Early risk-assessment of pathogen genomic variants emergence

Zachary Susswein[1,*], Kaitlyn E. Johnson[1], Robel Kassa[1], Mina Parastaran[1], Vivian Peng[1], Leo Wolansky[1], Samuel V. Scarpino[2], and Ana I. Bento[1,3*]

[1]The Rockefeller Foundation,, New York, New York, USA
[2]Northeastern University, Boston, Massachusetts, USA
[3]Indiana University, Bloomington, Indiana, USA
[*]zsusswein@rockfound.org; abento@rockfound.org

ABSTRACT

Accurate, reliable, and timely estimates of pathogen variant risk are essential for informing effective public health responses to infectious diseases. Despite decades of use for influenza vaccine strain selection and PCR-based molecular diagnostics, data on pathogen variant prevalence and growth advantage has only risen to its current prominence during the SARS-CoV-2 pandemic. However, such data are still often sparse: a novel variant is initially rare or a region has limited sequencing. To ensure real-time estimates of risk are available in these types of data-sparse conditions, we develop a hierarchical modeling approach that estimates variant fitness advantage and prevalence by pooling data across geographic regions. We apply this method to estimate SARS-CoV-2 variant dynamics at the country level and assess its stability with retrospective validation. Our results show that more stable and robust estimates can be obtained even when sequencing data are sparse, as compared to established, single-country estimation approaches. We discuss how this method can inform risk assessment of novel emerging variants and provide situational awareness on currently circulating variants, for a range of pathogens and use-cases.

Introduction

Virus emergence, circulation and diversity can impact outbreak dynamics and control efforts both globally and locally. Over the last two decades, the use of viral genome sequencing has shifted from primarily retrospective research toward near-real-time analyses. As we have seen during the SARS-CoV-2 pandemic, real-time variant characterization using genomic sequencing has the potential to inform public health practice[1–4], enhance disease forecasting models[5] and aid vaccines, therapeutics, and molecular diagnostic assay development[6,7]. An understanding of variant properties (i.e., immune evasion, transmissibility, and therapeutic efficacy) combined with accurate regional estimates of each variants' prevalence can be used to design, apply, and evaluate public health strategies to mitigate transmission and disease[1,3,8–13].

Disease systems as diverse as influenza, HIV, and dengue virus produce spatially and temporally structured competing populations[14–16]. These particularities lead to significant challenges in both the reliability of early characterization of emerging variants and the accessibility of real-time variant prevalence estimates in countries with limited sequencing capacity. However, modeling of the resulting complex dynamics is generally constrained by data sparsity and heterogeneous diagnostic and sequencing capacity across the globe. Robust methods that leverage data from across regions can improve our understanding of strain dynamics in real time for any pathogen and geographic context. Established analytic methods, such as multinomial or logistic regression, can provide reliable estimates of variant growth rates that account

for uncertainty due to changes in sampling intensity over time[17]. However, these approaches often require large numbers of sequences and cannot readily handle geographic heterogeneity in sampling intensity. As a consequence, many existing methods for estimating variant growth rates are inaccurate at small sample sizes[18,19]. These limitations can lead to under- or over-estimates of variant fitness advantage and/or risk posed by a variant.

Because genomic data are often quite sparse, these methods can be challenging to apply in practice. In addition to being limited in numbers, genomic data are often concentrated into a few geographies, both for emerging variants and for many countries as a whole[20]. For example, 90% of publicly available SARS-CoV-2 sequences come from just 10% of countries (Figure 1). These disparities are not only in the raw number of sequences available, but also in the percent of cases sequenced and the turnaround time from sample collection to sequence availability[20]. Because of the concentration of sequencing into a few countries, sample size limitations often make country-specific, regression-based approaches unreliable. We note that this sparsity can arise through a number of different mechanisms, including insufficient local capacity to sequence genomes (i.e., "sequencing capacity") or limited ability to divert existing sequencing capacity from existing use cases toward a pathogen of public health interest (i.e., "sequencing effort"). Established regression-based approaches are therefore most appropriate for monitoring variants in countries or regions with high sequencing capacity and effort and rapid turnaround times.

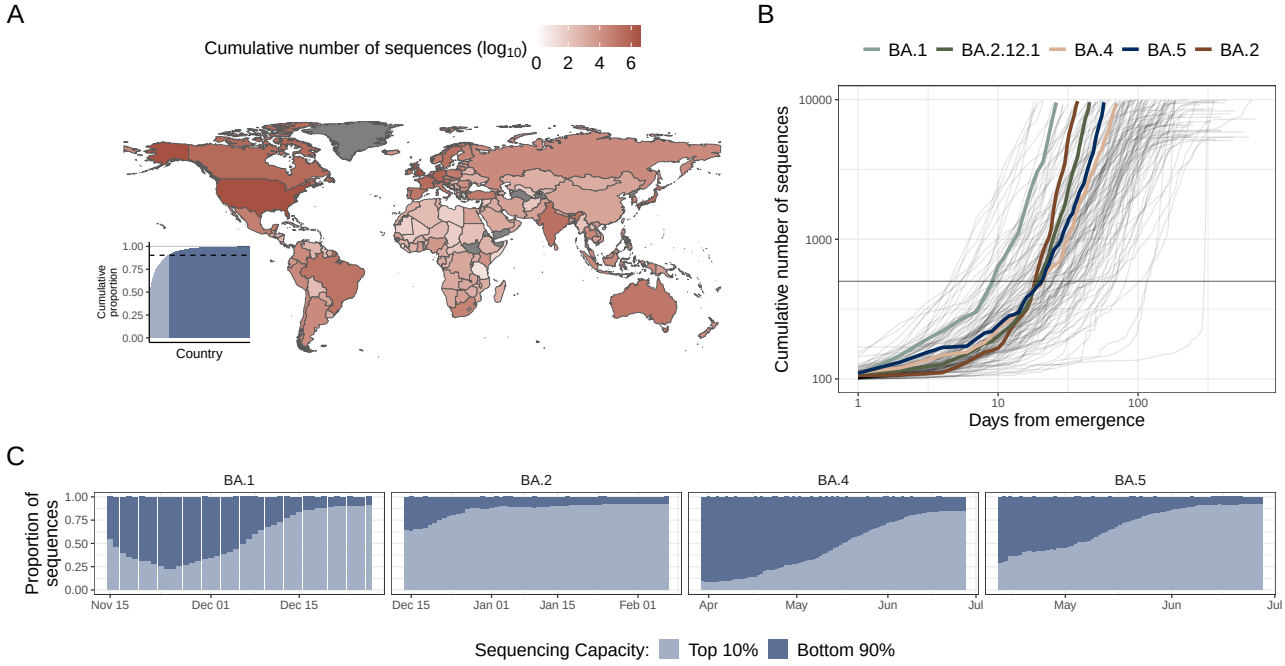

Figure 1. Landscape of SARS-CoV-2 genomic data and early emergent variant dynamics.(A) Shading indicates cumulative number, in log10 scale, of SARS-CoV-2 sequences obtained from GISAID as of July 1st, 2022[21]. (inset) Cumulative proportion of all sequences, with countries ordered by their relative contribution. Light blue indicates countries in the top 10th percentile of contributions, dark blue indicates countries in the bottom 90th percentile of countries . (B) Cumulative number of sequences versus days from variant emergence, with variants of interest which grew rapidly after emergence highlighted in color. Gray horizontal line at 500 sequences is included to highlight the time it took to reach this level for the key variants. (C) Proportion of sequences sampled by the high sequencing capacity countries (light blue) vs the lower sequencing capacity countries (dark blue) over time, starting from after a variant's emergence.

Although the available sample size can be increased by pooling sequences from multiple countries, different populations have different immunological landscapes and seeding dynamics of novel variants[4,22]. Consequently, relative variant fitness can differ substantially across regions. To address this challenge, we developed a hierarchical modeling approach that jointly estimates the trajectories of variants' growth over time. By pooling all the available sequences, it estimates variant properties globally even in regions with limited sequencing efforts and/or capacity (Figure 2). From this model, we identify the key quantities of: i.) within-country prevalence, ii.) within-country growth rate, and iii.) global fitness relative to circulating variants. The method accounts for the observed heterogeneity in variant fitness across geographic regions and sparsity in sequencing to provide more robust estimates of variant fitness shortly after emergence. By using sequences more efficiently, this approach complements investments aimed at increasing global sequencing capacity and/or can offset reductions in sequencing effort (e.g., as happened throughout 2022 with SARS-CoV-2 genomic sequencing[23]).

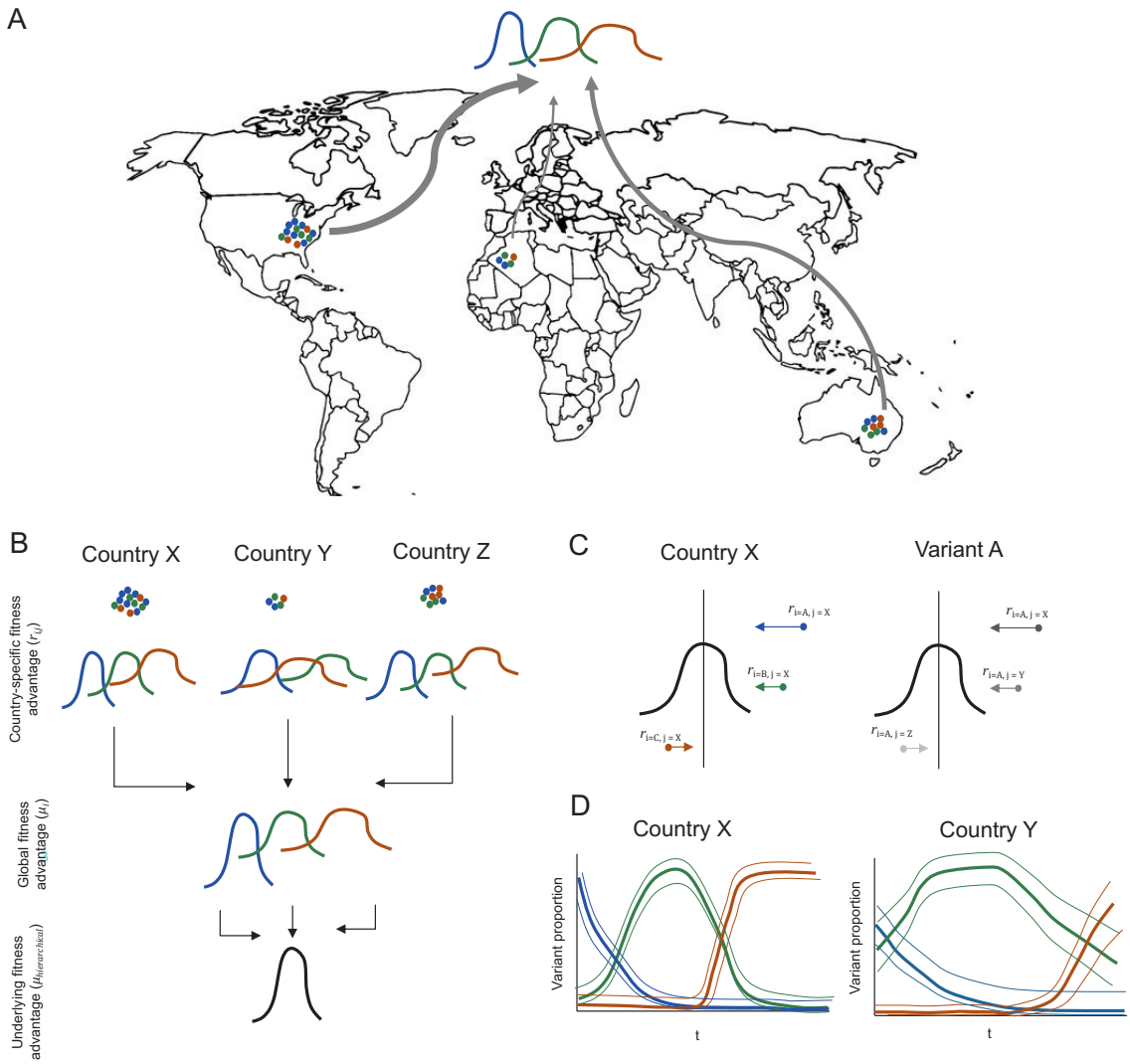

Figure 2. Schematic figure describing the multi-region model. (A) Observed sequences from across the globe are fit to a single Bayesian hierarchical multinomial model, with a relative growth rate parameter ($\beta_{1,i,j}$) that can be transformed into relative variant fitness advantage ($r_{i,j}$). (B) Posterior probability distributions of the country-specific variant fitness advantages with means $r_{i,j}$, global variant fitness advantage

distributions with means μ_i, and single global fitness advantage distribution with mean $\mu_{hierarchical}$. Arrows indicate the hierarchical layers of the model. The model assumes that the country-specific variant fitness advantage ($r_{i,j}$) is drawn from a shared distribution of that variant's global fitness advantage with means μ_i, and likewise that each variant's global fitness advantages are drawn from a shared normal distribution of fitness advantages across all variants with mean $\mu_{hierarchical}$. (C) Schematic illustration of how the mixed effects model structure results in shrinkage towards the mean of country-specific variant fitness advantages. Dots indicate estimates run without the pooling of information, arrows indicate where the estimate lands in the hierarchical structure. (D) The mixed effects structure means that, across variants and across countries, country-specific variant fitness advantages are shrunk towards the mean, and only allowed to deviate significantly if the data strongly supports it. This should result in more stable and robust estimates of variant fitness advantages. Schematic examples of the outputs from the model which include The model generates country-specific estimates of variant proportions over time. Lighter lines indicate credible intervals obtained from drawing from the posterior distributions of model parameters. Countries with fewer sequences (Country Y) will have more uncertainty in the estimated variant proportions. by drawing from the posterior distributions of the model parameters. Countries with more sequences will have more certainty in their variant proportions, and countries with less data will have higher uncertainty.

Results

Global landscape of SARS-CoV-2 genomic surveillance

In Figure 1, we examine the landscape of global SARS-CoV-2 genomic sequencing, identifying systemic variability in sequencing capacity/throughput. The majority of the world's SARS-CoV-2 sequencing has occurred in the United States and United Kingdom, which had contributed 54.9% of all SARS-CoV-2 sequences shared via the GISAID Initative[21] as of July 1, 2022. This disparity extends beyond these two countries — 10% of the countries that have shared sequences with GISAID have produced 90.3% of the SARS-CoV-2 sequences (Figure 1A). These high sequencing-capacity countries are unevenly geographically distributed; with the exception of India, all are members of the Global North. In contrast, most countries in the African continent and the Middle East have submitted only a few thousand SARS-CoV-2 sequences in total, with the notable exceptions of South Africa and Kenya (Figure 1A, inset).

In Figure 1B, we demonstrate a pattern in sequencing effort targeted towards emerging variants, finding that for each variant the rate of sequencing accelerates over time. Early on, when a variant is at low prevalence, the total number of sequences of that variant grows slowly — at a linear rate. In the early stages of variant emergence, the amount of information about the variant is limited, with only a few sequences available and the benefit provided by pooling information is highest. Later on, when that variant begins to spread more widely, we see exponential growth in the total number of sequences. This pattern of an increasing rate of sequencing over time leads to the characteristic elbow shape in the plot.

In Figure 1C, we show that this second phase of rapid sequence collection is largely driven by the concentration of genomic sequencing within a few high capacity countries. Shortly after variant emergence, the majority of the BA.1, BA.4, and BA.5 lineage sequences collected were from outside of these high sequencing capacity countries. However, once these variants reached high sequencing capacity countries, sequences were collected more rapidly, becoming a majority of the sequences collected. This change highlights that genomic surveillance for emerging variants has been largely limited to a few high-resource geographies.

We describe a general method to estimate multi-strain dynamics and relative fitness advantages by partially pooling information across patches (geographic subunits, i.e. countries) and strains (Fig 2). This statistical approach leverages a hierarchical mixed-effects Bayesian framework. The model has two levels of hierarchy. In the first level, country-specific variant fitness advantages are structured such that the fitness advantage of a variant in one location informs the expected fitness advantages of variants in other locations to formalize the assumption that variants' properties in one location are likely to be similar to others (Fig 2A). In the second level, variants' mean fitness advantages, averaged over countries, consist of a shared (hierarchical) normal distribution (Fig 2B). This approach shares information across variants to formalize the ecological assumption that most variants will be similarly fit to their recent ancestors and observing extreme deviations in fitness is uncommon. This assumption leads to shrinkage on extreme fitness advantage estimates for novel or otherwise infrequently observed variants for which we might otherwise overfit to noise in the data (Fig 1C).

Because of this sharing of information, robust estimates can be produced in settings with extremely sparse data. Estimated quantities include region-specific variant proportions over time with associated uncertainty as well as global and region-specific characterization of variant properties. The framework presented here is general enough to be applied to any geographic scale and can be extended to estimate multi-strain pathogen dynamics beyond SARS-CoV-2.

Identifying SARS-CoV-2 variant fitness and global emergence dynamics

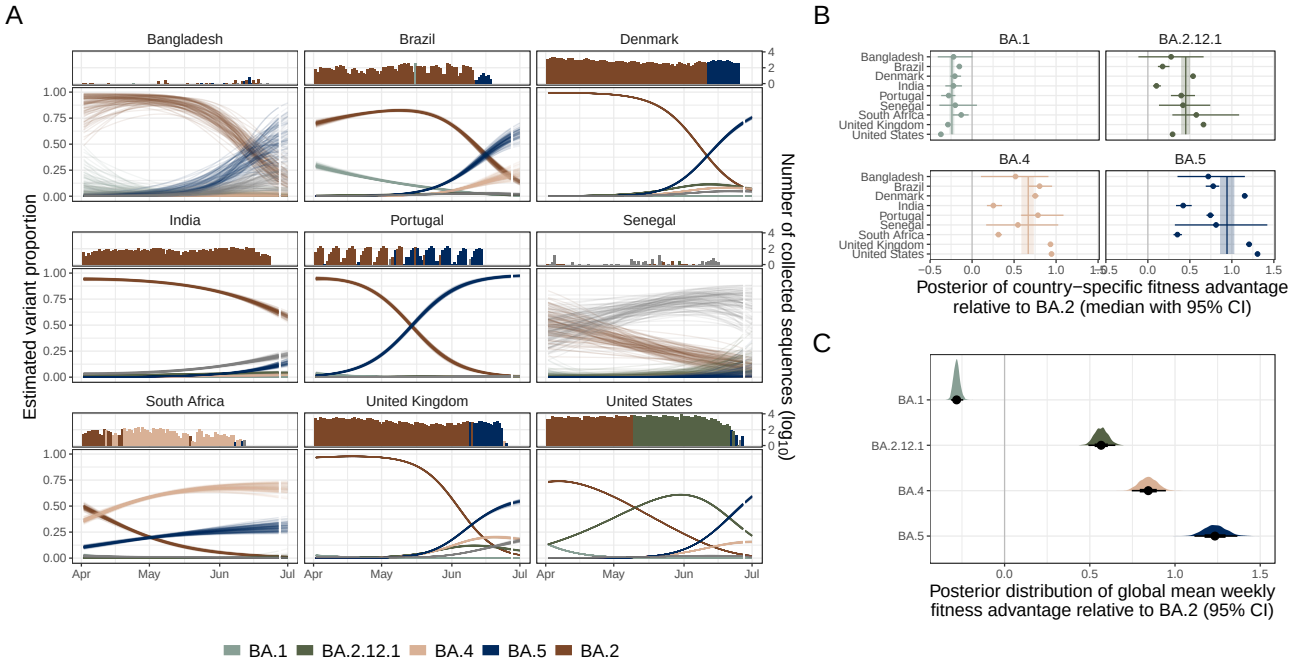

Figure 3. Estimating variant dynamics and fitness advantages(A) Model estimated variant dynamics in a subset of countries. Colors indicate variants, lines represent draws from the marginal posterior distributions of the country-specific estimates.The top panel shows the number of sequences collected over time, colored by the dominant variant at that time as is observed in the data. (B) Country-specific fitness advantages for selected variants (points). Vertical line indicates global estimate of variant fitness advantage. Bars and bands indicate 95% credible intervals. (C) Global posterior fitness advantage distributions for selected variants. Points indicate median, bars indicate 95% credible intervals.

In Figure 3A, we present the estimated SARS-CoV-2 variant proportions alongside the number of collected sequences over time, colored by the most frequently observed variant on each day, for a subset of countries. Global prevalence of the BA.2 lineage declined during the period from April to July 2022,as the BA.4 and BA.5 variants rapidly took over. The invasion dynamics of BA.4 and BA.5 were heterogeneous across countries: both took over in parallel in South Africa, but BA.5 had substantially higher observed growth in Israel and Bangladesh. Notably, in Bangladesh we estimate that BA.5 had reached a higher proportion of cases (56.5% CI: 21.2-83.4%) than in India as of July 1, 2022 (13.8%, CI: 1.1-17.6%) despite their geographic proximity. In contrast, Senegal had not yet experienced substantial BA.5 invasions as of July 1, 2022. Although BA.2 prevalence declined, this change was driven by BA.4 and other lineages rather than by BA.5.

In Figure 3B, we present country-specific variant growth rates, finding high country-specific fitness advantages for BA.4, BA.5, and BA.2.12.1. Country-specific fitness advantages are heterogeneous within variant type: the relative fitnesses of BA.4 and BA.5 are much higher in the United States (BA.5: 1.31 CI: 1.29-1.33, BA.4: 0.95 CI: 0.92-0.97) and the United Kingdom (BA.5: 1.21 CI: 1.17-1.24, BA.4: 0.93 CI:0.91- 0.96) than in India (BA.5: 0.42 CI: 0.34-0.52, BA.4: 0.25 CI: 0.18-0.36) and South Africa (BA.5: 0.35 CI: 0.31- 0.41, BA.4: 0.25 CI: 0.18-0.36). We note that while these estimated fitness advantages are often similar between the United States and United Kingdom (e.g., in the cases of BA.4 and BA.5), this relationship does not always hold true. In the case of BA.2.12.1, estimates of relative fitness advantage in the United States (0.29 CI: 0.29-0.30) are much lower than in the United Kingdom (0.66 CI: 0.64-0.69).

In Figure 3C, we present global relative fitness advantages for multiple variants of concern, finding that BA.5 is meaningfully more fit than BA.4 and BA.2.12.1. These expectations are the means of the Gaussian posterior distributions of country-specific estimates ($\mu_{hierarchical}$) in Figure 2 (i.e., the random effect distribution) and a robust measure of overall variant fitness advantage. Most of the posterior density of the expected relative fitness advantage for BA.5 (0.94 CI: 0.85-1.03%) is higher than that for BA.4 (0.67 CI: 0.59-0.73) which is itself higher than that for BA.2.12.1 (0.45 CI: 0.40 - 0.50) (Figure 3C). A weekly fitness advantage of 100% means that in one week the proportion of total cases caused by BA.5 is expected to double, relative to those of BA.2 (i.e. BA.5 would go from 1% to 2% of cases if the proportion of those due to BA.2 stay constant). All three variants (BA.5, BA.4, and BA.2.12.1) have fitness advantages substantially above 0, indicating that each is more fit than BA.2, while the expected fitness advantage of BA.1 is below 0 (-0.24 CI: -0.27 to -0.21) — indicating that it is less fit than BA.2.

Retrospective validation and comparison

We performed a retrospective validation of the multicountry model estimates from 5 successive reference dates. We compared these estimates to those from single country fits to a Maximum Likelihood Estimation (MLE) multinomial model (referred to as the single country model). Figure 4A depicts the results of this analysis, using BA.5's emergence in Portugal as a case study. Portugal was chosen for display because it had an early seeding of BA.5 prior to its widespread circulation in higher sequencing capacity countries, thus representing a good reliability test for our models ability to provide early, robust estimates of variant fitness. The top panels of Figure 4A depict the sequences each model was fit to, with shading corresponding to the calibration period (i.e., the date the last available sequence from that dataset was collected). We compare the model estimates (both of the past and current variant proportions, and the 21-day forecasted variant proportions) to the observed data as of July 1st, 2022. In Figure S1, we depict the same figure but compared to the data the model was fit to as of that reference date. This data changes over time as sequences from ear-

lier collection dates get submitted to GISAID and "back-fill" observed variant proportions on a specific date.

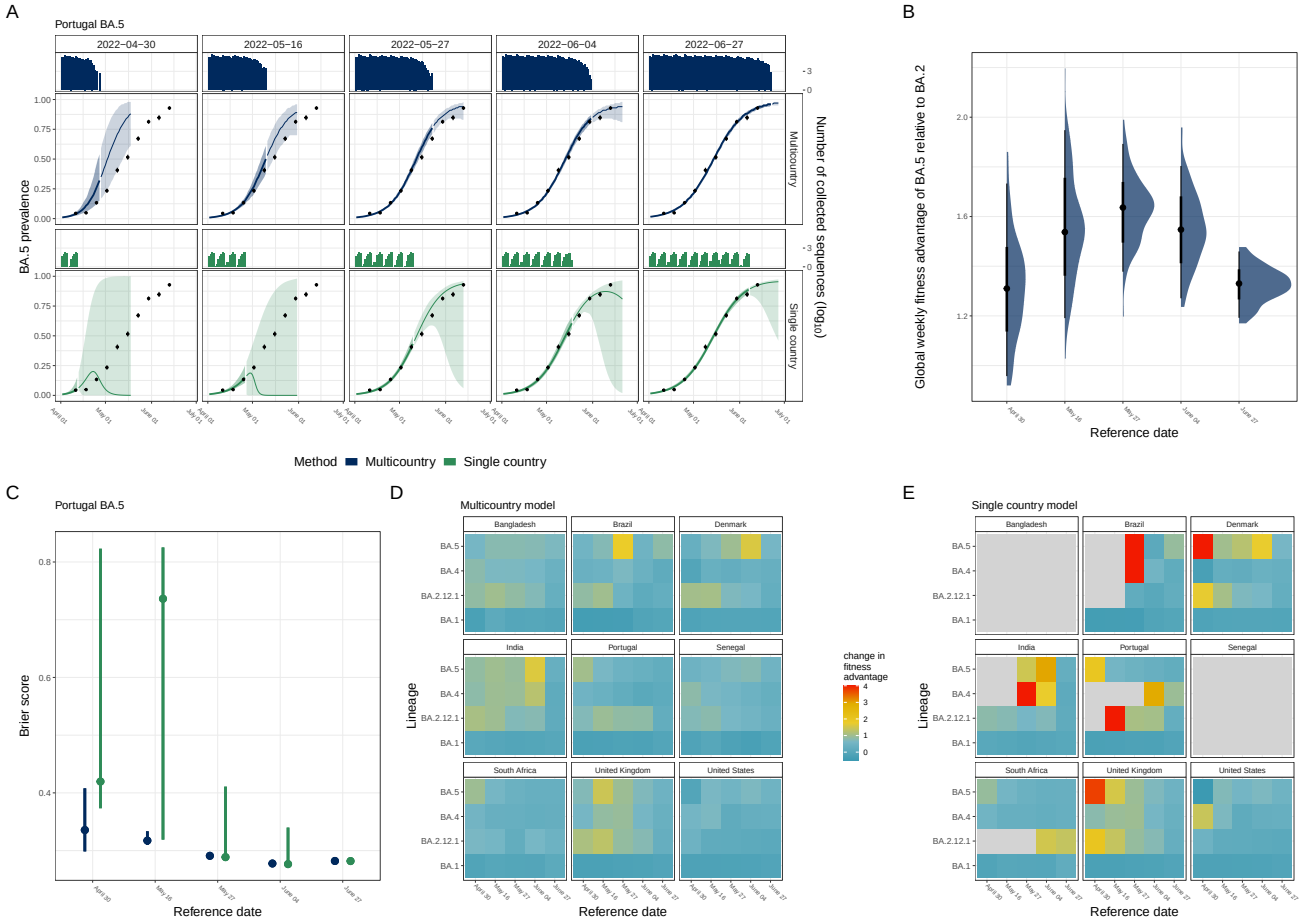

Figure 4. Historical validation to assess stability and predictive power of the multicountry model compared to a single country model applied to BA.5 emergence in Portugal. (A) Model estimated BA.5 prevalence in Portugal from 5 successive reference dates (columns). Rows indicate whether the multicountry (blue) or single country (green) model is applied, with the bar plots indicating the number of BA.5 sequences collected by date globally or in Portugal, respectively. Uncertainty bands represent the 95% credible interval of the multicountry model (blue) and the 95th percentiles of the non-parametric bootstrapped estimates for the single country model (green), with shading indicating the calibration period (darker) and the forecasting period (lighter). Black points indicate the observed variant prevalence as of July 1st, 2022, with error bars indicating the standard deviation. (B) Posterior distribution of the global estimate of BA.5's weekly fitness advantage compared to BA.2 over time. Points indicate median weekly fitness advantage, bands indicate the 66% and 95% credible intervals. (C) Brier score evaluating predictive accuracy of estimated BA.5 prevalence to the observed prevalence on July 1st, 2022. Points indicate the mean Brier score, bars indicate the 95% credible interval on the Brier score from the multicountry model, and the 95th percentiles of the non-parametric bootstrapped estimates for the single country model. (D)-(E). Country and variant specific changes in estimated fitness advantage from the July 1st, 2022 dataset compared to the estimates fit to the reference data set, for the multicountry model (D) and the single country model (E). Gray areas indicate that an estimate of the variant fitness advantage could not be made for that variant-country on that date due to insufficient data.

In Fig. 4B, we show the posterior distribution of the global estimates of the relative fitness advantage

of BA.5 over time, demonstrating that it is stable. Fig. S2 depicts how the multicountry fitness advantage estimates of BA.5 in Portugal compare to the single country model fitness advantage estimates over time. The single country model estimates experience wider uncertainty at early reference dates and a sharper decline.

In Fig. 4C, we evaluate the model predicted variant proportions from each reference date using the Brier score. Lower Brier scores indicate a more accurate probabilistic prediction. We compared the model predicted BA.5 proportion in Portugal to the observed proportions from the data as of July 1st 2022, using the Brier score over both the calibration and forecast period combined.The multicountry model exhibits lower Brier scores for the early reference dates during variant emergence, with scores of 0.34 [95% CI:0.30-0.31] and 0.32 [95% CI:0.31-0.33] as of April 30th, 2022 and May 16th, 2022 respectively compared to scores of 0.42 [95% CI:0.37-0.82] and 0.73 [95% CI:0.74-0.82] for the single country model on those same dates.

In Figs. 4D - E, we assess the stability of estimates from both the multicountry model (D) and the single county model (E) by showing the difference between the country-specific variant fitness estimate at each reference date compared to the country-specific variant fitness estimated at the final date (July 1st), for the two models. However, because the single country model cannot estimate country-specific variant fitness advantage if the particular variant has not been observed in sufficient quantities (<3 sequences of a variant) in that country, there are gaps in the single country model estimates indicated by the gray boxes in Fig. 4E. In contrast, the multicountry model is able to provide country-specific variant fitness advantage estimates, even before the variant has been observed in each country. Figure S3 provides the absolute transmission advantage estimates for each country-variant combination for both models.

The estimated fitness advantages from the multicountry model are more stable than those produced from the single country model (Figures S4 and S5). The estimated coefficients from the single country model are initially higher than those from the multicountry model and they decline to a stable estimate more slowly than those of the multicountry model (Figure S4). Although lower, the estimated country-specific fitness advantages from the multicountry model also have an initial transient positive bias, as can be seen in Figure 4D and Figure S5. However, the estimated global mean fitness advantages are quite stable over time (Figure S6), likely due to the additional shrinkage on these estimates from the partial pooling across global variant fitnesses (see Methods for additional information).

Discussion

Enabling broader access to information from genomic surveillance for emerging and re-emerging pathogens is and will remain a global health priority[5,7,20,24–26]. In addition to the crucial work of building global sequencing capacity, methodological advances in how sequences are analyzed can address gaps in the surveillance landscape. Here, we demonstrate a method to reliably estimate the growth of competing viral variants, generating both global and local estimates. We apply this method to emerging SARS-CoV-2 variants (Fig 3, 4, and Fig S3, S4, S5), highlighting its robustness both in regions with limited sequencing and for emerging variants with few sequences available. We also demonstrate its applicability across a number of settings, including influenza strain dynamics to show suitability across pathogens (Fig S7) and at the AL1 level in Brazil and Argentina to demonstrate relevance to local public health surveillance (Fig S8). We believe this approach is particularly relevant for lower and middle income countries (LMICs),

increasing the information available for local public health decisions in the present while capacity building continues to strengthen surveillance systems for the future[23].

Our method builds on the epidemiological (e.g.,[27]) and phylogenetic tools (e.g.,[28]) characterizing emerging variants, making two major contributions: stable characterization of emerging variant growth rate and improved nowcasting of variant prevalence. Characterizing the growth of emerging variants and accurate nowcasting of prevalence are challenging problems. There are often long lag times from sequence collection to submission and reporting, especially in locations with limited sequencing capacity ([20]). Emerging variants have few sequences available and, usually, have been detected in only a few countries. Consequently, characterizing expected growth and prevalence involves extrapolating from the limited information available and projecting yet-to-be-observed growth in new locales. Our method helps fill this information vacuum by pooling the information from all sequencing done globally to build a picture of the relationships between variant dynamics and countries' individual characteristics (i.e., empirically observed correlations between variant fitness in specific countries, potentially driven by similarities in the immune landscape or contact patterns). This global landscape of variant fitness is then applied to individual countries, informing our understanding of variant dynamics in data-sparse regions. It produces more reliable estimates of fitness advantage and improves the calibration of uncertainty intervals, particularly in the early days after a variant's emergence. It also generates more robust and stable estimates of variant dynamics over time than the standard approach of independent multinomial regressions[2,17,29,30].

Our estimates are robust even in settings with extremely sparse data (Fig 3 and Fig S3, S4, S5), such as LMICs with limited genomic surveillance. Early identification of high growth for emerging viral strains can be used to highlight potential new variants of concern or prepare health systems for the importation and impact of new SARS-CoV-2 variants. Identifying the viral strains likely to circulate in the upcoming months could also lead to improved vaccine strain selection for pathogens like influenza and SARS-CoV-2. The framework presented here is also general enough to be applied at the sub-country level, improving local estimates of variant prevalence at the municipality or state level and informing situational awareness for public health (Fig. S8).

We evaluate the pooling approach employed here by comparing it to the standard approach of country-independent multinomial regression models (Fig 4E)[17,31]. We show that our method outperforms this standard approach at estimating the growth of an emerging variant (Fig 4). Notably, the statistical regularization reduces overfitting to the noisy emergence data, reducing the biased overestimates of growth rates for novel variants produced by standard multinomial regression (Fig 4, Fig S3, S4, S5). While the mechanisms driving this transient bias are not fully understood, we speculate that this behavior is likely driven in part by potential biases in the observation process (e.g., prioritization of samples for sequencing), stochasticity in the disease-ecological process (e.g., superspreading dominating initial transmission), and overfitting to noisy data. However, these mechanisms are unlikely to completely explain the consistency in the direction of the bias across the observed lineages and settings and we hypothesize that additional mechanisms that have not yet been elucidated likely contribute as well (Fig S3, S5).

Although the modeling approach developed here can be more robust than standard approaches, it relies upon a number of assumptions and results and should be interpreted with care. We assume that sequences are randomly sampled from each geographic unit of cases (e.g., country), that samples of one variant are not prioritized for sequencing over those of another variant (e.g., due to S-gene target failure in PCR testing), and that sequences are perfectly assigned to Pango lineages. Further we assume even mixing of

variants within a patch (i.e., country), that the relative fitness of one variant to another is independent over time, and that the generation interval of each variant is the same. We note, however, that estimates of the fitness advantage can be robust to biases in sampling or sub-patch structure in disease dynamics provided that the sequences come from a consistent subset of the population. These estimated fitness advantages are likely also robust to misspecification of the generation interval distribution family[8], but the idealization of a constant generation interval across variants can be inaccurate[32,33] and can thus bias estimates of transmission advantage[34]. Further, we don't aim to explicitly identify the very first introductions of any given variant.

In future developments of our method, we intend to extend our preliminary work on influenza and extend to additional pathogens. Further we plan to explicitly include drivers of the immune landscape of a population, including previous infections and vaccine coverage, as well as mobility and assortative mixing. This will enable us to disentangle drivers of fine-scale epidemiological trends.

We believe this method complements investments in genomic sequence capacity building by better leveraging data across regions, to both incentivize data sharing and provide more equitable access to real-time localized variant dynamics. The global variant dynamics we uncover here improve our understanding of variant emergence and growth as well as provide an avenue to further develop our understanding across pathogens and geographic scales in the future. Methods such as the one we present here are critical for situational awareness, evidence-based decision-making, and policy design for pathogen spread. New analysis methods, such as the one we present here, are crucial to the continued maturation of genomic sequencing into real-time surveillance informing public health guidance.

Methods

Data processing

Line list SARS-CoV-2 sequence metadata was accessed via the GISAID EpiCov database[21]. The findings of this study are based on metadata associated with 2,032,779 sequences available on GISAID up to July 1, 2022, via https://doi.org/10.55876/gis8.230118ka. Using the Pango lineage annotations in the sequence metadata, we aggregated the observed sequences by country, collection date, and pango lineage to get daily counts of the number of lineages observed. We truncated lineage assignments to their root assignment (e.g., BA.5.1 to BA.5), but preserved specified lineages of interest (BA.1, BA.2, BA.2.12.1, BA.4, and BA.5). Lineages with fewer than 50 observed sequences in the time period of interest were marked as "Other". Additional details on data processing are available in the Supplemental Methods.

Hierarchical generalized linear modeling approach

We modeled the dynamics of competing variants of a directly transmitted infectious disease using a hierarchical multinomial regression approach (Figure 2). This approach can be used generally for competing strains, but here we used the example of SARS-CoV-2 variants using sequence metadata from GISAID to produce estimates of relative variant growth rates and true proportion of total cases in a given country.

We modeled the observed sequence counts $Y_{i,j,t}$ of variant i in country j on day t as drawn from a multinomial distribution (Eq.1), with the probability $p_{i,j,t}$ of observing a given variant in a country on a given day defined as the softmax of the linear predictor (Eq. 2).

$$Y_{i,j,t} \sim multinom(N_{j,t}, p_{i,j,t}) \tag{1}$$

$$p_{ijt} = \frac{e^{\beta_{0ij}+\beta_{1ij}t}}{\sum_{j=1}^{J} e^{\beta_{0ij}+\beta_{1ij}t}} \tag{2}$$

The intercepts of the linear predictor ($\beta_{0,i,j}$) are defined as random effects drawn from a normal distribution. The intercepts describe the initial variant prevalence on the scale of the linear predictor on day 0. The country-specific relative variant growth rates ($\beta_{1,i,j}$) are defined as j vectors of random effects drawn from a multivariate normal distribution (Eq. 3).

$$\begin{bmatrix} \beta_{11j} \\ \vdots \\ \beta_{1Ij} \end{bmatrix} \sim MVN(\begin{bmatrix} \mu_{\beta_{11}} \\ \vdots \\ \mu_{\beta_{1I}} \end{bmatrix}, \Sigma) \tag{3}$$

The country-specific relative growth rates $\beta_{1,i,j}$ describe the difference in intrinsic growth rates for the variant of interest and the reference variant (i.e., where $\beta_{ikj} = \phi_{kj} - \phi_{kj}$, where ϕ_{kJ} the intrinsic/Malthusian growth rate of the dominant variant in country K).

The elements of the vector of the mean relative variant growth rates ($\mu_{\beta_{1,i}}$) in the multivariate normal distribution are themselves random effects, drawn from a normal distribution with mean $\mu_{hierarchical}$ (Eq. 4).

$$\mu_{\beta_{1i}} \sim N(\mu_{hierarchical}, \sigma^2_{hierarchical}) \tag{4}$$

The hierarchical structure and its implications are described schematically in Figure 2. For interpretability, we presented the global and country-specific estimates of the relative growth rates as relative weekly fitness advantages using the relation as described by Davies et al.[30]

$$f = e^{7r} - 1 \tag{5}$$

Where r is the relative growth rate (either $\beta_{1,i,j}$ or $\mu_{\beta_{1i}}$) and f is the weekly fitness advantage displayed in the results.

This modeling approach made a number of assumptions about the data generating process. These included random sampling of sequences within countries, that variants are correctly assigned to lineages, and that sequence collection dates are correctly reported. The modeling approach also assumed that relative variant fitness does not change over the modeled time period, such as could potentially occur due to a vaccination campaign changing the immune landscape of the host population (i.e., during the 90-day window).

Additional detail on the observation process model, model structure, prior specification, and model limitations is available in the Supplemental Methods.

Retrospective validation of country-specific variant prevalence projections

Processing and fitting of historical datasets

SARS-CoV-2 line-list sequence data from GISAID was accessed on the following dates, which we refer to as "reference" dates: April 30th, 2022, May 16th, 2022, May 27th, 2022, June 4th, 2022, June 27th, 2022 and July 1st, 2022. A consensus set of lineages was found by identifying all lineages that exceed

the global threshold of 50 or more observed sequences in the past 90 days across any of the 6 reference datasets. For the dataset from each reference date, we defined a calibration period: the time period up to and including the day the last sequence was collected. For the multicountry model all sequences in the dataset were included and for the single country model only sequences from the country of interest were included. We also defined an associated forecast period: the days after the last sequence was collected. For example, if we accessed the data on April 30th, 2022, and the most recent collection date in that dataset was April 27th, 2022, than the calibration period would be any time before and including April 27th, and the forecast period would be any time after April 27th.To test the model's predictive power, we forecasted a maximum of 21 days out and compared this to any data observed by the last reference dataset from July 1st, 2022.

Estimation model comparison

For each reference dataset, we compared variant prevalence estimations from the multicountry model with estimations from the single country multinomial model. In order to get an estimate of the uncertainty of these outputs, non-parametric bootstrapping with replacement was performed, generating 100 bootstrapped datasets with time points randomly sampled with replacement from the true data. To fit the single country model to the data and the boot-strapped datasets, we used the nnet package[35] in R which returns the maximum likelihood estimation (MLE) of the multinomial model parameters, which we transform into variant fitness advantages and variant prevalence dynamics using equations 2 and 5.

Evaluation of model estimates

For both the multicountry and the single country model, we used 100 draws from the distributions of lineage prevalence estimates to evaluate the accuracy of the model predictions compared to the observed daily lineage prevalence from the July 1st, 2022 reference dataset. For countries that observed no sequences of a particular lineage in the consensus dataset during the 90 day time window, the single country estimation model does not estimate a prevalence for that lineage. To enable a fair comparison of the two model outputs at the country level, we collapsed all prevalence estimations from the multicountry model for these unobserved lineages into "other" for that country. We used the Brier score to evaluate the accuracy of each draw of the model output. The Brier score was calculated at the country-level for both the calibration period and the forecast period, and the combination. The result was a distribution of Brier scores at each reference time point.

References

1. Funk, T. *et al.* Characteristics of SARS-CoV-2 variants of concern b.1.1.7, b.1.351 or p.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Euro Surveill.* **26** (2021).
2. Kraemer, M. U. G. *et al.* Spatiotemporal invasion dynamics of SARS-CoV-2 lineage b.1.1.7 emergence. *Science* **373**, 889–895 (2021).
3. Tegally, H. *et al.* Sixteen novel lineages of SARS-CoV-2 in south africa. *Nat. Med.* **27**, 440–446 (2021).
4. Tegally, H. *et al.* Emergence of SARS-CoV-2 omicron lineages BA.4 and BA.5 in south africa. *Nat. Med.* 1–6 (2022).
5. Stockdale, J. E., Liu, P. & Colijn, C. The potential of genomics for infectious disease forecasting. *Nat Microbiol* **7**, 1736–1743 (2022).
6. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *Lancet* **395**, 497–506 (2020).

7. Hill, S., Perkins, M. & von Eije, K. Genomic sequencing of SARS-CoV-2. Tech. Rep., World Health Organization (2021).

8. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage b.1.1.7 in england. *Science* **372** (2021).

9. Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).

10. Borchering, R. K. *et al.* Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios - united states, April-September 2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 719–724 (2021).

11. Cramer, E. Y. *et al.* The united states COVID-19 forecast hub dataset. *Sci. Data* **9**, 1–15 (2022).

12. Kaiming Bi, Anass Bouchnita, Oluwaseun F. Egbelowo, Spencer Fox, Michael Lachmann, Lauren Ancel Meyers. Scenario projections for the spread of SARS-CoV-2 omicron BA.4 and BA.5 subvariants in the US and texas. (2022).

13. Johnson, K. *et al.* Real-Time projections of SARS-CoV-2 b.1.1.7 variant in a university setting, texas, USA. *Emerg. Infect. Dis. journal* **27**, 3188 (2021).

14. Earn, D. J. D., Dushoff, J. & Levin, S. A. Ecology and evolution of the flu. *Trends Ecol. Evol.* **17**, 334–340 (2002).

15. Taylor, B. S., Sobieszczyk, M. E., McCutchan, F. E. & Hammer, S. M. The challenge of HIV-1 subtype diversity. *N. Engl. J. Med.* **358**, 1590–1602 (2008).

16. Andraud, M., Hens, N., Marais, C. & Beutels, P. Dynamic epidemiological models for dengue transmission: A systematic review of structural approaches. *PLoS One* **7**, e49085 (2012).

17. Paul, P. *et al.* Genomic surveillance for SARS-CoV-2 variants circulating in the united states, december 2020-may 2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 846–850 (2021).

18. van Smeden, M. *et al.* No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med. Res. Methodol.* **16**, 163 (2016).

19. de Jong, V. M. T. *et al.* Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat. Med.* **38**, 1601–1619 (2019).

20. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. *Nat. Commun.* **13**, 7003 (2022).

21. Khare, S. *et al.* Gisaid's role in pandemic response. *China CDC Wkly.* **3**, 1049 (2021).

22. McCrone, J. T. *et al.* Context-specific emergence and growth of the SARS-CoV-2 delta variant. *Nature* **610**, 154–160 (2022).

23. SARS-CoV-2 genomics surveillance capacity map. https://www.finddx.org/covid-19/covid-19-genomic-surveillance/sars-cov-2-genomics-surveillance-capacity-map/ (2022). Accessed: 2022-12-20.

24. Lipsitch, M. & Santillana, M. Enhancing situational awareness to prevent infectious disease outbreaks from becoming catastrophic. In Inglesby, T. V. & Adalja, A. A. (eds.) *Global Catastrophic Biological Risks*, 59–74 (Springer International Publishing, Cham, 2019).

25. Fineberg, H. V. Pandemic preparedness and response–lessons from the H1N1 influenza of 2009. *N. Engl. J. Med.* **370**, 1335–1342 (2014).

26. Kucharski, A. J., Hodcroft, E. B. & Kraemer, M. U. G. Sharing, synthesis and sustainability of data analysis for epidemic preparedness in europe. *Lancet Reg Heal. Eur* **9**, 100215 (2021).

27. Chen, C. *et al.* CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics* **38**, 1735–1737 (2021).

28. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

29. SARS-CoV-2 variants of concern and variants under investigation in england: Technical briefing 43. (2022).

30. Davies, N. G. *et al.* Increased hazard of death in community-tested cases of SARS-CoV-2 variant of concern 202012/01. *medRxiv* (2021).

31. CDC. COVID data tracker. https://covid.cdc.gov/covid-data-tracker/ (2020). Accessed: 2022-9-15.

32. Hart, W. S. *et al.* Inference of the SARS-CoV-2 generation time using UK household data. *Elife* **11**, e70767 (2022).

33. Hart, W. S. *et al.* Generation time of the alpha and delta SARS-CoV-2 variants: an epidemiological analysis. *Lancet Infect. Dis.* **22**, 603–610 (2022).

34. Park, S. W. *et al.* The importance of the generation interval in investigating dynamics and control of new SARS-CoV-2 variants. *J. R. Soc. Interface* **19**, 20220173.

35. Ripley, B. Feed-Forward neural networks and multinomial Log-Linear models [r package nnet version 7.3-17]. (2022).

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

37. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22** (2017).

38. Mathieu, E. *et al.* Coronavirus pandemic (COVID-19). *Our World Data* (2020).

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative[21,36,37], on which this research is based. We acknowledge the financial support of The Rockefeller Foundation, who funded this work. We thank Nick Reich, Spencer Fox and Moritz Kraemer for their valuable comments. We also gratefully acknowledge Our World in Data[38], who curated data enabling this work.

Author contributions statement

ZS, KEJ, SVS, and AIB designed the study, ZS, KEJ, SVS, LW, and AIB conceptualized the work, ZS and KEJ wrote the code, conducted the analyses and wrote the first draft. ZS, KEJ, SVS, and AIB contributed to the model formulations and interpreted the results. All authors contributed to editing the manuscript and gave final approval for publication and agree to be held accountable for the work performed therein.