Title

Early detection of highly transmissible viral variants using phylogenomics

Permalink

https://escholarship.org/uc/item/3xs3b1gc

Journal

Science Advances, 10(33)

ISSN

2375-2548

Authors

May, Michael R
Rannala, Bruce

Publication Date

2024-08-16

DOI

10.1126/sciadv.adk7623

Peer reviewed

EPIDEMIOLOGY

Early detection of highly transmissible viral variants using phylogenomics

Michael R. May* **and Bruce Rannala**

As demonstrated by the SARS-CoV-2 pandemic, the emergence of novel viral strains with increased transmission rates poses a serious threat to global health. Statistical models of genome sequence evolution may provide a critical tool for early detection of these strains. Using a novel stochastic model that links transmission rates to the entire viral genome sequence, we study the utility of phylogenetic methods that use a phylogenetic tree relating viral samples versus count-based methods that use case counts of variants over time exclusively to detect increased transmission rates and identify candidate causative mutations. We find that phylogenies in particular can detect novel transmission-enhancing variants very soon after their origin and may facilitate the development of early detection systems for outbreak surveillance.

INTRODUCTION

The continuous emergence of novel genomic variants with the potential for increased transmissibility, virulence, and other traits is a universal feature of viral pandemic and endemic diseases (*1*). The primary measure of transmissibility, R_0, the basic reproductive rate, may be altered by many intrinsic and extrinsic factors (*2*). For example, D614G spike mutations in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) appear to enhance viral replication by increasing infectivity and stability of virions (*3*), epitope mutations may lead to immune escape increasing the population of susceptibles (*4*), and mutations may cause increased viral loads in individuals due to enhanced replication increasing infectivity (*5*) or may cause a longer period of host infectivity (*6*). The theoretical models we explore in this study do not specify a mechanism for enhancing R_0 but instead consider the most general case by assuming only that a change in R_0 arises from a change in genome sequence.

Evidence for a causal role of one or more specific mutations in increasing the R_0 of a strain (cluster of mutations) may come from a variety of sources, for example, epidemiological studies of case counts over time of individuals infected with particular strains (*7*); experimental studies of the transmissibility of different strains in an animal model system (*8*, *9*); biophysical predictions—for example, predictions of binding affinities of mutant SARS-CoV-2 receptor binding domain (RBD) sequences to the ACE2 receptor based on molecular modeling (*10*); phylogenetic analysis of the expansions of particular strains through time (*11*, *12*); etc. Although demonstrating enhanced transmission in an experimental animal system is often taken as the gold standard, no one source of evidence is definitive. Large differences in transmissibility may exist between human and model animal populations, casting doubt on the relevance of evidence from the model. On the other hand, epidemiological studies can directly infer enhanced transmission in human populations, but often sufficient data are available only after a variant has already become widespread (*13*).

The availability of genome sequences for infectious disease organisms, such as SARS-CoV-2, allows emerging mutations to be identified and monitored to assess their potential impacts (*14*, *15*). Phylogenetic information defining evolutionary relationships among strains is also available from genome sequences, and such data have been used with influenza, SARS-CoV-2, and other pathogens to predict the likely dominant variants (strains) of future pandemics (*16*, *17*). Such methods implicitly assume that phylogenetic information provides additional predictive power beyond that available from simply monitoring changing frequencies of variants among infected individuals. However, surprisingly little is known about the relative power of phylogenies versus frequencies for identifying variants destined to become widespread.

Most genomic variants do not influence transmissibility, and methods are urgently needed to identify (from among the hundreds or thousands of variants that do not influence transmissibility) the small subset of variants that do. Phylogenetic information from genome sequences could potentially be used for de novo identification of sites in viral genomes influencing transmissibility, but theoretical studies are needed to understand the potential of such approaches. The statistical problem of identifying transmission-enhancing sites is very similar to the challenging problem of identifying so-called "driver" mutations in genomes of cancer cells (*18*), although the smaller genomes of viruses greatly reduce the number of candidate mutations. Theoretical studies are needed both to demonstrate that such inferences are possible and to determine their prospective power.

Here, we explore the information available from viral genomic datasets for early detection of transmission-enhancing mutations (TEMs) that increase R_0, as well as for identification of the specific sites with TEMs in genomes. We focus on two types of data: (i) counts of viral variants sampled over time and (ii) molecular phylogenies relating viral samples. To study information content, we develop an explicit statistical model of how mutation events influence transmissibility. Critically, this model allows us to derive tractable probability distributions for both count and phylogenetic datasets.

By simulating datasets based on realistic epidemiological and mutation parameters for SARS-CoV-2, we find that phylogenetic data provide strong evidence supporting TEM status of variants days or weeks before case counts alone. This suggests that phylogenetic methods can identify emerging variants of concern (VOCs) sooner than methods using case counts. If epidemiological approaches for identifying VOCs based on case counts provide a lagging, or post hoc, indicator for the emergence of a new more transmissible variant, it is possible that phylogenetic analyses might allow candidate TEMs to be identified before they are widespread

Department of Evolution and Ecology, University of California Davis, Davis, CA, USA.
*Corresponding author. Email: mikeryanmay@gmail.com

enough for traditional epidemiological methods to estimate R_0, thus providing a much-needed leading indicator for an emerging variant of interest.

RESULTS

In the simulations that follow, we use a novel statistical framework (see Methods) to address two related statistical questions: (i) how early can a novel TEM be detected using either phylogenetic or count data, and (ii) how much information is there in phylogenetic data to identify particular TEM sites and nucleotides.

Phylogenetic data improve early detection of TEMs

Early detection of variants with increased transmission rates is critical for mitigating outbreaks before they can become established. We used simulation to compare the ability of methods using either phylogenies or genotype counts to detect mutations that confer increased transmission in the early phase of an outbreak, assuming the TEM site was known a priori.

We simulated the first few weeks of outbreaks of pathogen variants bearing a single mutation that increased the transmission rate by a factor δ (the "effect size"). We varied the effect size over a range of plausible values of R_0, from a 25% increase up to a doubling of R_0 (see section S4). We then computed the posterior probability of a neutral model (where we assume that there was no TEM site) and each possible TEM model (corresponding to different TEM nucleotides at the known TEM site).

While both methods perform well for large effect sizes (Fig. 1, bottom row/left column, "both"), the tree method performs much better when the effect size is more modest (Fig. 1, left column, "tree only"); notably, the count method never succeeds when the tree method fails (Fig. 1, left column, "count only"). In cases when both methods succeed, the tree method detects increased R_0 several days earlier than the count method (2.09 to 3.38 days on average for different effect sizes, δ), with detection using phylogenies occurring at least 9 days earlier in 5% of cases (with only a 25% increase of R_0) (Fig. 1, middle column). The absolute time to detection depends on the effect size (Fig. 1, right column), ranging from over 2 weeks for modest effect sizes to about 1 week for the more extreme ones; as before, the tree method (orange lines) detects increased R_0 earlier than the count method (blue lines). (We provide more details and results for this simulation in section S4.)

De novo identification of TEMs

In addition to quantitatively outperforming the count method when the TEM variant is hypothesized a priori, the tree method also provides a practical approach for identifying which specific sites in the genome confer increased transmission rates (i.e., when TEM sites are not know a priori). The ability to scan a sample of sequenced viral genomes for evidence of variants with increased R_0 is perhaps critical in identifying VOCs before other information about possible effects of variants (experimental studies, etc.) are available.

We performed a simulation to characterize the ability of the tree method to correctly reject a neutral model (where no site in the genome confers increased transmission) and to identify the true site—and the specific nucleotide state that confers enhanced transmission—from among the entire genome. We simulated outbreaks with a single TEM site over a range of effect sizes (as described previously) and with sample sizes ranging from 100 to 1600 viral samples. For each of these datasets, we also simulated the evolution of the neutral genome, comprising 29,999 sites (a genome size similar to SARS-CoV-2), using plausible values of the per-site mutation rate for SARS-CoV-2 (*17*, *19*). We then computed the posterior probability of the neutral model and each single-site TEM model (where a given TEM model corresponds to a particular TEM site/nucleotide combination).

Overall, the ability to decisively reject the neutral model increases as the effect size increases (Fig. 2, columns, blue lines) and as the number of samples increases (Fig. 2, *c*, *x* axis). Beyond the ability to reject the neutral model, the tree method demonstrates generally good power to identify the true model. The frequency with which the true site/nucleotide combination has the highest support increases as a function of effect size, δ, and sample size, *c* (Fig. 2, orange lines), exceeding 95% for TEMs of large effect and large sample sizes; in many of these cases, the true model is decisively supported (i.e., it has a posterior probability greater than 95%; Fig. 2, red lines). (We provide more details and results for this simulation in section S4.)

DISCUSSION

We have developed a novel modeling framework to understand the theoretical behavior of methods for inferring changes in viral transmission rates caused by mutation events. The model incorporates several simplifying assumptions to make this study analytically and computationally tractable. For example, we have assumed that changes in transmission rates are driven by a single point mutation event; of course, multiple sites in the genome may confer increased transmission rates, and some transmission-enhancing variants may involve epistatic interactions among multiple sites, or recombination (*20*, *21*). We have also assumed that patients do not recover (at least before the end of the monitoring date) and that viral sampling is stochastic and uniform; in reality, patients will recover over the course of days or weeks, and sampling effort naturally varies over time and space. While these assumptions may seem quite unrealistic, they may approximate the true process over short temporal and spatial scales, when multiple mutation and recombination events are unlikely, patient recovery is long relative to the monitoring period, and sampling effort is relatively homogeneous. In particular, these assumptions may be quite reasonable during the early stage of an epidemic or outbreak (*22*), which is the focus of our study. Nonetheless, state-dependent diversification models (*23*) could be developed, which would provide the necessary mathematical and computational basis for incorporating more realistic models of mutation, recovery, and sampling; our theoretical exploration suggests that such efforts are worthwhile, as viral genomes appear to contain substantial information about TEM events. Our results also assume that the true phylogeny is known, and therefore represent a best-case scenario for the advantages of phylogenetic methods over count methods; however, the count model corresponds to maximal phylogenetic uncertainty, so unbiased phylogenetic estimates should always increase power to identify TEMs. Nonetheless, biased phylogenetic estimates may lead to biased inferences about TEMs (*24*).

Our study demonstrates the theoretical and practical utility of phylogeny-based approaches for identifying emerging transmission-enhancing variants, which is a critical component of combating viral outbreaks. While experimental methods are currently the gold standard for identifying the mechanisms underlying changes in R_0 caused by mutations (*9*), these approaches are expensive and time

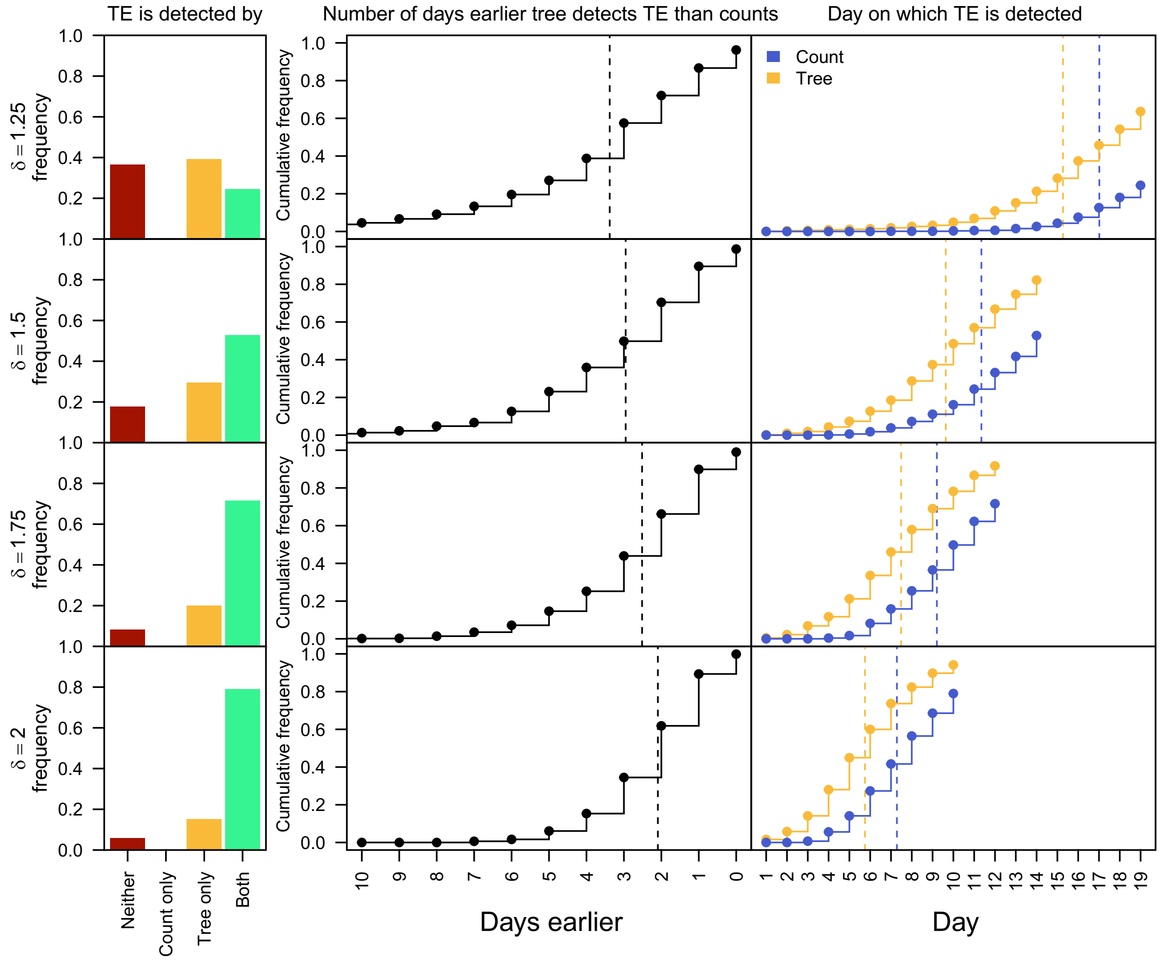

Fig. 1. Phylogenies outperform counts at early detection. We simulated pathogen outbreaks where a single-nucleotide change increased the transmission rate by a factor δ compared to the ancestral variant. We simulated each outbreak for a fixed number of days and collected both phylogenetic and count datasets at daily increments since the origin of the variant (10 to 19 days, depending on δ). We then compared the fit of a neutral model (the novel variant has the same transmission rate as the ancestral variant) against the true model (the mutation confers an increased transmission rate) for each pair of datasets. We measured the frequency with which the phylogenetic and count methods detected increased transmission (i.e., the true model had a posterior probability ≥95%) on at least 1 day of the outbreak (left column) as a function of the effect size (δ, rows). The count method never succeeds unless the tree method does as well (left column, second bar), while the tree method often succeeds where the count method fails (left column, third bar). If both methods succeeded, we computed (i) the distribution (dots/lines) and mean (dashed lines) of the number of days earlier the tree method succeeded than the count method (center column), and (ii) and the distribution/mean for the day on which each method succeeded (right column). Overall, the phylogenetic method detects increased transmission several days earlier than the count method.

consuming, and are difficult (if not impossible) to apply continuously and on short timescales. By contrast, mechanism-agnostic approaches—based on the frequency of genomic variants over time, or the phylogenetic relationship among variants—may provide earlier detection of variants with increased transmission rates before they become outbreaks. They may also identify specific variants that can then become targets of future experimental work.

We have demonstrated that, in theory, phylogeny-based approaches outperform frequency-based ones not only quantitively—they can detect an increased transmission rate for a given variant up to a week earlier, given realistic parameters for the SARS-CoV-2 pandemic—but also qualitatively: In contrast to frequency-based approaches, phylogenies allow us to detect increasing transmission rates when the variant is unknown a priori, even with relatively small sample sizes (on the order of hundreds or thousands of samples). These results support strategies for continuous sampling and genome sequencing of endemic viruses to monitor for emerging VOCs (*25*). While the model we have presented here is relatively simple and analytically tractable, more realistic models including host recovery, density dependence, and temporal and geographic variation in transmission-enhancing effects will undoubtedly entail computationally expensive numerical approximations (*26*, *27*). Nonetheless, we are optimistic that theoretical and computational advances in phylogenetic approaches can lead to the development of early detection systems for monitoring and predicting epidemics that will be of substantial value to the epidemiological community, and the world at large.

Methods

We developed a novel birth-death process to explore the information content available in genomic sequence datasets for answering key epidemiological questions. Birth-death processes have been

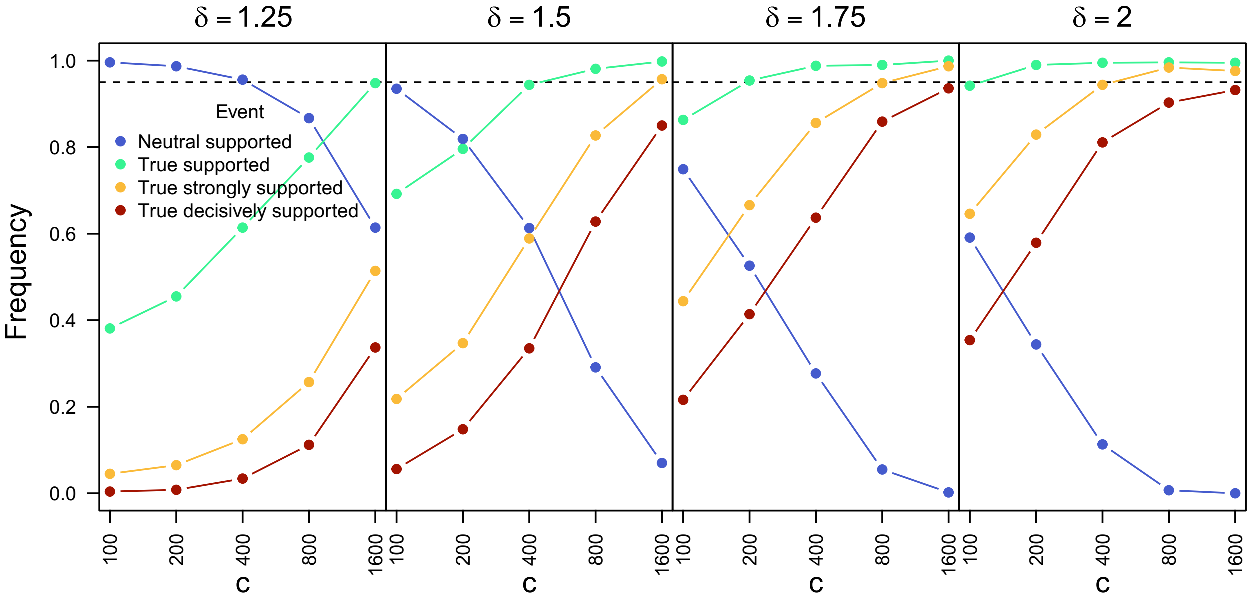

Fig. 2. De novo identification of TEMs. We simulated pathogen outbreaks in which a single-nucleotide change increased the transmission rate by a factor δ. We then computed the support for a neutral model (where all variants have the same transmission rate) and all possible transmission-enhancing models (each corresponding to a particular genomic site/nucleotide combination). A model is supported if it has a posterior probability of at least 5%, strongly supported if it has a posterior probability higher than any other model, and decisively supported if it has a posterior probability greater than 95%. Across simulated datasets, the frequency with which the neutral model is supported (blue) decreases as a function of both the effect size (δ, columns) and the number of samples (*c*, *x* axis). Conversely, the frequencies with which the true model is supported (green), strongly supported (yellow), or decisively supported increase as a function of the effect size and number of samples.

widely used as models of pathogen transmission for serially sampled data (*28*) to understand geographic variation in the transmission rates (*27*), density dependence (*29*), and variation in host susceptibility (*30*), among other important phenomena. Here, we extend the linear birth-death process of epidemic transmission dynamics to allow mutation events to change the transmission rate (see section S1). Under this process, individual viral infections with genome sequence $\boldsymbol{x}$ infect new hosts ("transmit") at rate $\lambda(\boldsymbol{x})$, mutate to a new sequence at rate ν per site, and are sampled (and removed from circulation) at rate ϕ. Viral lineages without any TEMs have a base transmission rate of λ_0; each TEM multiplicatively increases the transmission rate by a factor δ. The TEM model *M* specifies which base positions and nucleotides are TEMs.

This transmission-mutation process begins with a single viral lineage with some ancestral genome sequence and evolves forward in time, transmitting, mutating, and producing samples according to the dynamics described above, until the end of a predefined monitoring period. The outcome of this process is a set of sampled viral sequences, denoted $\mathbf{X}$, the times associated with those samples, denoted *T*, and the phylogenetic tree relating those sequences, denoted Ψ. While the phylogeny is typically unobserved, we assume that researchers can reliably estimate it from the available genomic sequences. We refer to the dataset composed of the sequences and times as the "count dataset" (the number of each genomic variant over time), and the dataset composed of the sequences, times, and phylogeny as the "phylogenetic dataset."

Our model unifies the analysis of count datasets and phylogenetic datasets in a common statistical framework. Specifically, we can compute the likelihood of both a phylogenetic dataset, given a TEM model, $f(\mathbf{X}, T, \Psi | M, \nu, \lambda_0, \delta, \phi)$, and a count dataset, $f(\mathbf{X}, T | M, \nu, \lambda_0, \delta, \phi)$. The count likelihood can be viewed as the phylogenetic likelihood, averaged over all possible phylogenies

$$f\left(\mathbf{X}, T | M, \nu, \lambda_0, \delta, \phi\right) = \int_{\Psi} f\left(\mathbf{X}, T, \Psi | M, \nu, \lambda_0, \delta, \phi\right) d\Psi \quad (1)$$

This relationship suggests that phylogenetic datasets should contain more information about the TEM model than count datasets, because the likelihood is not averaged across all possible phylogenies. Intuitively, the phylogeny provides more information about the age of mutations and transmission events than is available from count data alone; both of these pieces of information should improve our ability to estimate transmission rates of lineages with and without potential TEMs.

In practice, the TEM model *M* is unknown and of critical interest to the researcher. Under this model, the posterior probability of the model *M* for phylogenetic data is

$$f\left(M | \mathbf{X}, T, \Psi, \nu, \lambda_0, \phi\right) \propto f\left(\mathbf{X}, \Psi, T | M, \nu, \lambda_0, \phi\right) f(M)$$

The quantity $f(M)$ is the probability that *M* is the true model before collecting phylogenetic and sequence data, while $f(M | X, \mathbf{X}, T, \Psi, \nu, \lambda_0, \phi)$ is the probability that *M* is the true model, given the phylogenetic and sequence data. The posterior probability of the model therefore provides a basis for deciding which (if any) mutations are TEMs. In this equation, we assume the base transmission rate, mutation rate, and sampling rate are fixed to plausible values based on previous studies, while the TEM effect size, δ, is unknown; in practice, we average over all possible effect sizes in proportion to their posterior probability using numerical integration. The full details underlying this equation can be found in section S2; the corresponding equation for count datasets can be found in section S3.

Supplementary Materials

This PDF file includes:
Sections S1 to S4
Figs. S1 to S7
References

REFERENCES AND NOTES

1. O. G. Pybus, A. Rambaut, Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550 (2009).
2. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, 1991).
3. J. A. Plante, Y. Liu, J. Liu, H. Xia, B. A. Johnson, K. G. Lokugamage, X. Zhang, A. E. Muruato, J. Zou, C. R. Fontes-Garfias, D. Mirchandani, D. Scharton, J. P. Bilello, Z. Ku, Z. An, B. Kalveram, A. N. Freiberg, V. D. Menachery, X. Xie, K. S. Plante, S. C. Weaver, P.-Y. Shi, Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116–121 (2021).
4. G. Dolton, C. Rius, M. S. Hasan, A. Wall, B. Szomolay, E. Behiry, T. Whalley, J. Southgate, A. Fuller, The COVID-19 Genomics UK (COG-UK) consortium, T. Morin, K. Topley, L. R. Tan, P. G. R. Goulder, O. B. Spiller, P. J. Rizkallah, L. C. Jones, T. R. Connor, A. K. Sewell, Emergence of immune escape at dominant SARS-CoV-2 killer T cell epitope. *Cell* **185**, 2936–2951 (2022).
5. T. Mourier, M. Shuaib, S. Hala, S. Mfarrej, F. Alofi, R. Naeem, A. Alsomali, D. Jorgensen, A. K. Subudhi, F. B. Rached, Q. Guan, R. P. Salunke, A. Ooi, L. Esau, O. Douvropoulou, R. Douvropoulou, S. Perumal, H. Zhang, I. Rajan, A. Al-Omari, S. Salih, A. Shamsan, A. Al Mutair, J. Taha, A. Alahmadi, N. Khotani, A. Alhamss, A. Mahmoud, K. Alquthami, A. Dageeg, A. Khogeer, A. M. Hashem, P. Moraga, E. Volz, N. Almontashiri, A. Pain, SARS-CoV-2 genomes from Saudi Arabia implicate nucleocapsid mutations in host response and increased viral load. *Nat. Commun.* **13**, 601 (2022).
6. E. Boehm, I. Kronig, R. A. Neher, I. Eckerle, P. Vetter, L. Kaiser, Novel SARS-CoV-2 variants: The pandemics within the pandemic. *Clin. Microbiol. Infec.* **27**, 1109–1117 (2021).
7. L. M. A. Bettencourt, R. M. Ribeiro, Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PLOS ONE* **3**, e2185 (2008).
8. N. H. Leung, Transmissibility and transmission of respiratory viruses. *Nat. Rev. Microbiol.* **19**, 528–545 (2021).
9. Y.-I. Kim, M. A. B. Casel, Y. K. Choi, Transmissibility and pathogenicity of SARS-CoV-2 variants in animal models. *J. Microbiol.* **60**, 255–267 (2022).
10. S. Kumar, T. S. Thambiraja, K. Karuppanan, G. Subramaniam, Omicron and Delta variant of SARS-CoV-2: A comparative computational study of spike protein. *J. Med. Virol.* **94**, 1641–1649 (2022).
11. R. Viana, S. Moyo, D. G. Amoako, H. Tegally, C. Scheepers, C. L. Althaus, U. J. Anyaneji, P. A. Bester, M. F. Boni, M. Chand, W. T. Choga, R. Colquhoun, M. Davids, K. Deforche, D. Doolabh, L. du Plessis, S. Engelbrecht, J. Everatt, J. Giandhari, M. Giovanetti, D. Hardie, V. Hill, N.-Y. Hsiao, A. Iranzadeh, A. Ismail, C. Joseph, R. Joseph, L. Koopile, S. L. K. Pond, M. U. G. Kraemer, L. Kuate-Lere, O. Laguda-Akingba, O. Lesetedi-Mafoko, R. J. Lessells, S. Lockman, A. G. Lucaci, A. Maharaj, B. Mahlangu, T. Maponga, K. Mahlakwane, Z. Makatini, G. Marais, D. Maruapula, K. Masupu, M. Matshaba, S. Mayaphi, N. Mbhele, M. B. Mbulawa, A. Mendes, K. Mlisana, A. Mnguni, T. Mohale, M. Moir, K. Moruisi, M. Mosepele, G. Motsatsi, M. S. Motswaledi, T. Mphoyakgosi, N. Msomi, P. N. Mwangi, Y. Naidoo, N. Ntuli, M. Nyaga, L. Olubayo, S. Pillay, B. Radibe, Y. Ramphal, U. Ramphal, J. E. San, L. Scott, R. Shapiro, L. Singh, P. Smith-Lawrence, W. Stevens, A. Strydom, K. Subramoney, N. Tebeila, D. Tshiabuila, J. Tsui, S. van Wyk, S. Weaver, C. K. Wibmer, E. Wilkinson, N. Wolter, A. E. Zarebski, B. Zarebski, D. Goedhals, W. Preiser, F. Treurnicht, M. Venter, C. Williamson, O. G. Pybus, J. Bhiman, A. Glass, D. P. Martin, A. Rambaut, S. Gaseitsiwe, A. von Gottberg, T. de Oliveira, Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
12. V. Tchesnokova, H. Kulasekara, L. Larson, V. Bowers, E. Rechkina, D. Kisiela, Y. Sledneva, D. Choudhury, I. Maslova, K. Deng, K. Kutumbaka, H. Geng, C. Fowler, D. Greene, J. Ralston, M. Samadpour, E. Sokurenko, Acquisition of the L452R mutation in the ACE2-binding interface of Spike protein triggers recent massive expansion of SARS-Cov-2 variants. *J. Clin. Microbiol.* **59**, 10–1128 (2021).
13. P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang, K. H. Jacobsen, Complexity of the basic reproduction number (R_0). *Emerg. Infect. Dis.* **25**, 1–4 (2019).
14. R. P. Walensky, H. T. Walke, A. S. Fauci, SARS-CoV-2 variants of concern in the united states-challenges and opportunities. *JAMA* **325**, 1037 (2021).
15. C. H. van Dorp, E. E. Goldberg, N. Hengartner, R. Ke, E. O. Romero-Severson, Estimating the strength of selection for new SARS-CoV-2 variants. *Nat. Commun.* **12**, 7239 (2021).
16. R. A. Neher, T. Bedford, nextflu: Real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics* **31**, 3546–3548 (2015).
17. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford,R. A. Neher, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
18. B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, K. W. Kinzler, Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
19. T. Day, S. Gandon, S. Lion, S. P. Otto, On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* **30**, R849–R857 (2020).
20. A. M. Carabelli, T. P. Peacock, L. G. Thorne, W. T. Harvey, J. Hughes, COVID-19 Genomics UK Consortium, S. J. Peacock, W. S. Barclay, T. I. de Silva, G. J. Towers, D. L. Robertson, SARS-CoV-2 variant biology: Immune escape, transmission and fitness. *Nat. Rev. Microbiol.* **21**, 162–177 (2023).
21. Y. Turakhia, B. Thornlow, A. Hinrichs, J. McBroome, N. Ayala, C. Ye, K. Smith, N. De Maio, D. Haussler, R. Lanfear, R. Corbett-Detig, Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature* **609**, 994–997 (2022).
22. N. T. Bailey, Stochastic birth, death and migration processes for spatially distributed populations. *Biometrika* **55**, 189–198 (1968).
23. D. A. Rasmussen, T. Stadler, Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. *eLife* **8**, e45562 (2019).
24. R. A. Smith, E. L. Ionides, A. A. King, Infectious disease dynamics inferred from genetic data via sequential monte carlo. *Mol. Biol. Evol.* **34**, 2065–2084 (2017).
25. B. B. Oude Munnink, N. Worp, D. F. Nieuwenhuijse, R. S. Sikkema, B. Haagmans, R. A. Fouchier, M. Koopmans, The next phase of SARS-CoV-2 surveillance: Real-time molecular epidemiology. *Nat. Med.* **27**, 1518–1524 (2021).
26. W. P. Maddison, P. E. Midford, S. P. Otto, Estimating a binary character's effect on speciation and extinction. *Syst. Biol.* **56**, 701–710 (2007).
27. D. Kühnert, T. Stadler, T. G. Vaughan, A. J. Drummond, Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Mol. Biol. Evol.* **33**, 2102–2116 (2016).
28. T. Stadler, Sampling-through-time in birth–death trees. *J. Theor. Biol.* **267**, 396–404 (2010).
29. G. E. Leventhal, H. F. Günthard, S. Bonhoeffer, T. Stadler, Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol. Biol. Evol.* **31**, 6–17 (2014).
30. A. A. King, Q. Lin, E. L. Ionides, Markov genealogy processes. *Theor. Popul. Biol.* **143**, 77–91 (2022).
31. W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **115**, 700–721 (1927).
32. D. G. Kendall, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press Berkeley, 1956), vol. 4, pp. 149–165.
33. K. Dietz, The estimation of the basic reproduction number for infectious diseases. *Stat. Methods Med. Res.* **2**, 23–41 (1993).
34. O. Diekmann, J. A. P. Heesterbeek, J. A. Metz, On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365–382 (1990).
35. J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
36. E. Renshaw, Birth, death and migration processes. *Biometrika* **59**, 49–60 (1972).
37. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2022).
38. C. Moler, C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.* **45**, 3–49 (2003).
39. D. Eddelbuettel, R. François, Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40**, 1–18 (2011).
40. D. Bates, D. Eddelbuettel, Fast and elegant numerical linear algebra using the RcppEigen Package. *J. Stat. Softw.* **52**, 1–24 (2013).
41. K. Soetaert, T. Petzoldt, R. W. Setzer, Solving differential equations in R: Package deSolve. *J. Stat. Softw.* **33**, 1–25 (2010).
42. E. Fehlberg, Klassische Runge-Kutta-Formeln fünfter und siebenter ordnung mit schrittweiten-kontrolle. *Comput. Secur.* **4**, 93–106 (1969).
43. A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, N. Daveis, A. Gimma, K. van Zandvoort, H. Gibbs, J. Hellewell, C. I. Jarvis, S. Clifford, B. J. Quilty, N. I. Bosse, S. Abbot, P. Klepac, S. Flasche, Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *Lancet Infect. Dis.* **20**, 553–558 (2020).
44. B. Narasimhan, S. G. Johnson, T. Hahn, A. Bouvier, K. Kiêu, cubature: Adaptive multivariate integration over hypercubes. R package version 2.0.4.5 (2022).

Acknowledgments: We are grateful to B. R. Moore, J. Gao, and A. Thompson for their thoughtful discussion and feedback on this study. **Funding:** This research was supported by

NIH grant R01GM123306-S awarded to B.R. **Author contributions:** M.R.M. and B.R. conceived the research, developed the statistical model, and wrote the manuscript. M.R.M. implemented the methods and conducted the simulation studies. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. This study does not involve any empirical data. All code used in this study is available at the GitHub release (https://github.com/mikeryanmay/YuleSelectionArchive/releases/tag/final) and Zenodo (DOI: 10.5281/zenodo.8193217).

Submitted 8 September 2023
Accepted 9 July 2024
Published 14 August 2024
10.1126/sciadv.adk7623