frac{\boldsymbol{\lambda}}{\lambda}\right) \frac{\left(n'_a + \delta n_a\right)\left(n'_b + \delta n_b\right)}{\left(n' + \sum_e \delta n_e\right)^2} ,$$

and the term on the far right can be expanded as

$$\frac{\left(n'_a + \delta n_a\right)\left(n'_b + \delta n_b\right)}{\left(n' + \sum_e \delta n_e\right)^2} \approx \frac{n'_a n'_b}{(n')^2} + \frac{n'_a \delta n_b + n'_b \delta n_a}{(n')^2} - \frac{2 n'_a n'_b \sum_e \delta n_e}{(n')^3} .$$

These can be calculated in much the same way as the correction to the first moment, and all of the resulting terms are $\mathcal{O}(1/n^2)$.

2.6. Derivation of the selection coefficient estimator

The derivation in this section closely follows that given in ref.[79]. It is well known that a WF process can be approximated by a continuous-time continuous-frequency diffusion process in the large n limit. In the continuous-time limit the time variable t has units of n generations, with one generation in discrete time taking $\tau = 1/N$ continuous time units. The change in the frequency of variant a in one generation due to travel is δy_a, so the change of frequency in τ generations is $\delta y_a * \tau$. Furthermore, the selection coefficients w_a are assumed to scale with n such that $w_a = \tilde{w}_a/n$, where $\tilde{w}_a$ is a parameter independent of the population size n. In the limit of large population size, our generalized super-spreading model can, like the WF process, be approximated by a diffusion process, where the transition probability density ϕ is the solution to the Fokker-Planck equation

$$\frac{\partial \phi}{\partial t} = \left[-\sum_{a=1}^{M} \frac{\partial}{\partial x_a} \boldsymbol{d}(\boldsymbol{y}(t)) + \sum_{a=1}^{M}\sum_{b=1}^{M} \frac{\partial}{\partial y_a}\frac{\partial}{\partial y_b} C_{ab}(\boldsymbol{y}(t)) \right] \phi \,,$$

where M is the number of distinct genotypes, $\boldsymbol{y}$ is the genotype frequency vector, $\boldsymbol{d}$ is the drift vector, and C is the diffusion matrix. Ignoring recombination and mutation, since these are comparatively small and therefore unlikely to significantly affect estimates of changes in viral transmission (though these can be included and the solution remains tractable), the drift and diffusion have entries given by,

$$\begin{aligned}
\tilde{d}_a(\boldsymbol{y}(t)) &= \lim_{n\to\infty} n \langle \Delta y_a \rangle \\
&= \lim_{n\to\infty} \frac{nk}{nk+1} y_a(t) \left(w_a - \sum_{b=1}^{M} w_b y_b(t) \right) + \frac{nk}{nk-1}\frac{1}{R} \left(\delta y_a \tau - y_a \sum_b \delta y_b \tau \right) \\
&= y_a(t) \left(\tilde{w}_a - \sum_{b=1}^{M} \tilde{w}_b y_b(t) \right) + \frac{1}{R} \left(\delta y_a - y_a \sum_b \delta y_b \right) , \\
C_{ab} &= \frac{1}{2} \lim_{n\to\infty} n \langle \Delta y_a \Delta y_b \rangle \\
&= \frac{1}{2} \left[\frac{1}{k} + \frac{1}{R} \right] \begin{cases} y_a(t)\,(1-y_a(t)) & a = b \\ -y_a(t) y_b(t) & a \neq b \,. \end{cases}
\end{aligned}$$

The Fokker-Planck equation can be converted into a path integral approximation for the transition probability density

$$P\left(\boldsymbol{y}(t+1)|\boldsymbol{y}(t)\right) = \frac{\exp\left\{ -\frac{n}{4} \sum_{a=1}^{M}\sum_{b=1}^{M} \left[y_a(t+1) - y_a(t) - \tilde{d}_a(\boldsymbol{y}(t))\tau \right] \left(C^{-1}(y_a(t)) \right)_{ab} \left[y_b(t+1) - y_b(t) - \tilde{d}_b(\boldsymbol{y}(t))\tau \right] \right\}}{(4\pi)^{M/2} \sqrt{\det\left(C(\boldsymbol{y}(t))\right)}} \,.$$

We write the re-scaled drift vector as $d_a = \tilde{d}_a \tau$. Since we aim to infer selection coefficients for the SNVs, it is more convenient to work with the allele frequencies x_i instead of the genotype frequencies y_a. The allele frequency at site i is given by

$$x_i(t) = \sum_{a=1}^{M} g_i^a y_a(t) \,,$$

where g_i^a is a 1 if there there is a mutant allele at site i on genome a and zero if there is not. Similarly, if the selection coefficient for the genotype a is w_a and the allele level selection coefficient for allele i is s_i, then they are related by:

$$w_a = \sum_{j=1}^{L} g_j^a s_j \,,$$

where L is the length of the genome.

The allele level drift and diffusion terms will be linear combinations of the genotype level drift and diffusion, just as with the frequencies and the selection coefficients. The drift vector for the allele frequencies can be transformed by

$$\begin{aligned}
d_i(\boldsymbol{x}) &= \sum_{a=1}^{M} g_i^a d_a(\boldsymbol{y}) \\
&= \sum_{a=1}^{M} g_i^a y_a(t) \left(w_a - \sum_{b=1}^{M} w_b y_b(t) \right) + g_i^a \frac{1}{R} \left(\delta y_a - y_a \sum_b \delta y_b \right) \\
&= x_i(t)(1-x_i(t)) s_i + \sum_{j=1, j\neq i}^{L} \left(x_{ij}(t) - x_i(t) x_j(t) \right) s_j + \frac{1}{R} \left[\delta x_i - x_i \sum_{b=1}^{M} \frac{\delta n_b}{n} \right] .
\end{aligned}$$

The sum in the last term can be interpreted as the total number of infected individuals added or subtracted to the population due to travel, divided by the population size. This can be used, along with the transition probability density for genomes, in order to find an approximation for the mutant allele transition probability density:

$$P\left(\boldsymbol{x}(t+1)|\boldsymbol{x}(t)\right) = \frac{\exp\left\{-\frac{n}{4}\sum_{i=1}^{L}\sum_{j=1}^{L}\left[x_i(t+1)-x_i(t)-d_i(\boldsymbol{x}(t))\right]\left(C^{-1}(\boldsymbol{x}(t))\right)_{ij}\left[x_j(t+1)-x_j(t)-d_j(\boldsymbol{x}(t))\right]\right\}}{(2\pi/n)^{L/2}\sqrt{\det\left(C(\boldsymbol{x}(t))\right)}},$$

where here the diffusion C is derived similarly to the drift $\boldsymbol{d}$ and has entries

$$C_{ij}(\boldsymbol{x}(t)) = \left[\frac{1}{k}+\frac{1}{R}\right]\left(x_{ij}(t)-x_i(t)x_j(t)\right).$$

A path integral then gives the probability of observing a trajectory of allele frequencies $(\boldsymbol{x}(t_1),\boldsymbol{x}(t_2),...,\boldsymbol{x}(t_K))$, and is given by

$$P\left((\boldsymbol{x}(t))_{t=1}^{K}\,|\boldsymbol{x}(0)\right) = \prod_{t=0}^{K-1} P\left(\boldsymbol{x}(t+1)|\boldsymbol{x}(t)\right).$$

Bayesian analysis can then be used to show that the posterior probability of the selection coefficients $\boldsymbol{s}=(s_1,s_2,...,s_L)$ given an observed frequency path $\boldsymbol{x}(0),\boldsymbol{x}(1),...,\boldsymbol{x}(T)$ is

$$P\left(\boldsymbol{s}\,|\,(\boldsymbol{x}(t))_{t=0}^{K}\right) \propto P\left((\boldsymbol{x}(t))_{t=1}^{K}\,|\boldsymbol{x}(0)\right)\times P_{\text{Prior}}(\boldsymbol{s}), \tag{S3}$$

where we use a Gaussian prior distribution with zero mean and adjustable covariance determined by the parameter γ.

For the inferred coefficients, we take those that maximize the posterior probability. They can be analytically found by a simple application of the Euler-Lagrange equations to equation S3 and are given by

$$\hat{\boldsymbol{s}} = \left[\gamma+\sum_t n\frac{k^2R^2}{(R+k)^2}C(t)\right]^{-1}\left[\sum_t\frac{nkR}{k+R}\left(\Delta\boldsymbol{x}(t)-\frac{1}{R}\left(\delta\boldsymbol{x}-\boldsymbol{x}\sum_{i=1}^{L}\delta x_i\right)\right)\right]. \tag{S4}$$

The second term in the numerator is the correction due to travel of infected individuals into and out of the region, and is given by

$$\tau = -\sum_t\frac{nkR}{k+R}\frac{1}{R}\left(\delta\boldsymbol{x}-\boldsymbol{x}\sum_{a=1}^{M}\frac{\delta n_a}{n}\right).$$

2.7. Extension to multiple regions

In the SARS-CoV-2 pandemic, and in real disease outbreaks in general, there are frequently multiple different outbreaks in different regions that develop largely or entirely independently of one another. In order to find the best estimate for the selection coefficients using the data from multiple regions, the estimator can be generalized to find the maximum a posteriori estimate for the selection coefficients given the time series of allele frequencies in each of the regions. If the probability for a specific path in a specific region r is given by $P\left((\boldsymbol{x}_r(t_r))_{t_r=1}^{T_r}|\boldsymbol{x}_r(0)\right)$, where x_r is the allele frequency vector in region r, then the joint probability of the specific paths in all of the regions is simply the product of the individual region probabilities:

$$P\left((\boldsymbol{x}_1(t_r))_{t=1}^{T_1},...,(\boldsymbol{x}_M(t_r))_{t_r=1}^{T_M}\,|\,\{\boldsymbol{x}_r(0)\}_{r=1}^{M}\right) = \prod_{r=1}^{M}P\left((\boldsymbol{x}_r(t))_{t=1}^{T_r}|\boldsymbol{x}_r(0)\right),$$

where M is the number of different regions. Since this is a product of exponential functions, the log posterior will be the sum of the exponents and the regularization. This can be maximized with respect to the selection coefficient vector $\boldsymbol{s}$ as before, which, dropping the travel term, leads to the estimator:

$$\hat{\boldsymbol{s}} = \left[\gamma I+\sum_r\sum_{t_r}\frac{n_rk_r^2R_r^2}{(k_r+R_r)^2}C_r(t_r)\right]^{-1}\left[\sum_r\sum_{t_r}\frac{k_rn_rR_r}{k_r+R_r}\Delta\boldsymbol{x}_r(t_r)\right]. \tag{S5}$$

2.8. Simplification of the estimator

In real outbreaks the parameters k, R, and n are in general time-varying. In our simulations as well, R and n are time-varying (and k can be constant or time-varying). In order to accurately infer the selection coefficients according to Eq. (S4) or Eq. (S5), it would seem that we need to accurately infer the values of k, R, and N at every point in the time series. In practice, this would be extremely difficult. For general discussion about the effective reproduction number R and the basic reproduction number R_t as well as some attempts to infer this, see refs.[80–84]. In order to get an accurate estimate for k it is necessary to have pervasive contact tracing, so that the negative binomial distribution is well sampled, and there are other difficulties in inferring k as well[85–87]. Lastly, it can be difficult to estimate the number of new infections due to multiple factors, including the difference between the population that gets tested and the population that does not, test result inaccuracies, and delays between symptom onset, testing, and reporting[88,89].

We propose an alternative that lets us avoid these complications. The prefactor $nkR/(R+k)$, multiplies both the numerator and the denominator. Therefore, the only effect of the prefactor is to weight time points more heavily if the population size, the dispersion parameter, or the basic reproduction number, is larger. This makes sense in theory, because a larger n or k implies that there is less noise and the trajectories are more deterministic, while a larger R means that there are more new infections per generation and thus more data to use to infer the selection coefficients. This does hold with perfect information, that is, if all infected individuals are sampled at every time point. However, in practice, finite sampling is the source of significantly more noise than that due to a time-varying population size or dispersion, so weighting the time points based upon n, k, or R in fact leads to worse inference than assuming the parameters are constant in time and thus weighting the time points equally. However, in the special and unrealistic case of perfect sampling, using the actual parameters does lead to better inference than using constant parameters (see **Supplementary Fig. 4**). If the time points are weighted equally, then, provided that the regularization γ is scaled appropriately (and in general it must be determined by separate means, discussed below), the prefactors in the numerator and denominator cancel, and the estimator is independent of n, k, and R. Defining $\bar{C}$ by

$$C = \left[\frac{nkR}{k+R}\right]\bar{C},$$

so that

$$\bar{C}_{ij} = \begin{cases} x_{ij}(t) - x_i(t)x_j(t) & i \neq j \\ x_i(1-x_i) & i = j \end{cases},$$

Eqs. (S4) and (S5) for the selection coefficients become, respectively

$$\hat{\boldsymbol{s}} = \left[\gamma I + \sum_t \bar{C}(t)\right]^{-1} \left[\sum_t \Delta \boldsymbol{x}(t)\right],$$

$$\hat{\boldsymbol{s}} = \left[\gamma I + \sum_r \sum_{t_r} \bar{C}_r(t_r)\right]^{-1} \left[\sum_r \sum_{t_r} \Delta \boldsymbol{x}_r(t_r)\right],$$

which are the same as the MPL estimators for the Wright-Fisher model except that we have ignored the mutation term because the mutation rate for SARS-CoV-2 is small[79].

3. Simulations

We tested the inference using simulations of disease spread. We used two different kinds of simulations. The first is the super-spreader simulation based on the model described above, which is an analog of the Wright-Fisher model where the sampling distribution for the number of new infections per infected individual is drawn from a negative binomial distribution instead of a pure Poisson distribution The second is a standard deterministic susceptible-infected-recovered (SIR) model that has been adapted in order to include multiple variants of the virus with different transmission rates, described in Methods.

3.1. Description of simulations

We simulated disease spread as a branching process in which the number of individuals infected per currently infected individual is drawn from a negative binomial distribution whose shape is determined by the basic reproduction number R_0 (or the reproduction number, R, in a population that is not totally susceptible) and the dispersion parameter k. Because we sample in this way, the population size is not constant. However, if the population size is too small, then the population is extremely likely to die off stochastically, and if the population size is too large, then sampling from the negative binomial becomes too computationally expensive. In order to avoid both of these problems, once the population size is large enough R is adaptively adjusted so that the average reproduction number for the entire population will remain near 1, and the population size will oscillate around a fixed value. An explicit time-varying population size can also be used as input, and R will be adaptively

adjusted to remain near the given curve. Constant values can be used for the dispersion k or k can vary as a function of time, perhaps representing different degrees of social distancing or lockdown measures at different times. Since different interventions implemented to prevent the spread of disease would likely affect the shape of the distribution of the number of individuals infected by a single infected individual, time-varying values for k and R can be used to reflect these effects. We also implement travel of specified variants into or out of the population over time.

3.2. Inference

The simulations are run for a number of generations and genomes are sampled from the population of infected individuals at different times using a multinomial sampling distribution. This sampled time series is then used to infer the selection coefficients using Eq. (S4). Alternatively, multiple simulations can be run and the joint inference of the selection coefficients can be made using Eq. (S5). We find that, given good enough sampling, a long enough time series, and sampling that occurs at a sufficient number of times, the selection coefficients can be inferred very accurately (**Fig. 1**). The quality of inference is significantly improved if multiple simulations are combined and if mutated sites show up in more than one of the simulations, even under less than ideal sampling conditions. Beneficial coefficients are typically inferred more accurately than deleterious ones, likely because deleterious SNVs frequently die off and therefore there is less data to use for inference.

The inference is robust to shortening the time-series or lowering the number of samples taken per generation, though obviously if either of these conditions is too extreme (or worse, both), the inference starts to break down. The negative effects of a short time-series or poor sampling can be somewhat made up for by using multiple simulations, which is analogous to using data from outbreaks in multiple regions. In addition, the diffusion approximation is only valid in the large n limit. However, we tested the inference for small population sizes and found that inference is accurate even if the population of newly infected individuals per serial interval is as low as a few hundred (**Fig. 1**).

It is reasonable to expect that in a real outbreak there will be some travel of infected individuals into and out of the population. This can affect the estimation of the selection coefficients if the travel term, Eq. (S2), is ignored. As a simplified example, imagine there is a steady influx of a variant that has a SNV that is entirely neutral, and little to no outflow of this variant. In this case the selection coefficient for the SNV that migrates into the population will likely be overestimated because the frequency of the SNV will in general be increasing, even though this increase in frequency is not due to a selective advantage. Similarly, if there is an excess of outflow of a certain neutral SNV, then the fitness for this SNV will in general be underestimated. Testing this with simulations, we found that modest influx of a neutral SNV over a long time (25 importations per serial interval compared to $n = 10^4$ local transmissions, continuing for 100 serial intervals) produces a small but detectable bias in the inferred selection coefficient, which can be corrected by including the travel term and using the true flux of variants for δn_a (**Supplementary Fig. 8**). More generally, corrections due to travel should become significant when the term $\tau_{\rm int}$ becomes large compared to observed changes in SNV frequencies.

Supplementary References

76. Diekmann, O. & Heesterbeek, J. A. P. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, vol. 5 (John Wiley & Sons, 2000).
77. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
78. Hogg, R. V., McKean, J. & Craig, A. T. *Introduction to mathematical statistics* (Pearson Education, 2005).
79. Sohail, M. S., Louie, R. H. Y., McKay, M. R. & Barton, J. P. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature Biotechnology* **39**, 472–479 (2021).
80. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International journal of infectious diseases* **92**, 214–217 (2020).
81. Systrom, K., Vladek, T. & Krieger, M. Model powering rt.live. https://github.com/rtcovidlive/covid-model (2020).
82. Dietz, K. The estimation of the basic reproduction number for infectious diseases. *Statistical methods in medical research* **2**, 23–41 (1993).
83. D'Arienzo, M. & Coniglio, A. Assessment of the SARS-CoV-2 basic reproduction number, R0, based on the early phase of COVID-19 outbreak in Italy. *Biosafety and Health* **2**, 57–59 (2020).
84. Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T. & Jacobsen, K. H. Complexity of the basic reproduction number (R0). *Emerging Infectious Diseases* **25**, 1–4 (2019).
85. Clark, S. J. & Perry, J. N. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics* 309–316 (1989).
86. Saha, K. & Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* **61**, 179–185 (2005).
87. Hilbe, J. M. *Negative binomial regression* (Cambridge University Press, 2011).
88. Miller, A. C. *et al.* Statistical deconvolution for inference of infection time series. *medRxiv* 2020.10.16.20212753 (2020).
89. Manski, C. F. & Molinari, F. Estimating the COVID-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics* **220**, 181–192 (2021).