Frequency dynamics predict viral fitness, antigenic relationships and epidemic growth

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

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

Abstract

During the COVID-19 pandemic, SARS-CoV-2 variants drove large waves of infections, fueled by increased transmissibility and immune escape. Current models focus on changes in variant frequencies without linking them to underlying transmission mechanisms of intrinsic transmissibility and immune escape. We introduce a framework connecting variant dynamics to these mechanisms, showing how host population immunity interacts with viral transmissibility and immune escape to determine relative variant fitness. We advance a selective pressure metric that provides an early signal of epidemic growth using genetic data alone, crucial with current underreporting of cases. Additionally, we show that a latent immunity space model approximates immunological distances, offering insights into population susceptibility and immune evasion. These insights refine real-time forecasting and lay the groundwork for research into the interplay between viral genetics, immunity, and epidemic growth.

Introduction

The COVID-19 pandemic was marked by the successive emergence of SARS-CoV-2 variant viruses, driving repeated epidemics globally [1, 2]. While these repeated large waves occurred with the emergence of novel variants, the mechanism driving these variants' success changed over time. The spread of early variants such as Alpha, Beta, Gamma and Delta were largely driven by increases in intrinsic transmissibility [3]. The Omicron variant showed substantial immune escape [3] and subsequent derived lineages within Omicron including XBB, EG.5.1 and JN.1 appear to be driven by immune escape as evidenced through molecular studies of neutralization using human sera [4–7]. Since 2022, there has been repeated replacement by subsequent Omicron-derived lineages. This rapid viral population turnover is consistent with antigenic evolution and is observed in other viruses such as seasonal influenza [8], although SARS-CoV-2 currently remains an outlier in terms of pace of its evolution [9]. This transition from transmissibility-driven to immune escape-driven success is a consequence of the interplay between population immunity and variant fitness.

With the increased temporal and geographical scale of sequencing alongside a detailed genetic nomenclature [10] and bioinformatic tools for lineage assignment [11, 12], we have gained more data for SARS-CoV-2 than for other circulating viruses giving a unique

opportunity for insight into its evolution. Several models of variant frequency have been developed to estimate the fitness of emerging SARS-CoV-2 variants [13–18]. These models estimate the relative fitness (or selective advantage) of circulating variant viruses from their frequency in sequencing data, typically represented by counts of variant sequences over time within a geographic region. Relative fitness in these models is often assumed to be constant and intrinsic to the variant of interest. However, this may be an oversimplification of the transmission process.

It has been shown that these transmission advantages differ geographically and temporally, suggesting that variant transmission advantages are not necessarily fixed and may be informed by regional population differences [15,19]. In fact, heterogeneity in transmission advantages may be well explained by regional differences in immune structure as Dadonaite et al. [20] show deep mutational scanning estimates of immune escape are well correlated with estimated variant growth advantages. Existing models that allow variant transmission advantages to change in time often do not have a mechanistic underpinning for why transmission advantages exist and vary geographically and temporally [15,16]. Models that do include mechanistic grounding of transmission based on population immunity such as Meijers et al. [21] and Raharinirina et al. [22] have parameterized variant-specific immunity based on serological measurements or deep mutational scanning datasets. Timely experimental data is thus a requirement for these models to perform well for real-time evolutionary forecasting.

In response to this gap, we introduce a novel framework that links variant dynamics directly to transmission mechanisms using compartmental models of infectious diseases (Fig. 1). By modeling both intrinsic transmissibility and immune escape, we explain how shifts in population immunity shape the relative fitness of viral variants and select for immune escape over intrinsic transmissibility with increasing past exposure. Furthermore, including these mechanisms suggests that relative fitness varies in time, reflecting the evolving landscape of population immunity and exposure regardless of the underlying mechanism.

Here, we present a novel non-parametric method for estimating time-varying fitness regardless of the underlying transmission mechanism. Alongside this development, we introduce a "selective pressure" metric that quantifies the impact of variant turnover on population-level epidemic growth rates. Finally, we develop a latent immunity model that we use to estimate the underlying proportion of pseudo-immune groups within multiple geographies and pseudo-immune escape rates for circulating variants that predicts antigenic distances using sequence data alone. Overall, our framework bridges the gap between genetic data and transmission dynamics, offering a new way to predict and manage viral outbreaks.

Results

Variant dynamics and relative fitness in multistrain models

Multi-strain models of epidemics have been developed to understand the competition between different viral strains that exhibit different levels of cross-immunity [23, 24]. These models have typically been used to explain strain evolution in antigenically variable

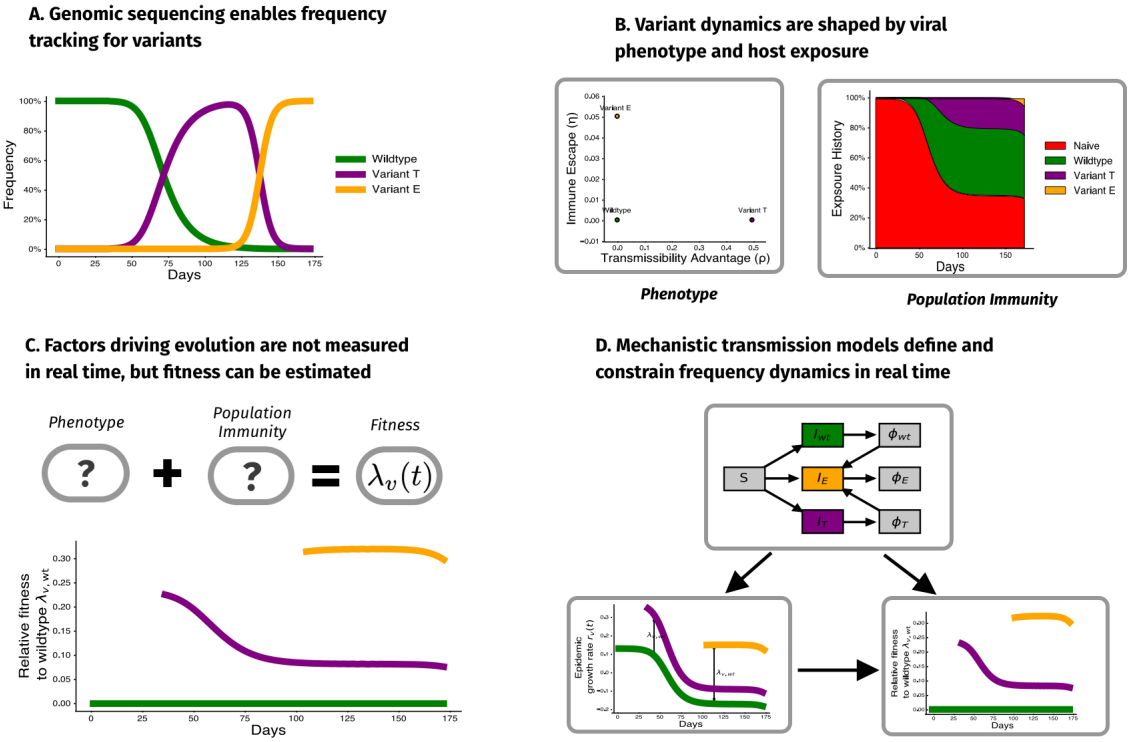

Fig. 1. Mechanistic transmission models inform frequency dynamics. (A) Genomic surveillance reveals changes in the frequency of genetic variations over time. (B) These frequency changes can arise due to differences in phenotypes related to transmission (e.g, immune escape, transmissibility, binding) and changes in population immunity due to recent exposure. (C) Despite being instrumental for real-time analysis, both variant phenotype and population immunity are rarely observed in real time. (D) We use mechanistic transmission models to infer relative fitness from frequency data alone, taking advantage of known structure in transmission dynamics. This enables us to quantify trade-offs between variant phenotypes and develop new methods for estimating fitness in populations undergoing antigenic evolution.

pathogens like seasonal influenza virus [8] and seasonal coronaviruses [25, 26].

We begin by modeling a population of V exponentially growing variant viruses v each with prevalence $I_v(t)$ and time-varying growth rate $r_v(t)$. By considering the difference in growth rates for variants v and u, we can define the relative fitness as $\lambda_{v,u}(t) = r_v(t) - r_u(t)$. This relative fitness determines the change in the frequencies of the variants in the population

$$f_v(t) = \frac{f_v(0) \exp\left(\int_0^t \lambda_{v,v^*}(s)ds\right)}{\sum_{u=1}^{V} f_u(0) \exp\left(\int_0^t \lambda_{u,v^*}(s)ds\right)}, \tag{1}$$

where v^* is a chosen pivot variant that has relative fitness zero.

In order to better understand frequency dynamics of pathogens with multiple co-circulating variants, we apply the above framework to compartmental models of epidemics, which can

be written as time-varying exponential growth (detailed in Supplementary Text S1). These models provide an intuition of how strain-level selection depends on the assumed transmission mechanism of the underlying epidemic model. This framework also generalizes several existing methods for relative fitness estimation and prediction (detailed in Supplementary Text S2). We summarize dynamics of a three-variant mechanistic transmission model in Fig. S1, where we compare a transmission variant T with a 50% increase in transmissibility ($\rho = 0.5$) to an escape variant E that infects 5% of hosts possessing wildtype immunity ($\eta = 0.05$).

Determining the transmissibility-escape tradeoff

To understand the fitness trade-off between transmissibility and immune escape, we consider dynamics with a wildtype virus W with $\rho_W = 0$ and $\eta_W = 0$, an increased transmissibility variant T with $\rho = \rho_T > 0$ and $\eta = \eta_T = 0$ and an immune escape variant E with $\rho_E = 0$ and $\eta_E > 0$.

Following Equation 35, we write relative fitnesses of the escape variant or transmissibility variant as

$$\lambda_{E,W} = \eta\beta\phi_W(t), \tag{2}$$
$$\lambda_{T,W} = \rho\beta S(t), \tag{3}$$

where β is the transmissibility coefficient, η is escape proportion against the wildtype, ρ is the variant's proportional increase to transmissibility, and $S(t)$ and ϕ_W are the proportions of the population who are susceptible and have wildtype immunity respectively.

In the simplest case where individuals are either susceptible or have wildtype immunity ($S(t) + \phi_W(t) = 1$), we can compute the critical immune fraction ϕ^* at which $\lambda_{E,W}(\phi^*) = \lambda_{T,W}(\phi^*)$ as

$$\phi^* = \frac{\rho}{\eta + \rho}. \tag{4}$$

For past exposure level greater than ϕ^* escape variants have a higher relative fitness. This trade off shows that increasing degree of escape entails that a lower proportion of past exposure is needed for escape variants to be preferred (Fig. 2). Additionally, this shows that when intrinsic transmissibility increases are limited escape is more likely to be a dominant mechanism for variant turnover.

Initial growth rates insufficient for predicting short-term frequency growth

One question of interest is whether knowledge of mechanism meaningfully informs our ability to forecast short-term frequency growth. The first step to addressing this is to understand how the relative fitness may change in time to understand the predictability of relative fitness in the short-term.

We find that the mechanistic forms analyzed in this paper (Supplementary Text S1) can be represented as weighted combinations of B time-varying functions $\Upsilon_b(t)$ with weights $\beta_b^{(v,u)}$. We can think of each of these functions Υ_b as an immune background and the

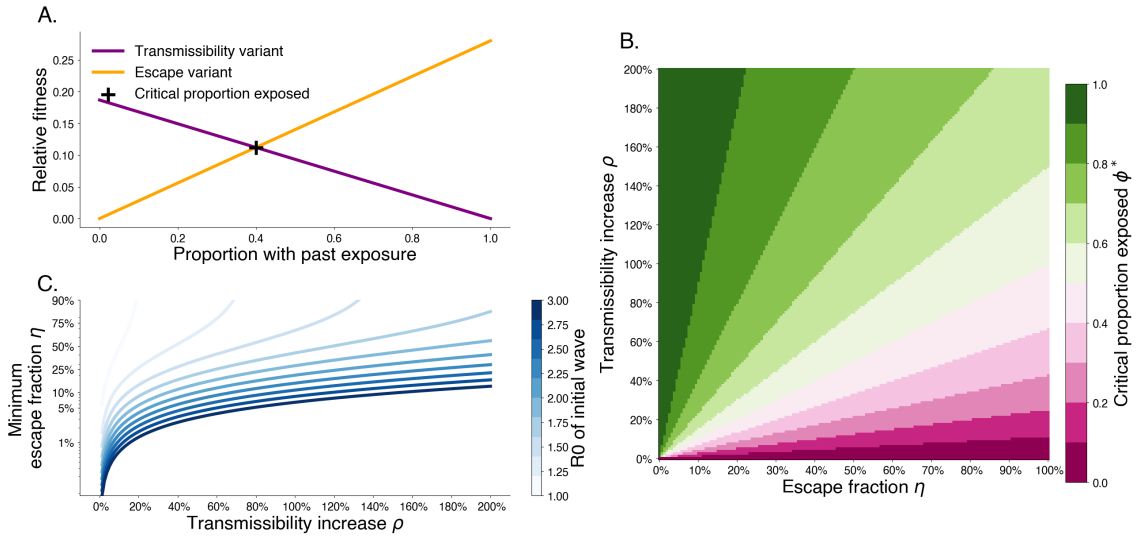

Fig. 2. Trade-off between degree of immune escape and increased transmissibility. A. Relative fitness for a transmissibility increasing variant T with $\rho = 0.2$ and an immune escaping variant E with $\eta = 0.3$ for $R_{0,W} = 2.8$ and $1/\gamma = 3.0$ days. The intersection point shows that after 40% of the population has wildtype immunity, the escape variant has higher fitness. B. The critical exposure proportion is shown for various escape fraction and transmissibility increase. Above the critical exposure proportion, we expect dominance of escape variants. C. The minimum escape fraction needed for second waves to be comprised of escape variant assuming competition with transmissibility increase variants and first wave with a given R_0.

coefficient $\beta_b^{(v,u)}$ as a transmission differential between variants v and u, so that

$$\lambda_{v,u}(t) = \sum_{1 \le b \le B} \beta_b^{(v,u)} \Upsilon_b(t). \tag{5}$$

Even in the case of complete knowledge of the relative fitness and the underlying fitness contributions in the present and past, we have that change in the relative fitness is determined by

$$\frac{d\lambda_{v,u}}{dt} = \sum_{1 \le b \le B} \beta_b^{(v,u)} \frac{d\Upsilon_b}{dt}(t). \tag{6}$$

By considering a Taylor expansion of the relative fitness about the point of estimation t_0, we can approximate the relative fitness in the future as

$$\lambda_{v,u}(t) \approx \lambda_{v,u}(t_0) + (t - t_0) \sum_{1 \le b \le B} \beta_b^{(v,u)} \frac{d\Upsilon_b}{dt}(t_0). \tag{7}$$

This suggests small differences in the form of $\lambda_{v,u}(t)$ can lead to meaningful differences in the future relative fitnesses through changes in the underlying immune backgrounds.

We investigate whether relative fitnesses vary predictably in the short-term regardless of mechanism. To do so, we apply the two-variant model developed in previous sections for

different mechanisms of immune escape and increased transmissibility. We fix the relative fitness of the novel variant at a prediction time t_0 using Equation 4 and assess the change in the relative fitness in the short-term. We find that although relative fitness trajectories share the same decreasing shape, they may decline at different rates depending on the mechanism (Fig. S3). This can lead to substantial changes in the predicted incidence depending on the assumed mechanism and affects to overall rate of turnover.

Correlations are insufficient for mechanism identification

Although correlations between vaccination uptake and variant growth advantage are often observed, these alone may not be sufficient to identify the mechanism behind a variant's success. A variant's fitness advantage may arise from increased transmissibility, immune escape, or a combination of both. Even in the absence of immune escape, the relative fitness of a variant depends on the proportion of the population that is susceptible to infection and therefore changes with both past exposure and vaccine uptake (Supplementary Text S1). To illustrate this, we simulate the spread of a variant with increased transmissibility in populations with varying initial vaccination levels.

In populations with lower vaccination levels, the variant's prevalence peaks more sharply and its relative fitness declines quickly as immunity accumulates within the population (Fig. 3A-C). In contrast, higher vaccination levels constrain relative fitness, leading to a delayed peak in prevalence and more stable relative fitness as the existing immunity limits the variant's spread (Fig. 3A-C). Even without immune escape, estimated growth advantages for this variant decrease with increasing vaccination uptake near the beginning of an epidemic (Fig. 3D). Later in the epidemic, this relationship reverses with estimated growth advantages over the full period increasing with initial vaccination levels, which may be mistaken as signal for immune escape (Fig. 3E).

This analysis shows that correlation-based methods alone may struggle to identify the true mechanisms driving a variant's success especially under the assumption of a fixed growth advantage. By explicitly considering how immunity and transmissibility interact within populations, models that incorporate these dynamics may provide a stronger foundation for understanding why certain variants spread.

Estimating relative fitness using approximate Gaussian processes

Our earlier approach shows that relative fitness is often dependent on the past exposure of a population (as discussed in Supplementary Text S1 and extended to full immune history models in Supplementary Text S3). This suggests that serology, vaccination history, and immunological data generally can be informative of relative fitness. Additionally, when working with variant classifications, non-neutral evolution within a variant will cause the relative fitness of that variant to change in time. However, even in the absence of external data that can inform relative fitness, there is still hope.

We develop a method for using approximate Gaussian processes to model variant relative fitness. Gaussian processes are probability distributions over functions, where the structure and smoothness of these functions are defined by a kernel that encodes correlations in time. These models are flexible and allow us to encode smoothness constraints, period-

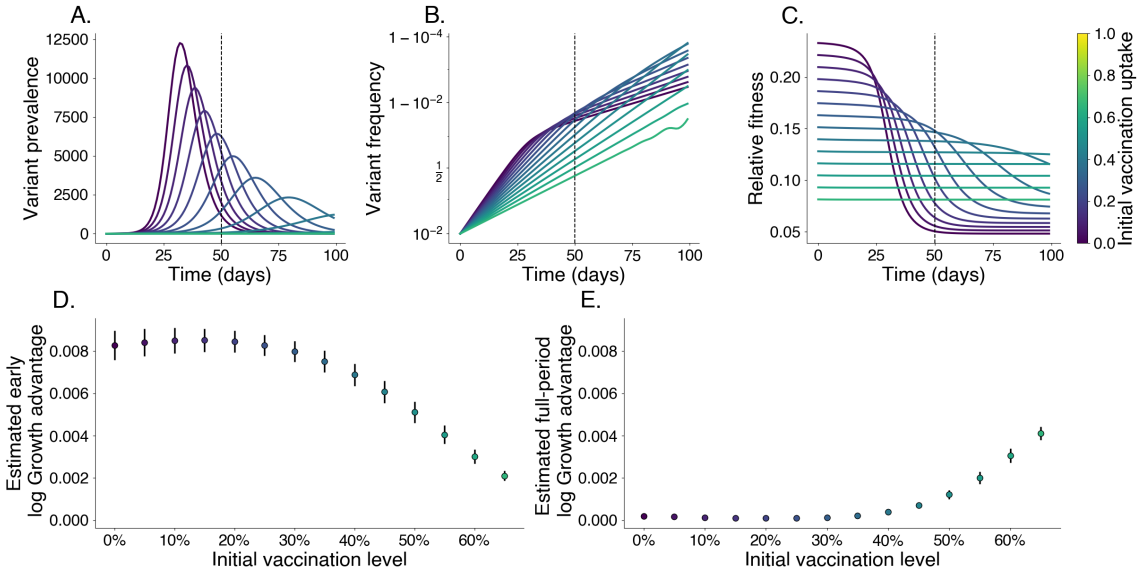

Fig. 3. Relative fitness is correlated with vaccination levels in the absence of immune escape. We simulate the growth of a pure transmissibility increased variant at varying levels of vaccination. Darker colors represent lower vaccine uptake. We identify an early growth period where relative fitness is at its highest; the cutoff for this period is denoted with a vertical dashed line. A. Prevalence of variant, each line is its own simulation. B. Frequency of variant. C. Relative fitness for variant over time. D. Estimated log growth advantage using linear regression of log relative frequency of variant over wildtype using only data before the early cutoff. E. Same as D. but using data from the entire period shown.

icity, and other structures [27]. Gaussian processes allow us a non-parametric estimate of the relative fitness for variants through time (see Materials and Methods).

Traditional Gaussian processes, while flexible, face challenges for large time series and large data sets. Our approach overcomes this using a Hilbert Space Gaussian Process (HSGP) approximation and shared eigenbasis, making the framework scalable for many variants and long time periods [28]. This enables real-time variant fitness estimation and can be applied to any frequency data regardless of the underlying transmission mechanism or immunity assumptions. This model is used in Fig. S2 to estimate the relative fitnesses of different variants through time based on simulated variant sequence counts from frequencies shown in Fig. S1.

Later, we also apply this model to empirical SARS-CoV-2 sequence data from 50 US states and England from 2021 to 2022 to estimate relative fitness for variants circulating in that period.

Quantifying selective pressure

Although it is useful to quantify the relative fitnesses of individual variants, we are often interested in quantifying the overall effects of selection in the population. With this in

mind, we define a population-level metric of overall selective pressure

$$\psi(t) = \mathbb{E}_{f(t)}\left[\frac{d\lambda_v}{dt}\right] + \mathbb{V}_{f(t)}[\lambda_v] \tag{8}$$

that describes the distribution of relative fitness in the population using the expectation of the fitness change and the variance of fitness. This selective pressure metric serves as an indicator for high fitness variants arising in the population as change. High fitness variants rising from initially low frequency leads to large increases in the variance of the fitness distribution and therefore increases in the selective pressure.

The selective pressure metric enables us to decompose changes in the average growth rate in the population, $d\bar{r}/dt$, to an evolutionary component ψ and a residual baseline growth rate r_W following

$$\frac{d\bar{r}}{dt} = \frac{dr_W}{dt} + \psi(t). \tag{9}$$

This shows that increased selective pressure through emerging high fitness variants can drive waves of infection. Further, this suggests that differences between growth rates based on selective pressure alone and observed rates are attributable to changes in baseline transmission over time. This mirrors ideas of Fisher's theorem of natural selection and its later interpretations with the variance of fitness contributing directly to the change in transmission rates (or fitness) [29, 30]. This definition of selective pressure captures how relative fitness contributes to epidemic growth. This is similar to ideas quantifying rates of adaptation via fitness flux [31].

In this case, the overall growth rate $\bar{r}$ and relative incidence $I(t)/I(0)$ can be written directly

$$\bar{r}(t) = \bar{r}(0) + [r_W(t) - r_W(0)] + \Psi(t), \tag{10}$$

$$\frac{I(t)}{I(0)} = \exp\left(\int_0^t [r_W(s) + \Psi(s)]ds\right), \tag{11}$$

using the cumulative selective pressure $\Psi(t) = \int_0^t \psi(s)ds$. In addition to estimating the relative fitness, metrics derived from these models can inform us of much more.

Our "selective pressure" metric allows us to model the contribution of evolution to changes in the epidemic growth rate of a population and is independent of pivot choice for relative fitness estimation. This metric acts as an early warning system for variant-driven outbreaks, especially in scenarios where case data are sparse or delayed. This metric can be computed using any method that estimates variant frequency and relative fitnesses and serves as a simple tool for understanding the contribution of selection to the overall population dynamics.

The full derivation of this metric and its contribution to the overall growth rate can be found in Supplementary Text S4.

Predicting epidemic growth rates using selective pressure

Motivated by the relationship between epidemic growth rate and selective pressure demonstrated above, we develop a predictive model of epidemic growth rate using estimates of

selective pressure. Using empirical SARS-CoV-2 sequence data from 50 US states between January 2021 and November 2022, we first estimate selective pressure through time using our approximate Gaussian process model on sequence counts (Fig. 4 A–C.) Here, we group variants at the granularity of Nextstrain clades [12] resulting in 28 distinct variants over this time period. As expected we see that relative fitness increases through time and that selective pressure corresponds to speed of clade turnover where the sweep of Omicron BA.1 (clade 21K) yields the strongest signal of selective pressure (Figs. S4–S8). Using case counts from each state, we estimate epidemic growth rates between January 2021 and November 2022. We use these epidemic growth rates to fit a gradient-boosted regressor to predict epidemic growth rates using selective pressure from the most recent 28 days, reserving data between July 2022 and November 2022 for testing (Fig. 4 D–I, Fig. S9). This regressor is chosen via time series cross-validation among model architectures and grid-search parameter tuning (Fig. S10).

We observe a strong correspondence between observed epidemic growth rate and model predictions with Pearson R^2 in the training period of 0.576 and a weaker Pearson R^2 in the testing period of 0.077. As case reporting declined over this period, we expect weaker correspondence between our predictions and epidemic growth rates computed from case data. To address this, we sought to evaluate the out-of-sample fit on case data from other countries e.g. South Africa, South Korea, and the United Kingdom, achieving an R^2 of 0.196.

To address the potential for this method under steady reporting rates, we validate this method by predicting the epidemic growth rates in England derived from the Office for National Statistics (ONS) Coronavirus Infection Survey between February 2022 and November 2022. The ONS Infection Survey represented a randomly sampled panel survey of households where nasal swabs were collected regardless of symptom status allowing for prevalence estimates despite faltering case reporting [32]. Our model is able to replicate patterns seen in epidemic growth rates in England derived from ONS data (Fig. 4 J–L), achieving a coefficient of variation of $R^2 = 0.329$ and mean absolute error of 0.026. Performance is significantly better for the first two subsequent waves, falling off in accuracy for the fall 2022 BQ.1 (clade 22E) wave.

Although these predictions can be biased by non-evolutionary effects on the epidemic growth, this approach provides a simple measure of epidemic growth in the absence of high quality case counts using sequence data alone.

Latent factor model of relative fitness

The representation of relative fitness using discrete immune backgrounds suggests that there may be low-dimensional structure to variant relative fitness when tranmission is shaped by past host exposure. To elucidate these factors, we develop and implement a method for latent factor analysis of relative fitness from sequence data alone. This model assumes that variants intrinsically escape the immune responses with particular groups and that differences in a variant's relative fitness between geographies is attributable to differences in immunity between populations. This enables us to estimate pseudo-escape rates for variants as well as pseudo-immunity groups within geographies over time.

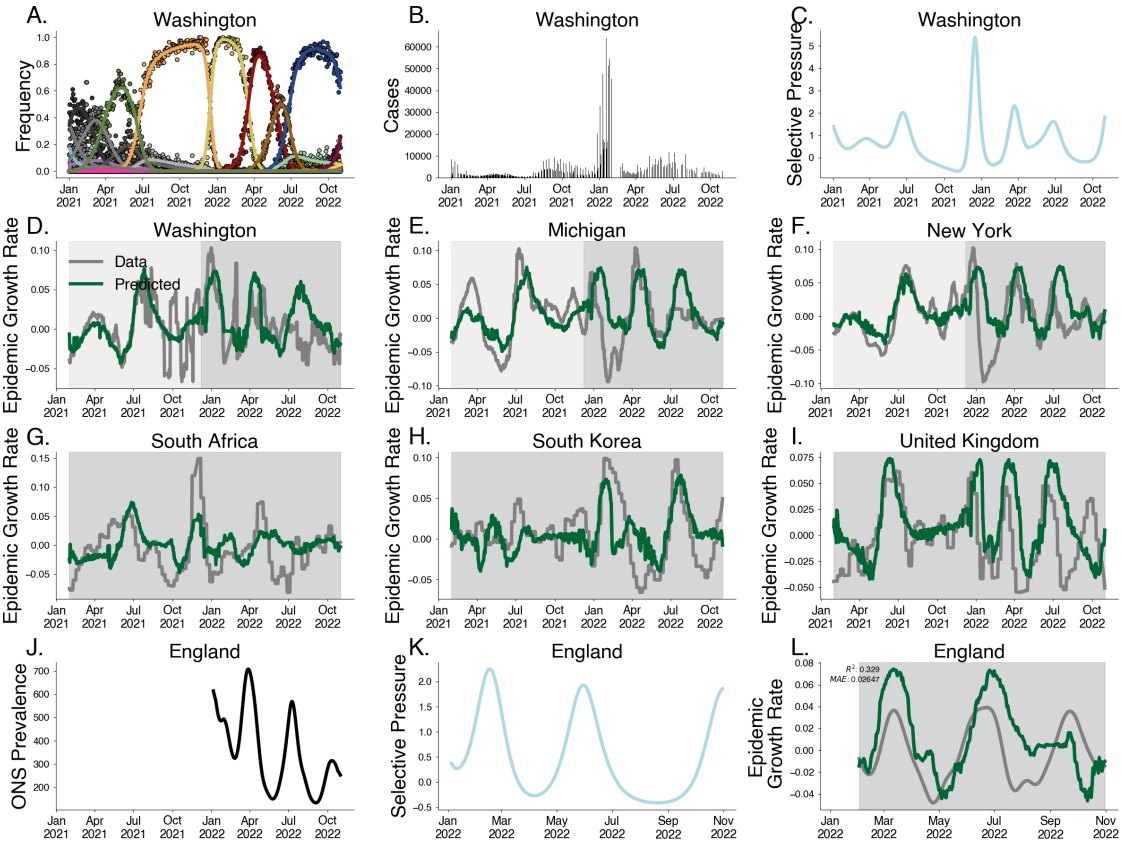

Fig. 4. Predicting epidemic growth rate using estimated selective pressure. A. Variant frequency estimated using the Gaussian process relative fitness model between January 2021 and November 2022 for sequence count data from Washington state. B. Case counts from Washington state. C. Selective pressure computed using estimated variant frequencies and relative fitnesses from Washington state. D-F. Predictions for empirical growth rate from selective pressure for selected US states. The light gray period is the training period and the darker gray is the testing period. G-I. Predictions for empirical growth rate from selective pressure for countries South Africa, South Korea and the UK. J. Prevalence estimates for England from ONS Infection Survey. K. Estimated selective pressure in England. L. Empirical growth rates (gray) computed from prevalence estimates and predictions from our model (green) computed from selective pressure.

We generate Pango lineage-level sequence counts for 18 countries and 53 variants between March 2023 and March 2024. These 18 countries were chosen based on availability of sequence data. Small lineages that do not meet a count threshold are collapsed into their parent lineages. This leaves us with a total of 53 variants, so that each variant met a threshold for number of sequences available.

Using these sequence counts, we apply our latent factor model to estimate the relative fitness of each variant over time in each country, pseudo-escape rates for each variant, and pseudo-immunity for each country simultaneously for $D = 8$ pseudo-immune groups (Fig. 5). This model is significantly constrained relative to estimating the time-varying fitness independently in each location, resulting in a model with 2,752 parameters compared to 7,488 parameters in the independent model. The results of this model are visualized in

Fig. 5 for several selected variants and countries of interest.

We chose $D = 8$ for our primary analysis by noting the point at which the loss function seems to stagnate with increasing D, i.e., the "elbow" method (Fig. S11A). Further, we observe that Bayesian Information Criterion (BIC) is minimized between 7 and 9 groups (Fig. S11D). However, the exact choice of latent immune dimensionality is necessarily somewhat arbitrary and we observe significant correlations with empirical titer data for fewer dimensions as well, although $D = 8$ also maximizes this correlation (Fig. S11B) and its significance is maintained for all dimensions $D > 8$ tested. Analogous figures showing pseudo-immunity and pseudo-antigenic relationships across variants can be seen for $D = 2$ in Fig. S12, $D = 4$ in Fig. S13, $D = 6$ in Fig. S14 and $D = 10$ in Fig. S15.

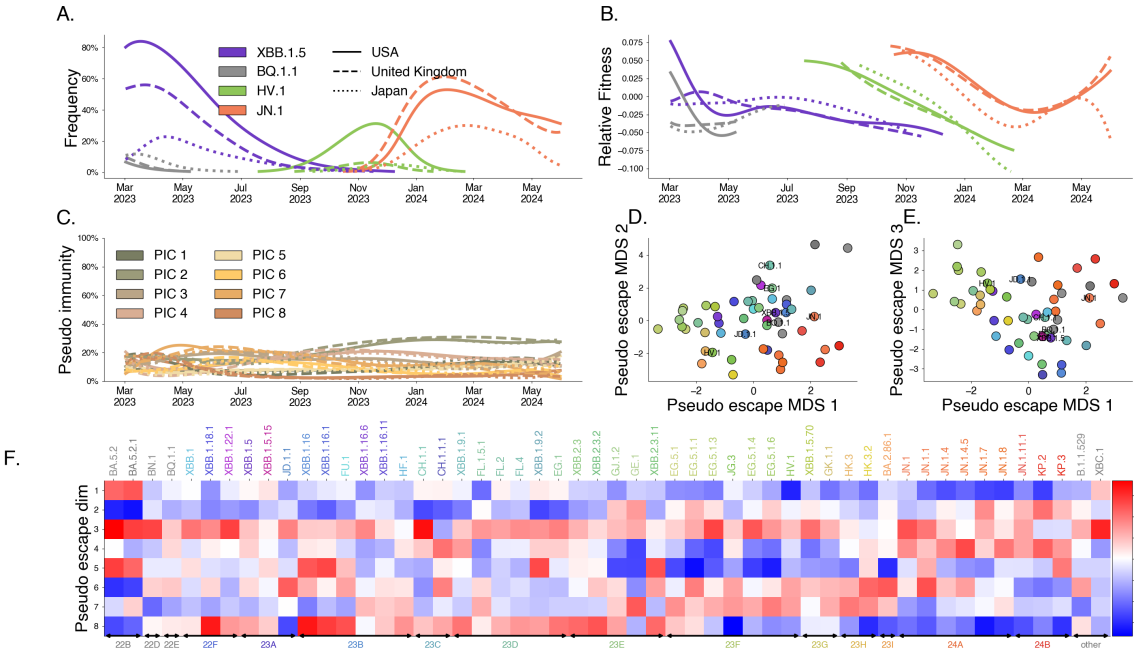

Fig. 5. Latent factor models of immunity describe variant dynamics. We fit our latent immunity factor model for $D = 8$ pseudo-immune groups using only SARS-CoV-2 sequence count data. A. Variant frequency. Lines are colored to show 4 variants of interest (of 53 total variants) with the style of the line denoting 3 countries of interest (of 18 total countries). B. Estimated relative fitness for selected variants and countries. These variant-specific relative fitnesses are similar across countries, but not identical. C. Estimated pseudo-immunity cohorts (PIC) over time for multiple countries ordered by decreasing share in the first geography using sequence data alone. D, E. Dimensionality-reduced pseudo-escape rates using multidimensional scaling (MDS). F. Estimated pseudo-escape rates for each variant relative to pivot variant "other".

Our results show that closely related Pango lineages are often assigned similar pseudo-escape values. This is visible as a clustering of lineages with similar colors into similar coordinates in Fig. 5D-E suggesting our pseudo-escape values broadly align with evolutionary structure. Further, our model shows that these groups of lineages tend to target particular immune groups such as clade 24A (JN.1, JN.1.1, JN.1.4) has high pseudo-escape in dimensions 3 and 4. If immune escape is the dominant mechanism for relative fitness difference, we expect that differences in immune response between variants from serological

data would mirror differences in our pseudo-escape space.

To examine how the learned immune structure relates to subject-level serology, we projected titers from Jian et al. [7] into principal-component space then performed clustering with k-means, arriving at $k = 5$ clusters using the elbow method (Fig. S16). Our learned immune clusters yields clear separation (Fig. 6A) in PCA-space, whereas as a subset of subject-level exposure histories show overlap in titer measurements (Fig. 6B). This overlap can be seen in a co-occurrence matrix linking learned clusters to reported exposure histories (Fig. 6C). These groups split and aggregate multiple exposure histories. For example, individuals with a BA.5/BF.7 breakthrough infections split into both clusters 1 and 2, and cluster 3 contains individuals with XBB infection, JN.1 infection, and wildtype vaccine. This indicates that coarse exposure histories may be an imperfect proxy for serological phenotype, and that titer-based clusters may better capture immunological heterogeneity within exposure categories.

We next relate measured serological titers from Jian et al. [7] to our estimates of pseudo-escape that derive solely from variant frequency dynamics across countries. To assess whether pseudo-escape captures titer measurements, we compute serological titer distances as average log2 differences in titer values between pairs of variants. We compare these titer distances to distances in our pseudo-escape space (Fig. 6D), finding the distances between distinct variant pairs in the pseudo-escape space are correlated with these titer differences between variants ($R^2 = 0.402$). We bootstrap this analysis among 1,000 replicates to assess significance of this relationship (Fig. S17, $p < 0.001$). Next, we subset by exposure history and find a similar relationship in most cohorts with the stronger correlations between serological titer distance and pseudo-immune escape in general being earlier exposures (WT, BA.5) and a weaker correlations observed in cohorts with later exposures (JN.1 and XBB) (Fig. S18).

We quantified which immune clusters are preferentially targeted by circulating lineages using the model-predicted pseudo-escape $\eta_{v,\cdot}$ for each variant v to predict titers against that variant for individuals in each immune cluster and exposure history group (Fig. 6E-F). This relationship is delineated in Methods section 'Regression of pseudo-escape onto neutralization titers'.

To assess variant-level escape against each cluster, we predict an escape burden as the negative of the predicted titer as a proxy for the escape potential against individuals of a certain background. Variants within JN.1 clade show elevated escape burden in specific clusters, with immune cluster 4 standing out as the most broadly targeted. This is consistent with the pseudo-escape patterns in Fig. 5F, where JN.1-family lineages exhibit high pseudo-escape along the dominant dimensions. Further, we find that immune cluster 4 shows a strong negative escape burden again viruses in clade 23B, likely owing the fact that individuals immune cluster 4 corresponds to a subset of individuals with past XBB infection (Fig. 6G). This is further supported by the fact that among all infection histories, pseudo-escape best predicts titers in individuals with past XBB infection including those with BA.5/BF.7 breakthrough infection (Fig. S19-S20).

In short, this shows that a sequence-only latent immunity model can recover an antigenic geometry that agrees with human serology and that learned pseudo-escape values can be

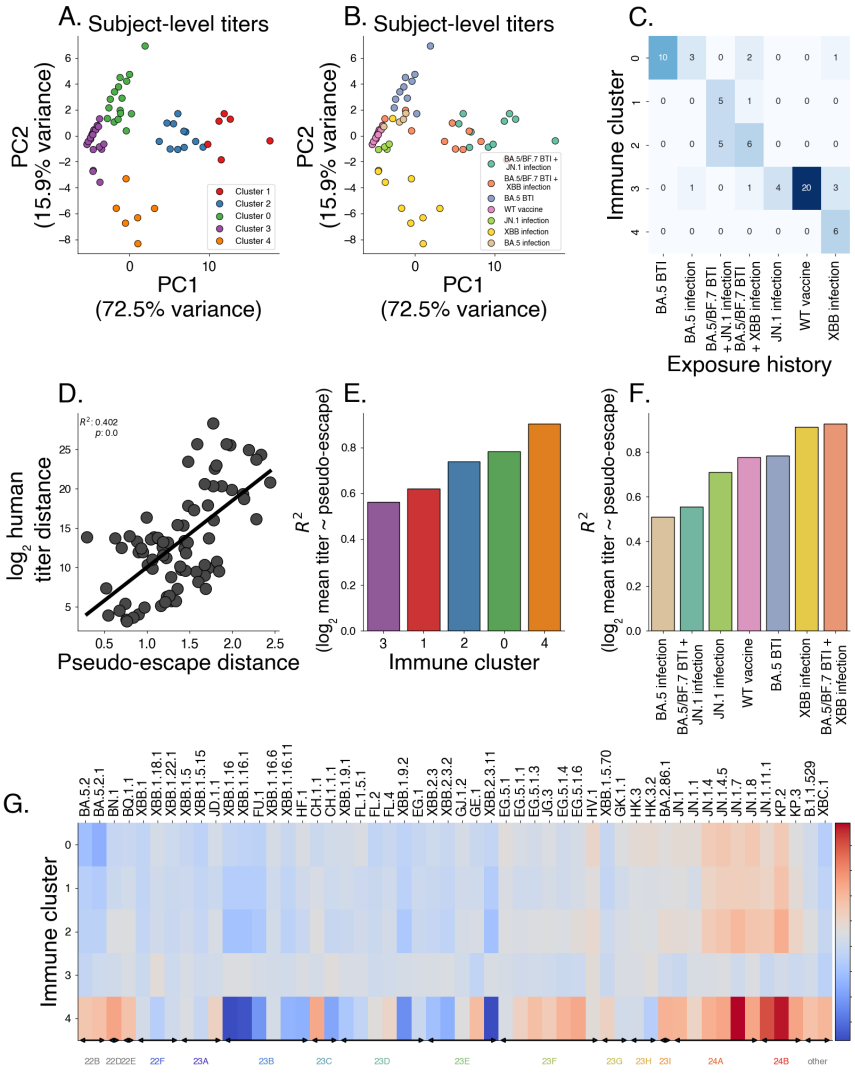

Fig. 6. Latent factor models of immunity predict titers across varying exposure histories. Using pseudo-escape values from our latent factor model (Fig. 5) and human titer data, we show that pseudo-escape values predict antigenic distance and titers. (A-B) Principal component analysis of subject-level titers, colored by learned immune clusters (A) and subject infection history (B). (C) Co-occurrence matrix between learned immune clusters and infection histories. (D) Comparing pairwise distance between variants in the pseudo-immune space to observed distances in human titer data. (E-F) Correlation between log mean titer against a variant and that variant's pseudo-escape by group. (G) Estimated escape burden by immune cluster, showing variant escape potential against individual immune clusters.

used to predict cohort-level neutralization patterns. These properties make the latent representation useful for analyzing population susceptibility and explaining geographic variation in variant success, even when contemporaneous titer data are unavailable.

This approach can be applied to other antigenically variable pathogens, such as seasonal influenza, making it broadly applicable beyond SARS-CoV-2. In fact, there is more utility for pathogens with larger geographic differences in immunity since this approach enables

to estimate the proportion of these latent immune pools in the population and how they vary geographically and over time alongside variant difference. By approximating antigenic differences using sequence data alone, this method offers for a deeper understanding of immune dynamics and how they shape variant success in the presence of immune escape. This enables an embedding similar to those from antigenic cartography but without the need for serological data and based purely on observed variant frequencies and estimated variant fitness.

Discussion

Our study demonstrates the utility of multi-strain mechanistic models in interpreting variant frequency dynamics. This enables a more detailed picture of variant success in environments with heterogeneous population immunity. Our mechanistic grounding of variant fitness allows for investigations into trade-offs between intrinsic transmissibility increase and immune escape, prediction of epidemic dynamics from sequence data alone and inference of antigenic relatedness among variants from differences in success across geographies.

In particular, our latent factor model is most easily compared to the approaches of Meijers et al. [21] and Raharinirina et al. [22] that use cross-neutralization and deep mutational scanning data respectively to parameterize variant fitness. However, our approach differs significantly in that our model does not require any data other than sequence counts for each variant over time, enabling real-time analysis of fitness and heterogeneity in population immunity before cross-neutralization and deep mutational scanning data are available.

Despite these advances, there are limitations to our approach. Long-term forecasts remain difficult, particularly as new variants with unknown fitness profiles emerge. This framework suggests that considering both the escape against individual immune backgrounds and the diversity in human immune escape is most useful for improving forecasts of relative fitness. Our models, while powerful in estimating short-term variant dynamics, rely on assumptions about transmission mechanisms that may not always hold across different pathogens or contexts. In fact, as we've shown, it's entirely possible for shifts in population immunity to change the dominant transmission mechanism.

Furthermore, the models considered here are deterministic in nature and do not explicitly model the emergence of variant viruses only the dynamics after their successful introduction. In reality, there are biological constraints on the types of variants that are produced in nature and even if there is a 'true' fitness boost, the chance for stochastic extinction of beneficial variants remains. These constraints present trouble for long-term forecasting as it will require a model of mutation or emergence, tying the potential for a variant to emerge with its potential to transmit in the current environment. Future work should focus on improving the integration of real-time genomic data with serological and epidemiological data, providing a more comprehensive understanding of variant dynamics over time.

In conclusion, our framework represents a significant advance in our understanding of viral evolution and transmission dynamics. By linking variant fitness to specific transmission

mechanisms, we provide a more nuanced and accurate prediction of how variants will spread and impact population-level epidemic growth. The selective pressure metric and latent immunity model offer new tools for public health agencies to monitor viral evolution in real time, enabling proactive intervention and insight into the variant difference and wave potential. While our work has been applied to SARS-CoV-2, the methods developed here are broadly applicable to other evolving pathogens, offering a versatile approach for improving epidemic forecasting, variant monitoring, and overall pandemic preparedness.

Materials and Methods

Correlations are insufficient for mechanism identification To assess how vaccination uptake affects the growth advantage of a variant with increased transmissibility, we simulate the spread of a more transmissible variant across populations with different initial past exposure and vaccination levels. This enables us to isolate the effects of transmissibility within different immunity landscapes, examining how relative fitness and growth advantage shift based on population vaccination coverage alone in the absence of immune escape. We begin with the 2-variant SIR model described in Supplementary Text S1. We simulate this model for 100 days with generation time $\tau = 1/\gamma = 3.0$ days, $R_{0,W} = 1.4$, $I_W(0) = 100$ individuals, $I_v(0) = 1$ individual, a 50% transmissibility increase $\rho = 0.5$, and no immune escape $\eta = 0.0$. We divide the period into early and late epidemic with the breakpoint being $t = 50$. In Fig. 3D-E, we estimate the log growth advantage for the variant in the early and full periods using a logit-linear model

$$\log\left(\frac{f_v(t)}{1 - f_v(t)}\right) = \beta t/\tau + \alpha, \tag{12}$$

where we take the model slope β to be our log growth advantage.

We repeat these simulations for a range of vaccination levels starting from 0% and ending at 65%.

Generating sequence counts We prepared sequence count data sets using the Nextstrain-curated SARS-CoV-2 sequence metadata [33] which is created using the GISAID EpiCoV database [34]. These sequences were tallied according to either their annotated Nextstrain clade or Pango lineage [12] depending on the data set to produce sequence count for each variant, for each day over the period of interest, and in each country analyzed.

Likelihood of sequence counts given frequencies The models discussed in this paper use observed counts of variant sequences to inform the underlying variant frequency in the population. This is accomplished using a multinomial likelihood, so that given count of sequences $S_v(t)$ of variant v at time t and total sequences $N(t)$ collected at time t, we have that

$$S_v(t) \sim \text{Multinomial}(N(t), f_v(t)), \tag{13}$$

where $f_v(t)$ is the frequency of variant v at time t. This is a simple model of sequence counts to frequencies and does not account for over-dispersion of sequence counts relative

to a multinomial. However, all models can be extended to estimate and account for over-dispersion by replacing the above likelihood with a Dirichlet-Multinomial likelihood.

Approximate Gaussian processes for relative fitness estimation To generate smooth non-parametric estimates of variant growth rates, we develop a Gaussian process based model for relative fitnesses. That is, we model the relative fitness for each variant over time $\lambda_v(t)$ as a multivariate normal distribution:

$$\boldsymbol{\lambda}_v \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{14}$$

$$\boldsymbol{\Sigma}_{s,t} = K_\theta(s,t), \tag{15}$$

where K_θ is a potentially parameterized kernel function. This induces a structure on the covariance of the relative fitness values over time points s and t, causing relative fitness to vary smoothly in time.

For computational efficiency, we implement a Hilbert Space Gaussian Process (HSGP) approximation instead of fitting V independent Gaussian processes. This approximation allows us to share basis functions between variants [28]. Under this approximation, the relative fitnesses are computed as

$$\lambda_v(t) \approx \sum_{j=1}^{m} S_\theta(\sqrt{\mu_j})^{1/2} \cdot \phi_j(t) \cdot \beta_j, \tag{16}$$

where S_θ is the spectral density of the kernel K_θ, μ_j and ϕ_j are the m eigenvalues and eigenfunctions of the Laplacian, and $\beta_j \sim \text{Normal}(0,1)$ [28]. Since the eigenvalues and eigenfunctions are shared across variants, this allows us to re-use values across variants, simplifying the computation to a matrix multiplication as

$$\boldsymbol{\lambda}_t = \boldsymbol{\Phi}_t \sqrt{\boldsymbol{S}_\theta} \boldsymbol{\beta}. \tag{17}$$

For the analyses in this paper, we use this approximate Gaussian process with a Matérn 5/2 kernel and shared hyperparameters across variants. We demonstrate this model for simulated data from Fig. S1 and show resulting relative fitnesses through time in Fig. S2.

Predicting epidemic growth rate from selective pressure The derivation of the selective pressure metric shows that the selective pressure can be a useful tool in predicting the epidemic growth rate. To develop a predictive model of epidemic growth rate using selective pressure, we begin by generating estimates of selective pressure and epidemic growth rate from a period with high sequencing and case surveillance.

We take sequence count and case count data from all states in the United States between January 2021 and November 2022. State-level daily case counts were obtained from USAFacts downloaded on August 7, 2024 at https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

Using the sequence counts, we compute selective pressure estimates from relative fitness and frequencies estimated with our approximate Gaussian process relative fitness model.

From the case data, we derive the empirical growth rate using a 14-day moving average on case counts $\hat{C}_t$ and computing the empirical growth rate as $\hat{r}_t = \log(\hat{C}_t) - \log(\hat{C}_{t-1})$. We then use the past 28 days of selective pressure to predict the empirical growth rate.

We use a gradient boosting regressor model which is fit using a mean absolute error loss function. This model was selected as it achieved the minimal error via time series cross-validation averaged across 10 splits among candidate models (Fig. S10). The candidate models include linear regression, ridge regression, Lasso regression, random forests, and gradient-boosted trees as implemented in scikit-learn [35]. We additionally tune the hyperparameters of this model using grid search cross-validation.

We validate our model by comparing our predicted epidemic growth rates to held-out case data for US states, South Africa, South Korea, the United Kingdom, and additionally to estimates of the epidemic growth rates in England derived from data from the Office for National Statistics (ONS) Coronavirus Infection Survey [32]. Estimates of prevalence from the ONS Infection Survey were obtained for January 2022 to September 2022 from www.ons.gov.uk/peoplepopulationandcommunity/ healthandsocialcare/conditionsanddiseases/datasets/coron
Epidemic growth rates are computed on this data in the same way as the state-level analysis.

Latent immune factor model We show that relative fitness dynamics can be explained by low-dimensional immunity when transmission dynamics are described with compartmental models (Supplementary Text S1). This motivates a model to learn this low-dimensional structure that is inspired by latent-factor models. We start by assuming that the relative fitness of variant v at time t and in geographic location g can be described by D latent factors so that

$$\lambda_v^g(t) = \sum_{d=1}^{D} \eta_{v,d}\, \phi_d^g(t). \tag{18}$$

As the structure here resembles Equation 43, we call $\eta_{v,d}$ "pseudo-escape" of variant v from group d and ϕ_d^g "pseudo-immunity" group d in geographic location g. To make this more consistent with our intuition here, we model ϕ_d^g to be in $[0, 1]$ and model it as smoothly varying in time. We model $\text{logit}(\phi_d^g)$ using 4th order splines with 6 knots placed uniformly over the time period modeled. Though we choose to model these latent factors with splines, other models would work here. For example, one alternative would be the approximate Gaussian processes described above. Additionally, in order to ensure identifiability of the parameter estimates, we fix some base variant v^* which fitness is defined relative to, so that $\eta_{v^*,d} = 0$ for all $1 \leq d \leq D$. For the same reason, we fix the order of components, so that the components are numbered in decreasing order by their share in the arbitrarily defined base geography.

We apply this model to SARS-CoV-2 sequence counts in the period between March 2023 to March 2024 for 14 countries. To access the necessary number of immune dimensions, we vary the number of immune dimensions between $D = 2$ to $D = 12$. Looking at the loss for the latent factor model for increasing D, we choose $D = 8$ for our primary analysis by

noting the point at which the loss function seems to stagnate with increasing D i.e. the "elbow" method (Fig. S11).

We compare the distances between variant pairs in our estimated pseudo-escape space to distances in log2 titer. Using human titer data from Jian et al [7], we compute neutralization titer distances as the average of differences in log2 neutralization titers between pairs of variants for a cohort of individuals. This analysis is repeated among 1,000 bootstrapped samples to create a distribution of R^2 values (Fig. S17). Additionally, we subset this by exposure history and repeat this analysis to find which exposure groups best explain distances in pseudo-escape space (Fig. S18).

Regression of pseudo-escape onto neutralization titers To assess the relationship between our estimated pseudo-escape values and empirical neutralization titers, we use human titer data from Jian et al. [7]. For each exposure group, we have a set of neutralization titers measured against multiple variants. Our goal is to determine whether variation in titers across variants can be explained by the pseudo-escape values inferred from our latent immune factor model.

Let $t_{i,v}$ denote the neutralization titer of individual i against variant v. For each exposure group G, we first aggregate titers by computing the group-level mean and applying a $\log_2$ transformation with a pseudocount of one, i.e.

$$y_v^G = \log_2 \left(\text{mean}_{i \in G}(t_{i,v}) + 1 \right). \tag{19}$$

Let $\eta_v \in \mathbb{R}^D$ denote the pseudo-escape vector of variant v estimated from the latent immune factor model across D latent factors. We then fit an ordinary least squares (OLS) model separately for each exposure group,

$$y_v^G = \beta_0 + \boldsymbol{\beta}^\top \eta_v + \varepsilon_{g,v}, \tag{20}$$

where β_0 is an intercept term and $\boldsymbol{\beta}$ is a vector of regression coefficients mapping latent factors to predicted group-level titers. This yields a coefficient of determination R_g^2 for each group, quantifying how well variation in pseudo-escape explains variation in group-level titers.

To assess statistical significance, we performed a permutation test in which the association between variants and their pseudo-escape vectors was randomly permuted 1,000 times. For each permutation, we re-fit the OLS model and recorded the resulting R_g^2, yielding a null distribution under the hypothesis that pseudo-escape and titers are unrelated. The empirical p-value for each group was computed as the fraction of permutations with R_g^2 greater than or equal to the observed R_g^2. This procedure controls for the correlation structure in the titer data while directly testing whether the inferred pseudo-escape values have predictive power for neutralization measurements.

We summarize the results across groups by reporting both the observed R^2 values and the permutation-derived p-values (Fig. S19 and S20). This analysis identifies exposure groups where pseudo-escape provides a statistically significant explanation of the neutralization response profile.

Acknowledgements

We thank Ivana Bozic, Betz Halloran, Mark Kot and Erick Matsen, as well as members of the Bedford Lab for their feedback on this work. We gratefully acknowledge all data contributors, ie 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, on which this research is based. We have included an acknowledgements table in the associated GitHub repository under `data/final_acknowledgements_gisaid.tsv.xz`.

Funding

This work is supported by NIH NIGMS award R35 GM119774 to TB and a Howard Hughes Medical Institute COVID-19 Collaboration Initiative award to TB. MDF is an ARCS Foundation scholar and was supported by the National Science Foundation Graduate Research Fellowship Program under grant No. DGE1762114. TB is a Howard Hughes Medical Institute Investigator.

Author contributions

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

Competing interests

All authors declare no competing interests.

Data and materials availability

Source code used to generate figures, model implementations, and sequence count data are available at github.com/blab/relative-fitness-mechanisms.

References

1. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, et al. (2021) Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 592: 438–443.
2. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, et al. (2021) Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 593: 266–269.
3. Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, et al. (2023) SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nat Rev Microbiol 21: 162–177.
4. Cao Y, Wang J, Jian F, Xiao T, Song W, et al. (2021) Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature 602: 657–663.
5. Cao Y, Jian F, Wang J, Yu Y, Song W, et al. (2022) Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. Nature 614: 521–529.

6. Bekliz M, Essaidi-Laziosi M, Adea K, Hosszu-Fellous K, Alvarez C, et al. (2024) Immune escape and replicative capacity of Omicron lineages BA.1, BA.2, BA.5.1, BQ.1, XBB.1.5, EG.5.1 and JN.1.1. bioRxiv 2024.02.14.579654.

7. Jian F, Feng L, Yang S, Yu Y, Wang L, et al. (2023) Convergent evolution of SARS-CoV-2 XBB lineages on receptor-binding domain 455–456 synergistically enhances antibody evasion and ACE2 binding. PLOS Pathog 19: e1011868.

8. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, et al. (2014) Integrating influenza antigenic dynamics with molecular evolution. eLife 3: e01914.

9. Kistler KE, Bedford T (2023) An atlas of continuous adaptive evolution in endemic human viruses. Cell Host Microbe 31: 1898–1909.

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

11. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, et al. (2021) Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet 53: 809–816.

12. Aksamentov I, Roemer C, Hodcroft EB, Neher RA (2021) Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw 6: 3773.

13. Annavajhala MK, Mohri H, Wang P, Nair M, Zucker JE, et al. (2021) Emergence and expansion of SARS-CoV-2 B.1.526 after identification in New York. Nature 597: 703–708.

14. Piantham C, Ito K (2022) Predicting the time course of replacements of SARS-CoV-2 variants using relative reproduction numbers. medRxiv 2022.03.30.22273218.

15. Figgins MD, Bedford T (2022) SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers. medRxiv : 2021.12.09.21267544.

16. Susswein Z, Johnson KE, Kassa R, Parastaran M, Peng V, et al. (2023) Leveraging global genomic sequencing data to estimate local variant dynamics. medRxiv : 2023.01.02.23284123.

17. Lefrancq N, Duret L, Bouchez V, Brisse S, Parkhill J, et al. (2023) Learning the fitness dynamics of pathogens from phylogenies. medRxiv : 2023.12.23.23300456.

18. Abousamra E, Figgins M, Bedford T (2024) Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency. PLoS Comput Biol 20: e1012443.

19. van Dorp C, Goldberg E, Ke R, Hengartner N, Romero-Severson E (2022) Global estimates of the fitness advantage of SARS-CoV-2 variant Omicron. Virus Evolution 8: veac089.

20. Dadonaite B, Brown J, McMahon TE, Farrell AG, Asarnow D, et al. (2024) Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. Nature 631: 617–626.

21. Meijers M, Ruchnewitz D, Eberhardt J, Łuksza M, Lässig M (2023) Population immunity predicts evolutionary trajectories of sars-cov-2. Cell 186: 5151–5164.e13.

22. Raharinirina NA, Gubela N, Börnigen D, Smith MR, Oh DY, et al. (2025) Sars-cov-2 evolution on a dynamic immune landscape. Nature 639: 196–204.

23. Gog JR, Grenfell BT (2002) Dynamics and selection of many-strain pathogens. Proceedings of the National Academy of Sciences 99: 17209–17214.

24. Bedford T, Rambaut A, Pascual M (2012) Canalization of the evolutionary trajectory of the human influenza virus. BMC Biol 10: 38.

25. Kistler KE, Bedford T (2021) Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. eLife 10: e64509.

26. Eguia RT, Crawford KHD, Stevens-Ayers T, Kelnhofer-Millevolte L, Greninger AL, et al. (2021) A human coronavirus evolves antigenically to escape antibody immunity. PLOS Pathog 17: e1009453.

27. Görtler J, Kehlbeck R, Deussen O (2019) A visual exploration of gaussian processes. Distill .

28. Riutort-Mayol G, Bürkner PC, Andersen MR, Solin A, Vehtari A (2023) Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. Statistics and Computing 33.1: 17.

29. Ewens W (1989) An interpretation and proof of the fundamental theorem of natural selection. Theor Popul Biol 36: 167–180.

30. Ewens WJ (2024) The fundamental theorem of natural selection: the end of a story. Evolution 78: 803–808.

31. Mustonen V, Lässig M (2010) Fitness flux and ubiquity of adaptive evolution. Proc Natl Acad Sci USA 107: 4248–4253.

32. Pouwels KB, House T, Pritchard E, Robotham JV, Birrell PJ, et al. (2021) Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. Lancet Public Health 6: e30–e38.

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

34. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, et al. (2021) GISAID's role in pandemic response. China CDC weekly 3: 1049.

35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12: 2825–2830.

36. Łuksza M, Lässig M (2014) A predictive fitness model for influenza. Nature 507: 57–61.

37. Huddleston J, Barnes JR, Rowe T, Xu X, Kondor R, et al. (2020) Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. eLife 9: e60067.

38. Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, et al. (2022) Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science 376: 1327–1332.

39. Wallinga J, Lipsitch M (2006) How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B 274: 599–604.

40. Lazebnik T, Bunimovich-Mendrazitsky S (2022) Generic approach for mathematical model of multi-strain pandemics. PLOS ONE 17: e0260683.

Supplementary Information for

Frequency dynamics predict viral fitness, antigenic relationships and epidemic growth

Supplementary Text

S1 Exponentially growing populations to frequency dynamics

We consider a viral population consisting of V exponentially-growing variant viruses each with prevalence I_v. Defining the time-varying growth rate for the prevalence of variant v as $r_v(t)$, we can model the prevalence using an ordinary differential equation

$$\frac{dI_v}{dt} = r_v(t)I_v(t), \quad v = 1, 2, \ldots, V. \tag{21}$$

The above differential equation has a known solution in terms of the integral of the time-varying growth rate and initial prevalence,

$$I_v(t) = I_v(0) \exp\left(\int_0^t r_v(s)ds\right), \tag{22}$$

where $I_v(0)$ is the initial prevalence of variant v.

Now turning to the frequency dynamics of the population, we write the frequency of variant v in the population as $f_v(t) = I_v(t)/\sum_{u=1}^{V} I_u(t)$. This allows us to derive an ODE for variant frequency in terms of the variant growth rates using the quotient rule for differentiation

$$\frac{df_v}{dt} = f_v\left(\sum_{u=1}^{V}[r_v(t) - r_u(t)]f_u\right) \tag{23}$$

$$= f_v\left(r_v(t) - \sum_{u=1}^{V} r_u(t)f_u\right). \tag{24}$$

This system of differential equations resembles a logistic growth equation and can be shown to have the following solution in terms of the initial frequencies $f_v(0)$ and the variant growth rates

$$f_v(t) = \frac{f_v(0)\exp(\int_0^t r_v(s)ds)}{\sum_{u=1}^{V} f_u(0)\exp(\int_0^t r_u(s)ds)}. \tag{25}$$

The above representation of the variant frequency will serve as a centerpiece for many of the arguments to follow. We see that by tracking the rate at which variant viruses are spreading, we can construct the corresponding frequency dynamics without knowing the absolute prevalence of any variant.

Relative frequency and relative fitness Using the above equation for the variant frequencies, we can write the relative frequency of variant v over u as $x_{v,u}(t) = f_v(t)/f_u(t)$ to see

$$x_{v,u}(t) = \frac{f_v(t)}{f_u(t)} = \frac{f_v(0)}{f_u(0)} \exp\left(\int_0^t [r_v(s) - r_u(s)]ds\right) \tag{26}$$

$$= x_{v,u}(0) \exp\left(\int_0^t \lambda_{v,u}(s)ds\right). \tag{27}$$

Notice this relative frequency change depends on the initial relative frequencies and the *relative fitness* $\lambda_{v,u}(t) = r_v(t) - r_u(t)$ of v over u. This relative fitness has the same units as the exponential growth rate (e.g. per day). Using the definition of relative fitness, we can notice that

$$\lambda_{v,u}(t) = r_v(t) - r_u(t) = \frac{d}{dt}\left[\log\left(x_{v,u}(t)\right)\right] = -\lambda_{u,v}(t). \tag{28}$$

We can see that there is a symmetry in the relative fitnesses and that the associated frequency dynamics depend on the differences between relative fitnesses. This suggests that absolute fitness (in terms of the growth of infections) may not be inferable from frequencies alone. This definition of relative fitness becomes essential in describing various existing modeling approaches for frequency dynamic data and motivates possible extensions since we can represent these models as having the form:

$$f_v(t) = \frac{f_v(0)\exp(\int_0^t \lambda_{v,v^*}(s)ds)}{\sum_{u=1}^V f_u(0)\exp(\int_0^t \lambda_{u,v^*}(s)ds)}, \tag{29}$$

where the relative fitness of v is expressed as relative to an arbitrary pivot variant v^*.

Cumulative relative-fitness and frequency change Above we saw that within our framework frequency change over time intervals depends only on the cumulative relative fitness over time intervals $\Lambda_{v,u}(0,t) = \int_0^t \lambda_{v,u}(s)ds$. We can then characterize approaches for modeling frequency change in terms of how they represent, estimate, and forecast these relative fitnesses. This framework includes various existing methods for analyzing frequency data such as the seasonal influenza forecasting models of Lässig and Łuksza [36] and Huddleston et al [37], multinomial logistic regression for frequency estimation [13] and the SARS-CoV-2 mutational fitness model of Obermeyer et al [38].

Though this framework can be used to describe existing statistical methods for frequency modeling, it is also applicable to traditional compartmental models of epidemics. In fact, applying these ideas to compartmental models enables to see how mechanistic assumptions on the transmission process determine relative fitness of variant viruses.

Two-strain SIR For simplicity, we will begin by analyzing a two-strain SIR model in which a variant virus v can differ from wildtype virus wt by increased intrinsic transmissibility (via η_T) and immune escape against wild-type immunity (via η_E). This gives a system of 5 ordinary differential equations

$$\frac{dS}{dt} = -\beta S I_W - \beta(1+\rho) S I_v, \tag{30}$$
$$\frac{dI_W}{dt} = \beta S I_W - \gamma I_W, \tag{31}$$
$$\frac{dI_v}{dt} = \beta(1+\rho) S I_v + \beta(1+\rho)\eta\phi_W I_v - \gamma I_v, \tag{32}$$
$$\frac{d\phi_W}{dt} = \gamma I_W - \beta(1+\rho)\eta\phi_W I_v, \tag{33}$$
$$\frac{d\phi_v}{dt} = \gamma I_v, \tag{34}$$

where I_W denotes wild-type prevalence, I_v denotes variant prevalence and ϕ_W denotes immunity derived from wild-type infection. In this model, the variant virus can infect both susceptible individuals S and individuals with immunity to wild-type virus ϕ_W. Increased intrinsic transmissibility increases the baseline transmission rate from β in wild-type to $\beta(1+\rho)$ in the variant virus and immune escape increases the transmission rate against those with wildtype immunity, so that the at-risk population is $\eta\phi_W$.

Writing that $r_W(t) = \beta S - \gamma$ and $r_v(t) = \rho\beta S + \beta(1+\rho)\eta\phi_W - \gamma$, we can then write the relative fitnesses as:

$$\lambda_{v,W}(t) = \rho\beta S(t) + (1+\rho)\eta\beta\phi_W(t). \tag{35}$$

From this representation of relative fitness, we can see that given fixed increases to overall transmission ($\rho > 0$) or immune escape ($\eta > 0$), the observed fitness boost at the level of variant relative fitness still depends on the proportion of the population at risk for infection.

n-strain SIR This model can also be extended to an n-strain SIR model where each variant strain v_i with $2 \le i \le n$ is described by its own advantage parameters $\theta_i = (\rho^{(i)}, \eta^{(i)})$ relative to the wildtype ($\theta_W = \theta_1 = (0,0)$)

$$\frac{dS}{dt} = -\beta S I_W - \sum_i \beta(1+\rho_i) S I_{v_i} \tag{36}$$
$$\frac{dI_W}{dt} = \beta S I_W - \gamma I_W \tag{37}$$
$$\frac{dI_{v_i}}{dt} = \beta(1+\rho_i) S I_{v_i} + \beta(1+\rho_i)\eta_i\phi_W I_{v_i} - \gamma I_{v_i} \tag{38}$$
$$\frac{d\phi_W}{dt} = \gamma I_W - \sum_i \beta(1+\rho_i)\eta_i\phi_W I_{v_i} \tag{39}$$
$$\frac{d\phi_{v_i}}{dt} = \gamma I_{v_i}, \quad i \in \{2,\dots,n\}. \tag{40}$$

In this formulation, the variant viruses compete only for susceptible population and those with previous wild-type infection. This formulation can be generalized to allow for competition between all variants for any exposure history and will be discussed in the following sections. In Fig. S1, we implement and simulate a 3-strain model with wildtype as

above, an escape variant E with $\theta_2 = (0, \eta)$, and a transmissibility increase variant T with $\theta_3 = (\rho, 0)$.

Models of immune escape against heterogeneous backgrounds We'll now consider a model where all hosts are assumed to fall into one of B immune backgrounds ϕ_b for $b = 1, \ldots, B$. We assume that infection by each variant v then leaves recovered hosts in the corresponding immune background of the most recent infection b_v. Variant transmission then occurs via immune escape against a background leading to a matrix of escape rates $\boldsymbol{\eta} = \eta_{v,b}$ for variants v and background b.

We can then write the system of ordinary differential equations as

$$\frac{dI_v}{dt} = \beta \sum_{1 \leq b \leq B} \eta_{v,b} \phi_b I_v - \gamma I_v, \quad v = 1, \ldots, V \tag{41}$$

$$\frac{d\phi_b}{dt} = -\beta \sum_{1 \leq v \leq V} \eta_{v,b} \phi_b I_v + \sum_{v:\ b_v = b} \gamma I_v. \tag{42}$$

With this model, susceptible and recovered compartments in the standard SIR model can be thought of as immune backgrounds. This allows us to represent the standard SIR model as $S = \phi_S$, $I = I_W$, $R = \phi_W$ and $\eta_{W,S} = 1, \eta_{W,W} = 0$ and $b_W = W$. We can also think of the two-strain SIR with $\rho = 1$ as a special case of this model where we set $S = \phi_S, \eta_{W,S} = 1, \eta_{W,W} = 0, \eta_{v,S} = 1, \eta_{v,W} = \eta$ and keep all other parameters the same.

With this formulation of immune escape, we can then write the relative fitnesses in terms of the escape rates $\eta_{v,b}$ and the immune background proportions ϕ_b as

$$\lambda_{v,u}(t) = \beta \sum_{1 \leq b \leq B} (\eta_{v,b} - \eta_{u,b}) \phi_b(t). \tag{43}$$

Under this model of immune escape, we can see relative fitness among variants can be decomposed into differences in immune escape among immune backgrounds within a population. Due to the dependence here on the proportion of each immune background in determining fitness, this suggests that the overall distribution of susceptibility to strains is potentially an important consideration when translating individual-level measures of immune escape to population-level estimates of variant fitness. Understanding the size and complexity of this immune space may therefore be useful for parameterization and forecasting of variant frequencies. However, the extent to which modeling this complexity affects estimates of relative fitness also depends on how quickly the distribution of immune backgrounds change i.e. $\frac{d\phi_b}{dt}$.

Though the derivation above uses a simplified model of using most recent infection to sort individuals into an immune group, we show that a more complicated model that accounts for the entire exposure history of the host also gives a similar decomposition to relative fitness in Supplementary Text S3.

S2 Revisiting existing models for frequency growth

Using the theory developed for exponentially-growing variant populations, we now re-visit existing methods for modeling viral frequency dynamics.

Multinomial Logistic Regression We begin with multinomial logistic regression (MLR) with fixed relative fitness. This model can be written as

$$f_v(t) = \frac{f_v(0)\exp(\lambda_v t)}{\sum_u f_u(0)\exp(\lambda_u t)}, \tag{44}$$

where $f_v(t)$ is the frequency of variant v at time t and λ_v is the relative fitness of variant v. This provides estimates of the relative fitness compared to some reference strain u^* for which $\lambda_{u^*} = 0$. In this model, initial frequencies $f_v(0)$ and relative fitness λ_v are estimated from frequency dynamics. Converting this estimate to an estimate of transmission advantage (relative effective reproduction number) requires assuming a Dirac delta distribution of the generation time [39].

Comparing this to equation 35, we can see this model of fixed relative fitness results from assuming that the at-risk populations are constant over-time. This assumption is useful since it requires no outside knowledge of the at-risk population and relative infection rates, though this may be less useful for longer forecasts or when there is large turnover in at-risk populations due to infection.

Fitness models of seasonal influenza Motivated by the observed antigenic evolution of seasonal influenza, Lässig and Łuksza [36] and Huddleston et al [37] approximate the cumulative relative fitness between influenza seasons on the level of individual strains as

$$\Lambda_{v,u}(t+\Delta t, t) = (\beta_1 x_{v,1} + \cdots + \beta_p x_{v,p})\Delta t = (\boldsymbol{\beta}\cdot\boldsymbol{x}_v)\Delta t, \tag{45}$$

where the relative fitness is determined by strain-specific predictors $\boldsymbol{x}_v$ and the regression parameter $\boldsymbol{\beta}_v$ are estimated.

This formulation fits neatly into the framework we've developed as the cumulative fitness here can be written as the integral of a relative fitness $\lambda_{v,u} = \boldsymbol{\beta}\cdot\boldsymbol{x}_v$ over the time period of interest:

$$\Lambda_{v,u}(t+\Delta t, t) = \int_t^{t+\Delta t} \lambda_{v,u}(s)ds \;= \int_t^{t+\Delta t} (\boldsymbol{\beta}\cdot\boldsymbol{x}_v)ds. \tag{46}$$

Therefore, these models can be thought as regression-based predictors of relative fitness where frequency and external covariates contribute to estimated relative fitness.

S3 Relative fitness for full immune history models

We show that the simple background model is consistent with an expanded immune history model. Beginning with the model from Lazebnik and Bunimovich-Mendrazitsky 2022 [40], we consider the differential equation for the individuals with strain infection history J and current infecting strain i $R_J I_i$

$$\frac{dR_J I_i}{dt} = -\gamma_{J,i} R_J I_i + \beta_{J,i} R_J \sum_{K \in P(M), i \notin K} R_K I_i \tag{47}$$

where $P(M)$ is the collection of all finite infection histories.

Here, infection can occur from any individual infected with strain i assuming their past immune history does not include i and the infected are any recovered individual with immune history J R_J. To compute the strain growth rate, we can sum over all possible immune histories for individuals infected with strain i, so that

$$\begin{aligned}
\frac{dI_i}{dt} &= \sum_{J \in P(M), i \notin J} \frac{dR_J I_i}{dt} && (48)\\
&= -\gamma_i I_i + \sum_{J \in P(M), i \notin J} \beta_{i,J} R_J \sum_{K \in P(M), i \notin K} R_K I_i && (49)\\
&= -\gamma_i I_i + \sum_{J \in P(M), i \notin J} \beta_{i,J} R_J I_i && (50)\\
&= \left(-\gamma_i + \sum_{J \in P(M), i \notin J} \beta_{i,J} R_J \right) I_i && (51)\\
&= \left(-\gamma + \beta \sum_{J \in P(M), i \notin J} \eta_{i,J} R_J \right) I_i. && (52)\\
& && (53)
\end{aligned}$$

Assuming that the transmission rate can be decomposed as a base transmission rate β and a strain i and immune history J specific escape rate $\eta_{i,J}$ and that the recovery rate is constant, we notice this is identical to our previous immune background model. Therefore, our relative fitnesses simplify to

$$\lambda_{i,j} = \beta \sum_{B \in P(M)} (\eta_{i,B} - \eta_{j,B}) R_B, \tag{54}$$

where for simplicity we define $\eta_{v,B} = 0$ if $v \in B$.

S4 Selective pressure and contribution to epidemic growth rates

In this section, we derive our selective pressure metric $\psi(t)$ and show how it contributes to the overall epidemic growth rate in the population.

Beginning again from our assumption of inhomogeneous exponential growth, we can write a differential equation for the total prevalence $I(t) = \sum_v I_v(t)$,

$$\frac{dI}{dt} = \sum_v \frac{dI_v}{dt} = \sum_v r_v(t) I_v(t) \tag{55}$$

$$= \left(\sum_v r_v(t) f_v(t) \right) I(t), \tag{56}$$

where we've used that $I_v(t) = f_v(t) I(t)$. This allows us to see that $\overline{r}(t) = \sum_v r_v(t) f_v(t)$ is the average growth rate of the prevalence. Re-writing the average in terms of some base exponential growth rate and the relative fitnesses so that $r_v(t) = \lambda_v(t) + r_W(t)$, we get that $\overline{r}(t) = \sum_v \lambda_v(t) f_v(t) + r_W(t)$. We can simplify this by writing $\overline{r}(t) = \overline{\lambda}(t) + r_W(t)$ where $\overline{\lambda}(t) = \sum_v \lambda_v(t) f_v(t)$ is the mean fitness of the population. We can now look at the rate of change in the average growth rate by taking its derivative

$$\frac{d\overline{r}}{dt} = \frac{dr_W}{dt} + \frac{d\overline{\lambda}}{dt} \tag{57}$$

$$= \frac{dr_W}{dt} + \sum_v \left[\frac{d\lambda_v}{dt} f_v(t) + \lambda_v(t) \frac{df_v}{dt} \right] \tag{58}$$

$$= \frac{dr_W}{dt} + \sum_v \left[\frac{d\lambda_v}{dt} f_v + \lambda_v f_v (\lambda_v - \overline{\lambda}) \right] \tag{59}$$

$$= \frac{dr_W}{dt} + \sum_v \frac{d\lambda_v}{dt} f_v(t) + \sum_v \lambda_v (\lambda_v - \overline{\lambda}) f_v \tag{60}$$

$$= \frac{dr_W}{dt} + \sum_v \frac{d\lambda_v}{dt} f_v(t) + \sum_v \lambda_v^2 f_v - \overline{\lambda} \sum_v \lambda_v f_v \tag{61}$$

$$= \frac{dr_W}{dt} + \mathbb{E}_{f(t)} \left[\frac{d\lambda_v}{dt} \right] + \mathbb{V}_{f(t)}[\lambda_v]. \tag{62}$$

Here, we've written the last line in terms of expectations relative to sampling according to the frequency distribution. This shows us that the change in the average growth rate of the epidemic can be written in terms of the growth rate of the pivot category, the mean rate of change in the relative fitness, and the variance of the relative fitnesses. We will call terms which can be computed in terms of quantities derived from frequencies alone the selective pressure

$$\psi(t) = \mathbb{E}_{f(t)} \left[\frac{d\lambda_v}{dt} \right] + \mathbb{V}_{f(t)}[\lambda_v]. \tag{63}$$

We can use this idea to directly write the prevalence in terms of the selective pressure and the base growth rate. First, we define a cumulative selective pressure

$$\Psi(t) = \int_0^t \psi(s) ds. \tag{64}$$

We can then use this to reconstruct the relative incidence

$$\frac{I(t)}{I(0)} = \exp\left(\int_0^t \overline{r}(s)ds\right) \tag{65}$$

$$= \exp\left(\int_0^t [r_W(s) + \Psi(s)]ds\right). \tag{66}$$

Supplementary Figures

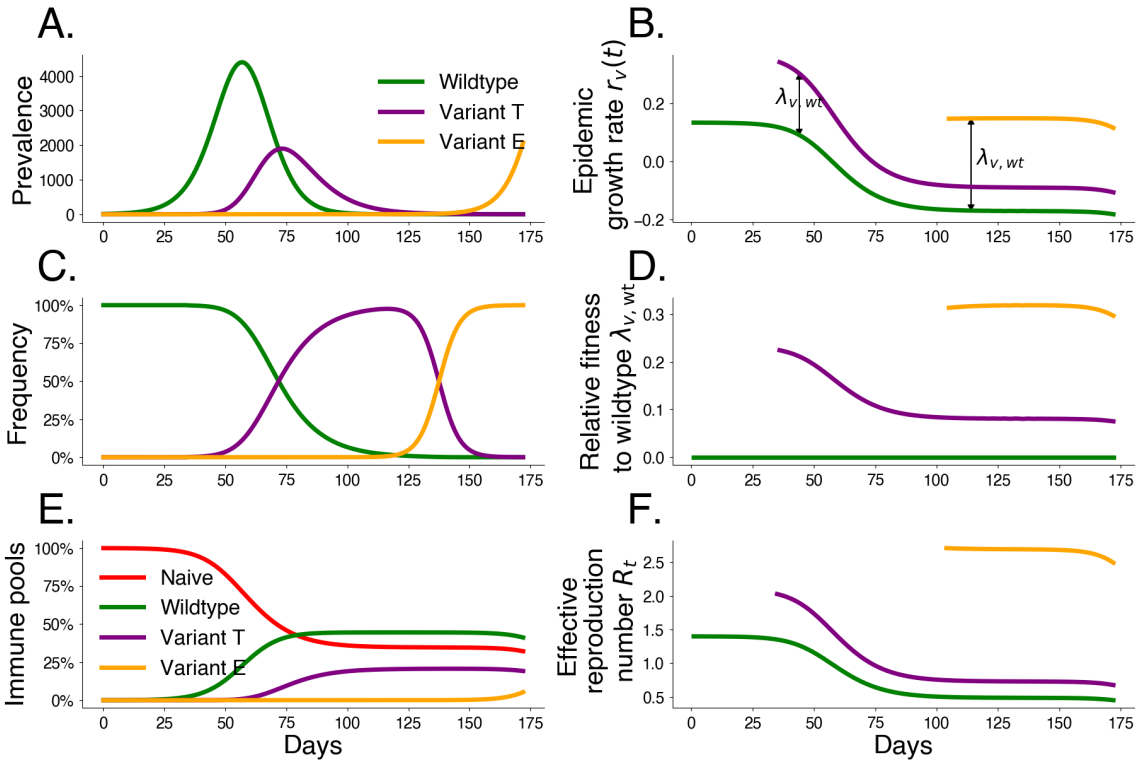

Fig. S1. Simulated variant dynamics in a mechanistic model. Mechanistic transmission models constrain variant frequency dynamics by specifying a functional form for relative fitnesses. Simulations of a three-variant model including wildtype W, an intrinsic transmission variant T, and an immune escape variant E show the relationship between population-level transmission and selection. We begin the simulation with initial wildtype prevalence $I_W(0) = 1$, effective reproduction number $R_{0,W} = 1.4$, and duration of infection $1/\gamma = 3.0$ days. We introduce transmissibility variant T at $t = 20$ with frequency $f_T(20) = 10^{-5}$ and a 50% increase in transmissibility $\rho = \rho_T = 0.5$. We introduce escape variant E at $t = 70$ with frequency $f_E(70) = 10^{-6}$ that infects 5% of hosts possessing wildtype immunity $\eta = \eta_E = 0.05$. A. Prevalence I by variant. B. Exponential growth rate r by variant. C. Variant frequency f. D. Fitness relative to wildtype λ. E. Underlying immune pools. F. Effective reproduction number R_t by variant.

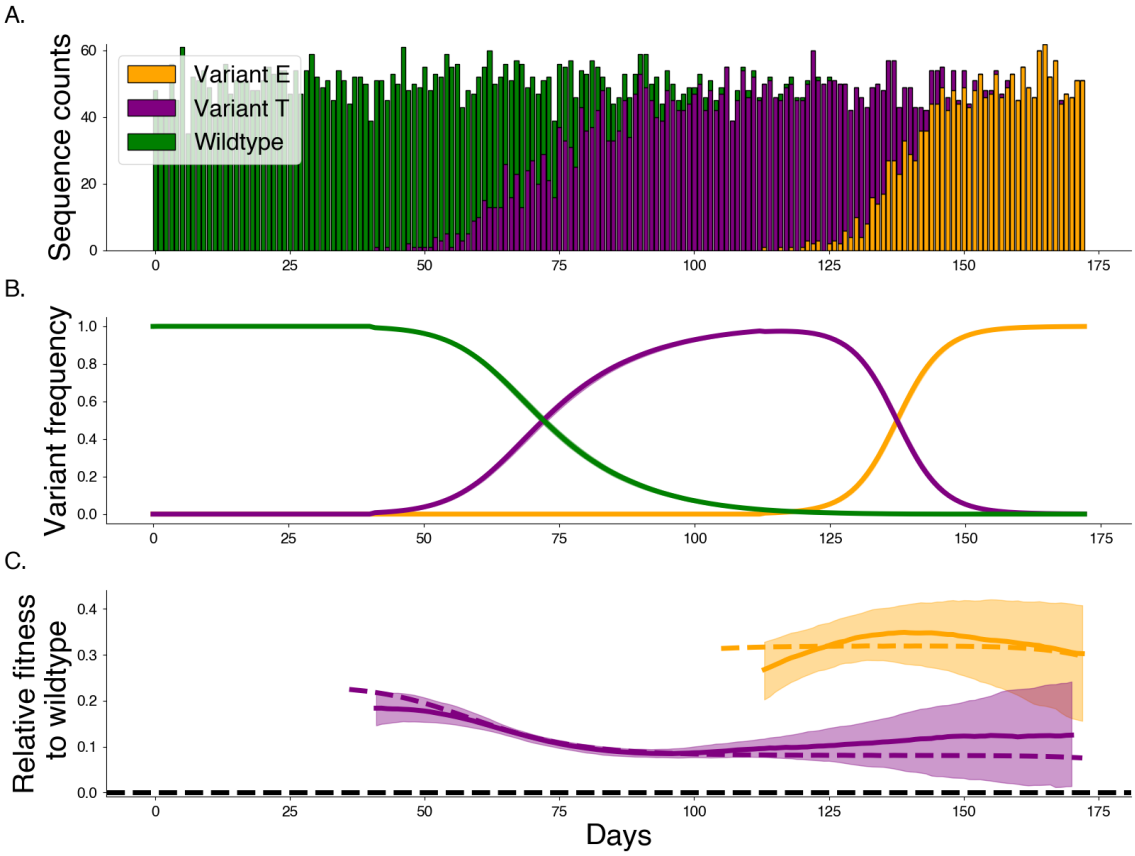

Fig. S2. Estimating relative fitness with Gaussian processes. Gaussian processes allow us a non-parametric estimate of the relative fitness for variants through time. This figure uses Gaussian processes to model the 3 variant example shown in Fig. S1. A. Synthetic sequence counts generated using a multinomial distribution with frequencies from Fig. S1C. B. Frequencies and posterior frequencies according to Gaussian process model. Intervals show the 80% credible interval. C. Posterior relative fitnesses. Intervals show the 80% credible interval. Dashed line shows true relative fitnesses from underlying mechanistic model.

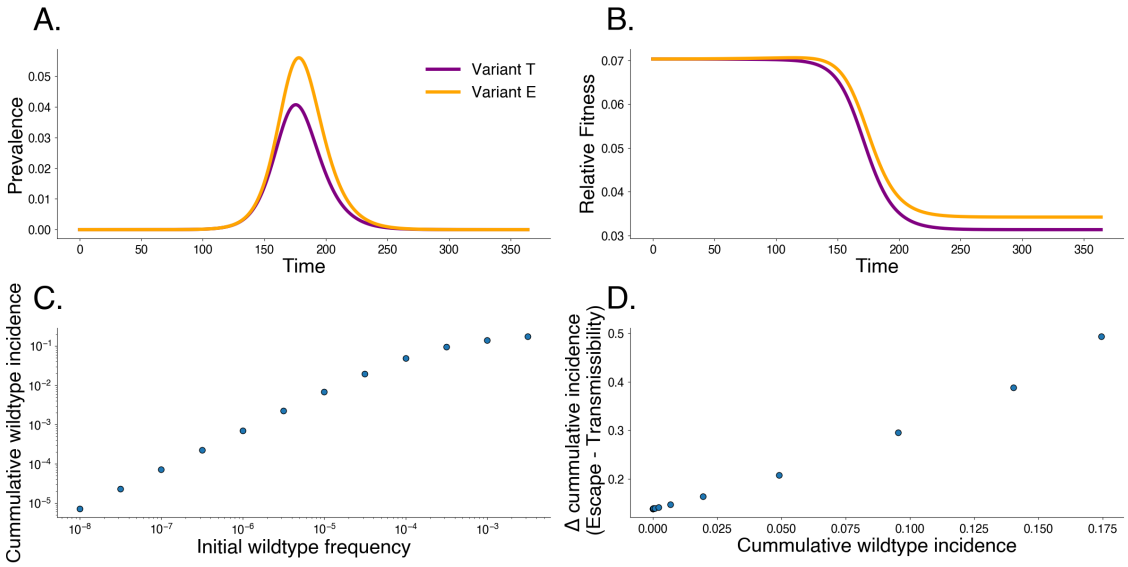

Fig. S3. Differences in fitness mechanisms impact frequency and prevalence in the short-term. Comparing simulations from two independent two-variant systems with either an escape variant E (orange) or a transmissibility variant T (purple). We fix the initial relative fitness for the two variants using Equation 4 and simulate dynamics for 365 days. A. The prevalence for the variants. B. The relative fitness from the variants. C. The cumulative wildtype incidence as a function of the initial wildtype frequency. D. The difference between the cumulative incidence between the escape variant and the transmissibility variant as a function of wildtype incidence.

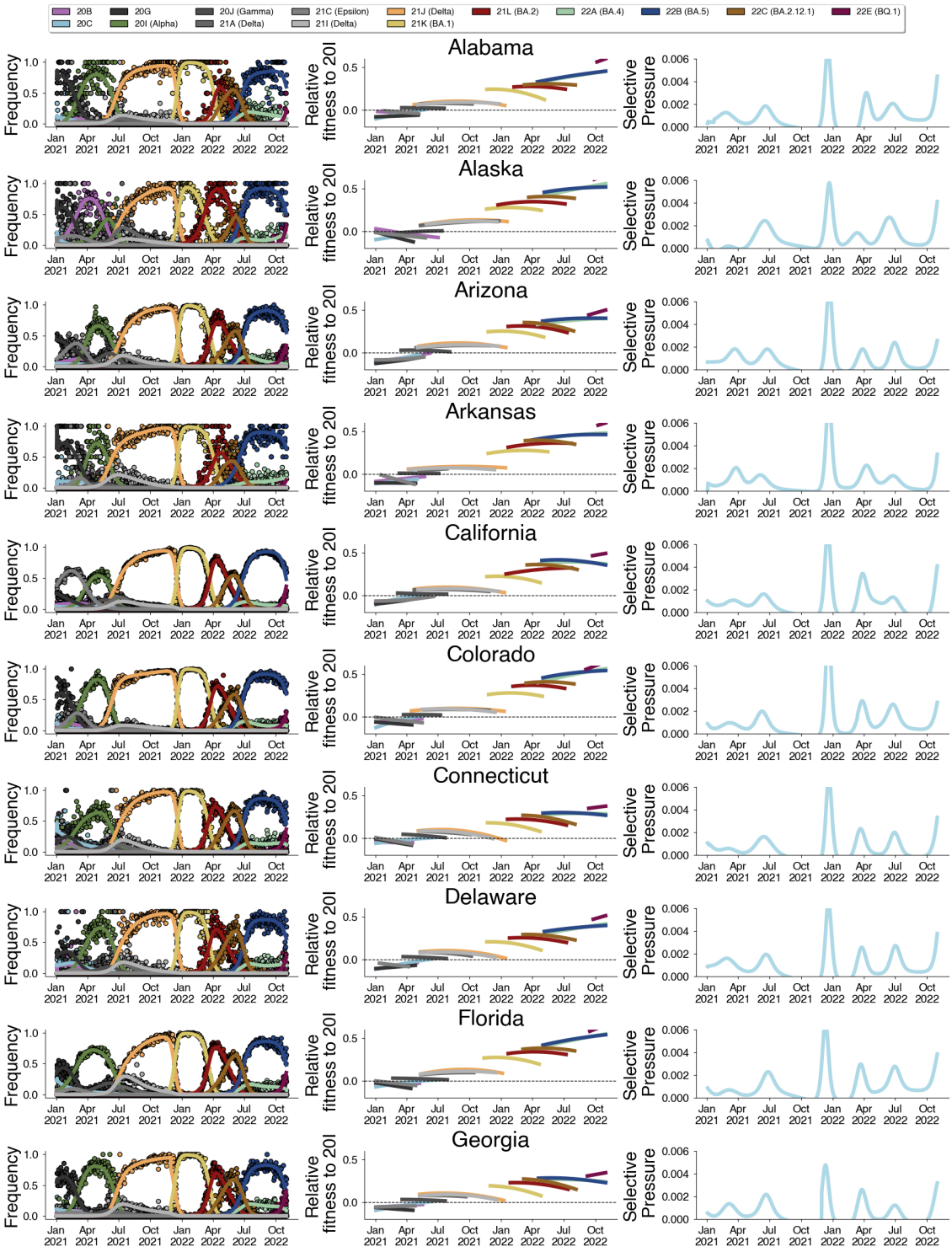

Fig. S4. Estimated variant frequencies, relative fitnesses, and selective pressure. Alabama through Georgia.

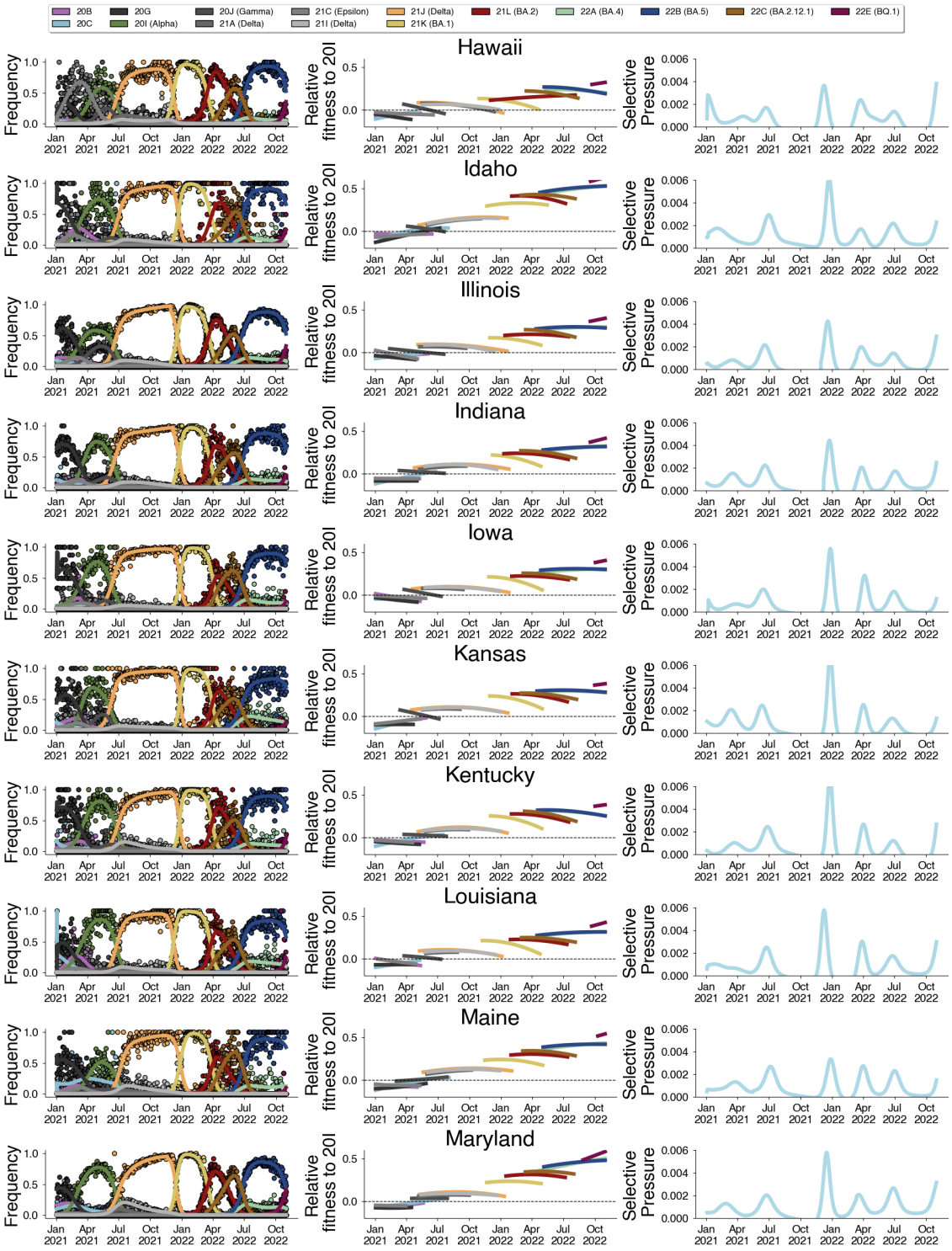

Fig. S5. Estimated variant frequencies, relative fitnesses, and selective pressure. Hawaii to Maryland.

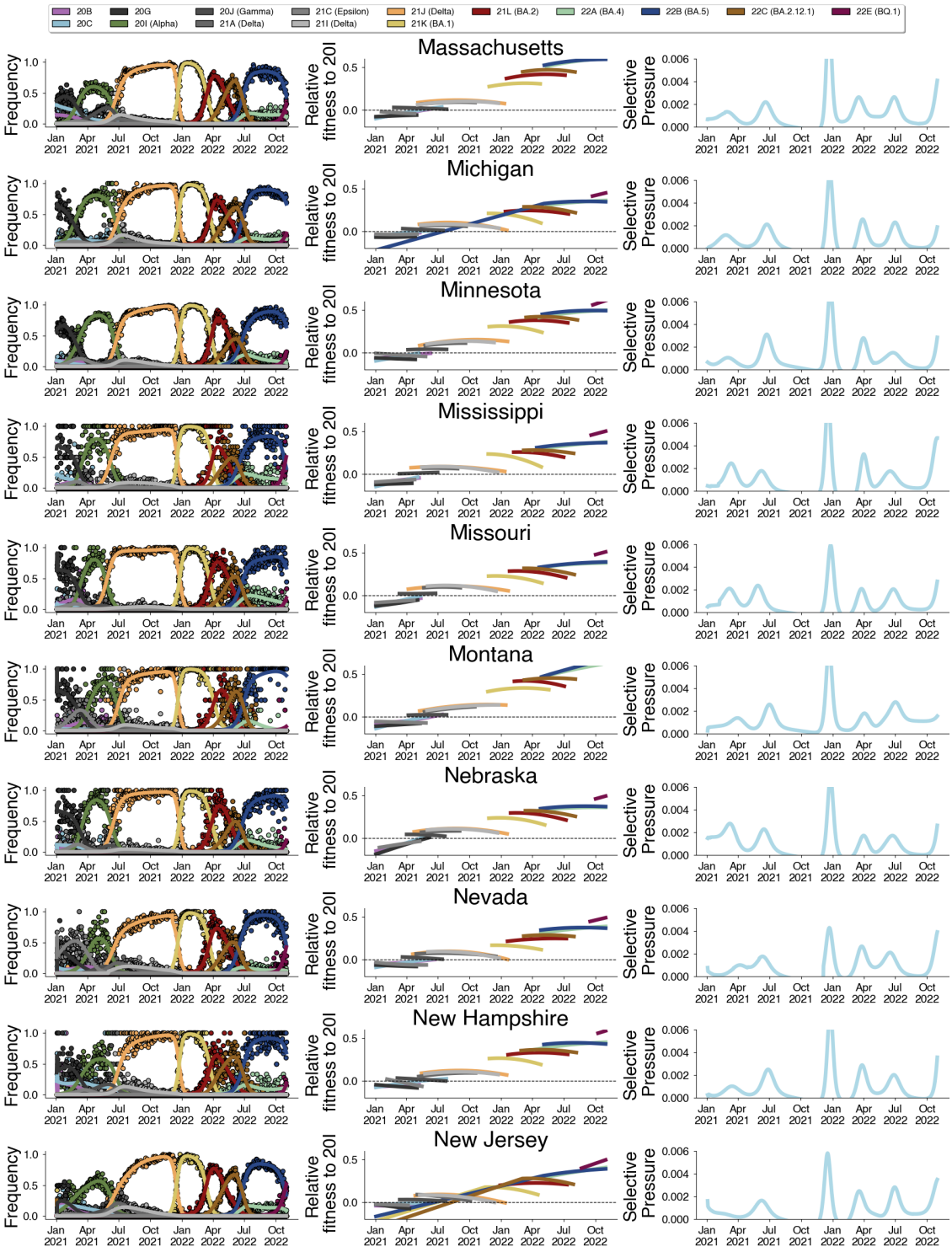

Fig. S6. Estimated variant frequencies, relative fitnesses, and selective pressure. Massachusetts to New Jersey.

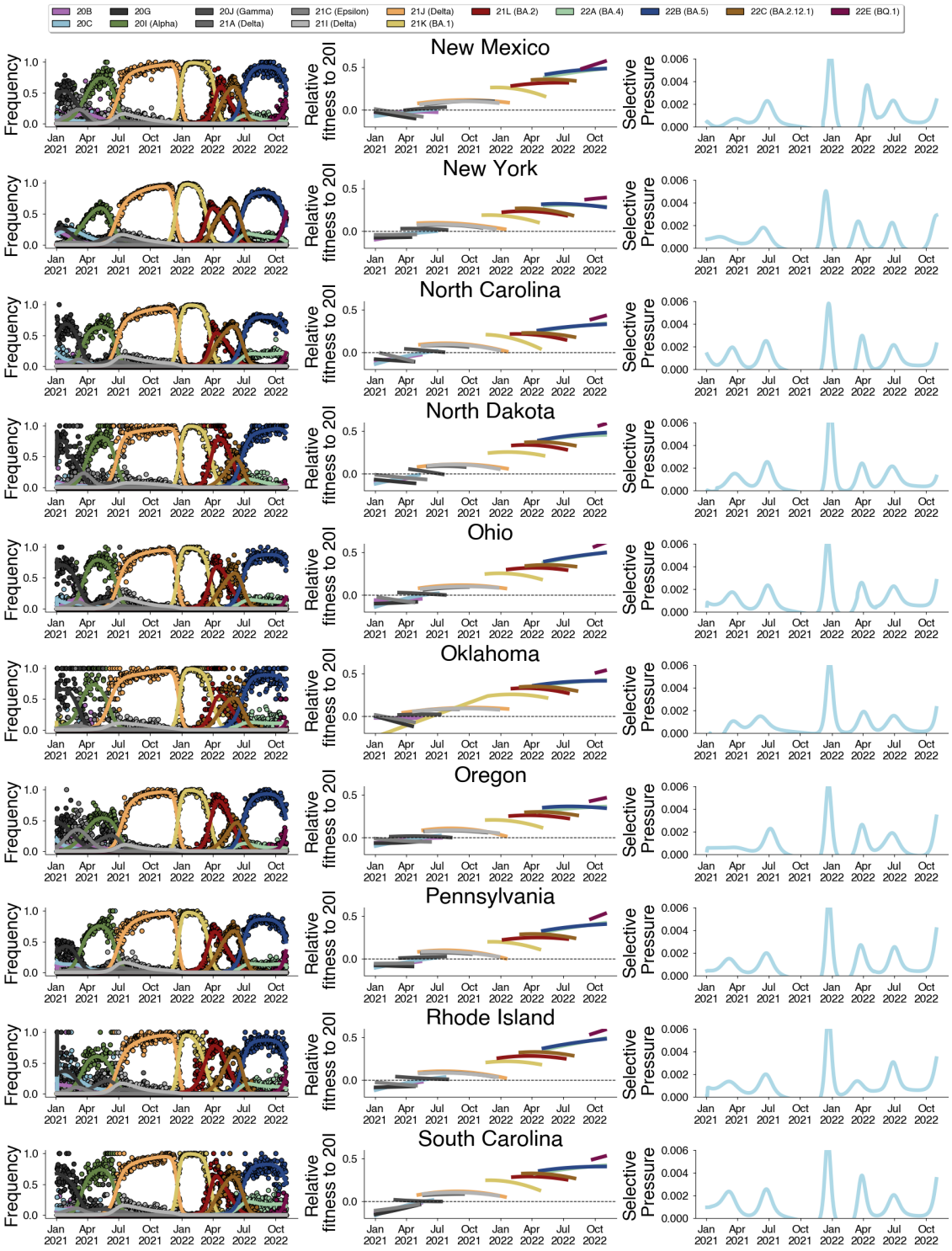

Fig. S7. Estimated variant frequencies, relative fitnesses, and selective pressure. New Mexico to South Carolina.

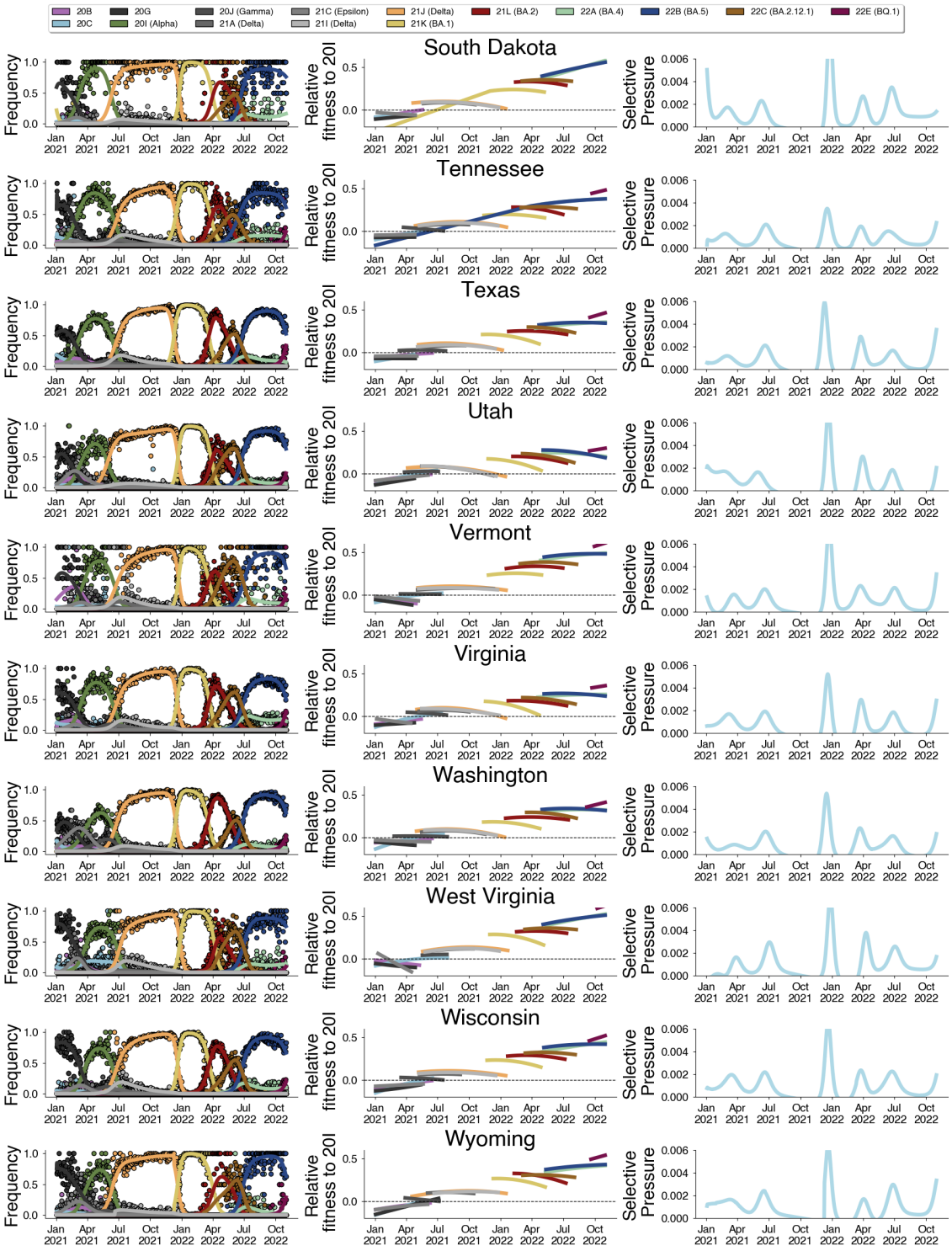

Fig. S8. Estimated variant frequencies, relative fitnesses, and selective pressure. South Dakota to Wyoming.

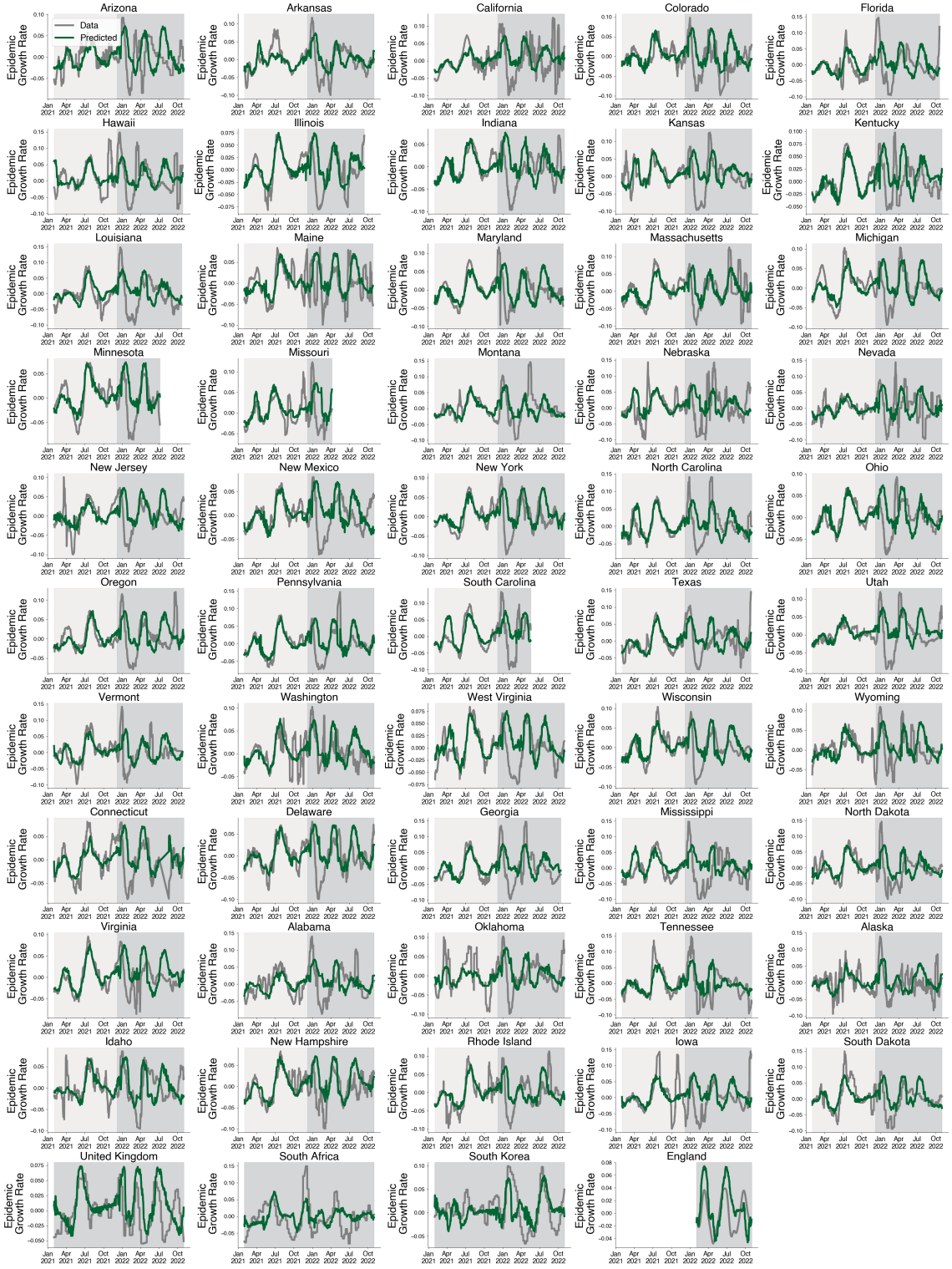

Fig. S9. Predictions for empirical growth rate using selective pressure for all locations.

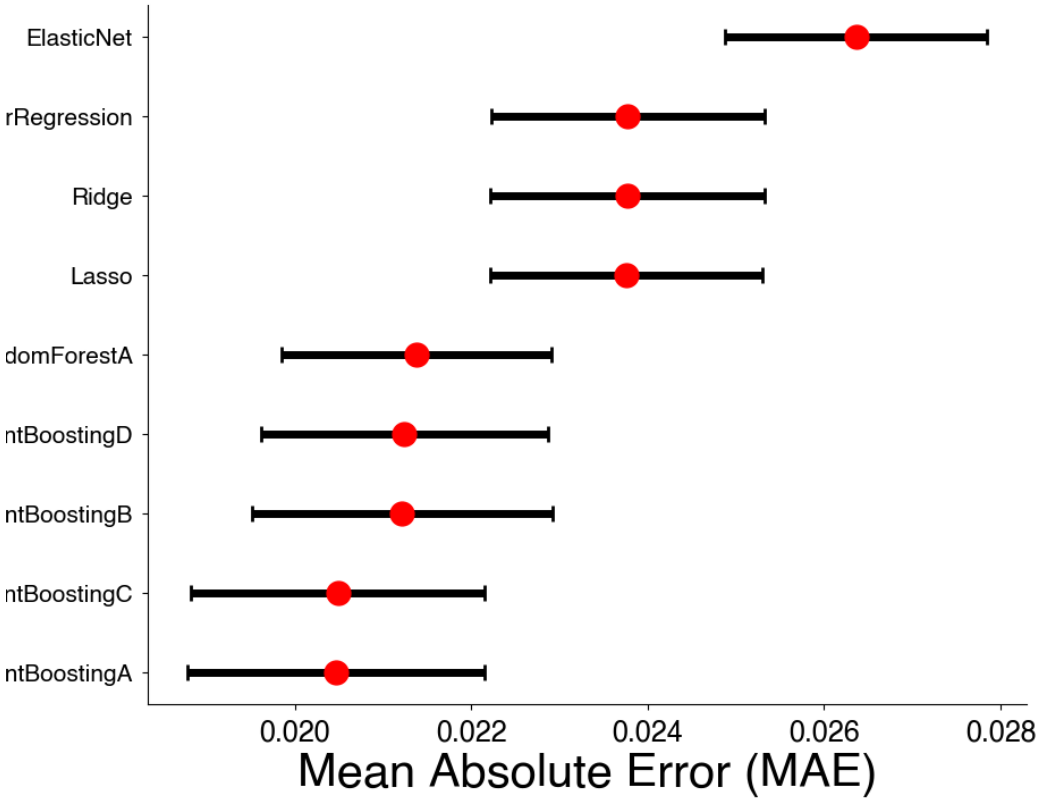

Fig. S10. Cross-validation error by model We compare the errors between models fit on 10 time series cross-validation splits.

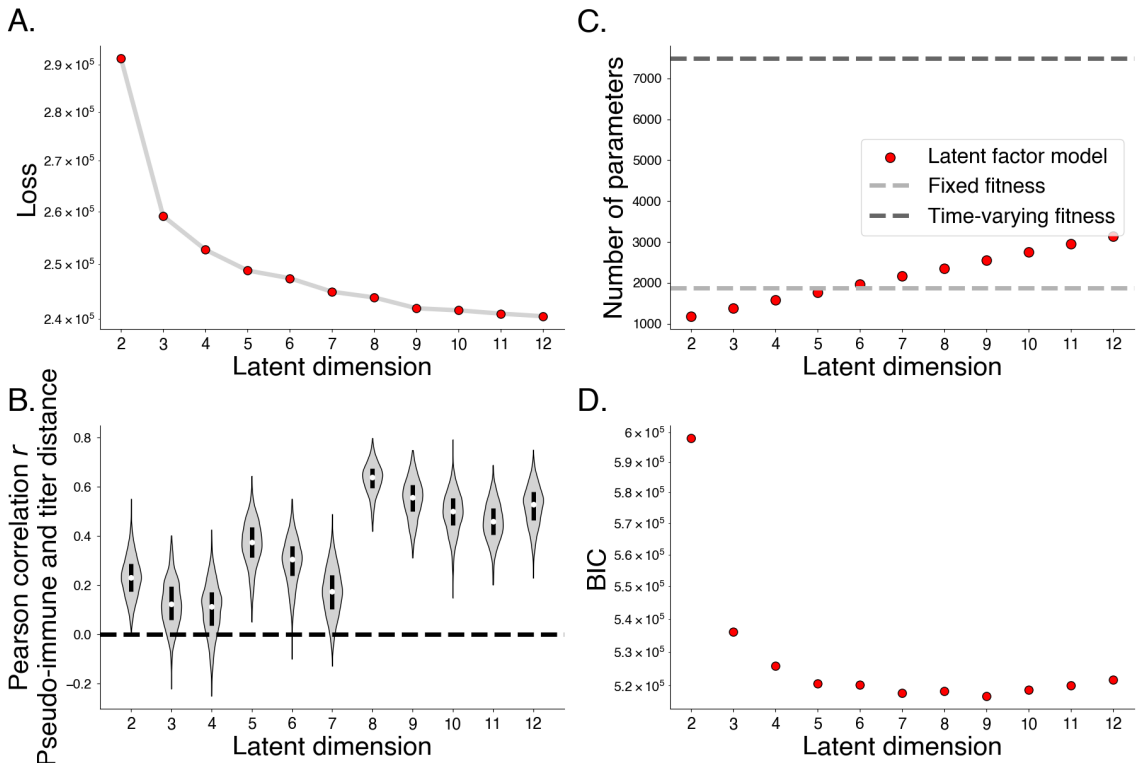

Fig. S11. Comparing latent factor model by number of latent immune dimensions. A. Maximum a posteriori loss by number of latent immune dimensions. B. Spearman correlation between pseudo-immune and titer distance by number of latent immune dimensions. C. Number of parameters by number of latent immune dimensions. D. Bayesian Information Criterion (BIC) by number of latent immune dimensions.

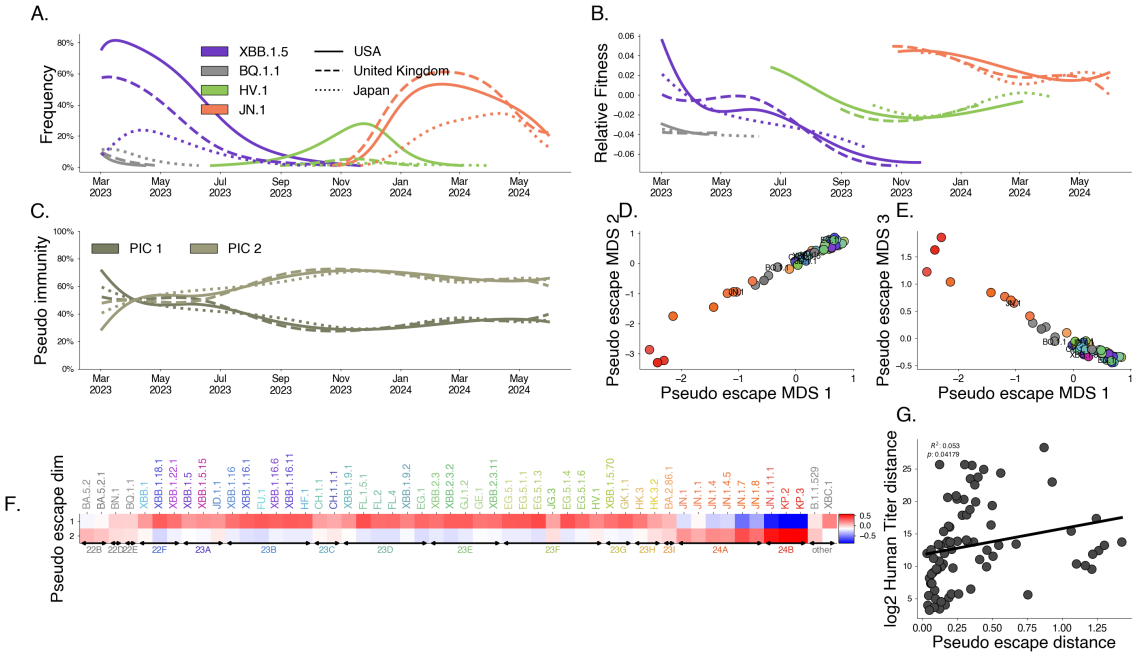

Fig. S12. Latent factor model with $D = 2$ pseudo immune dimensions.

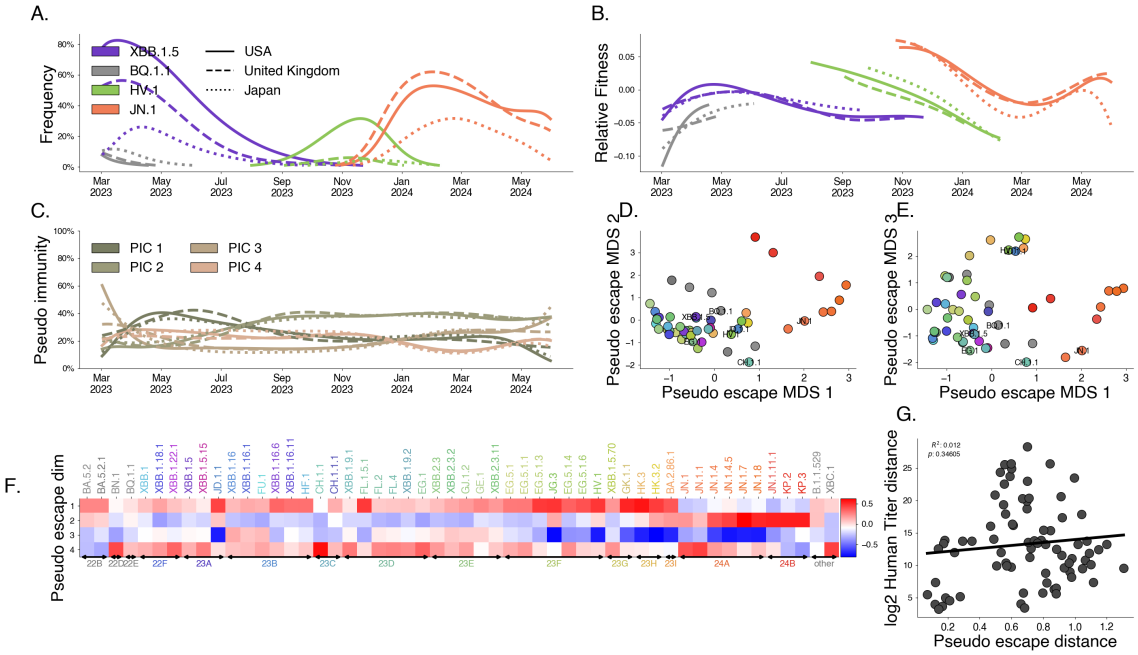

Fig. S13. Latent factor model with $D = 4$ pseudo immune dimensions.

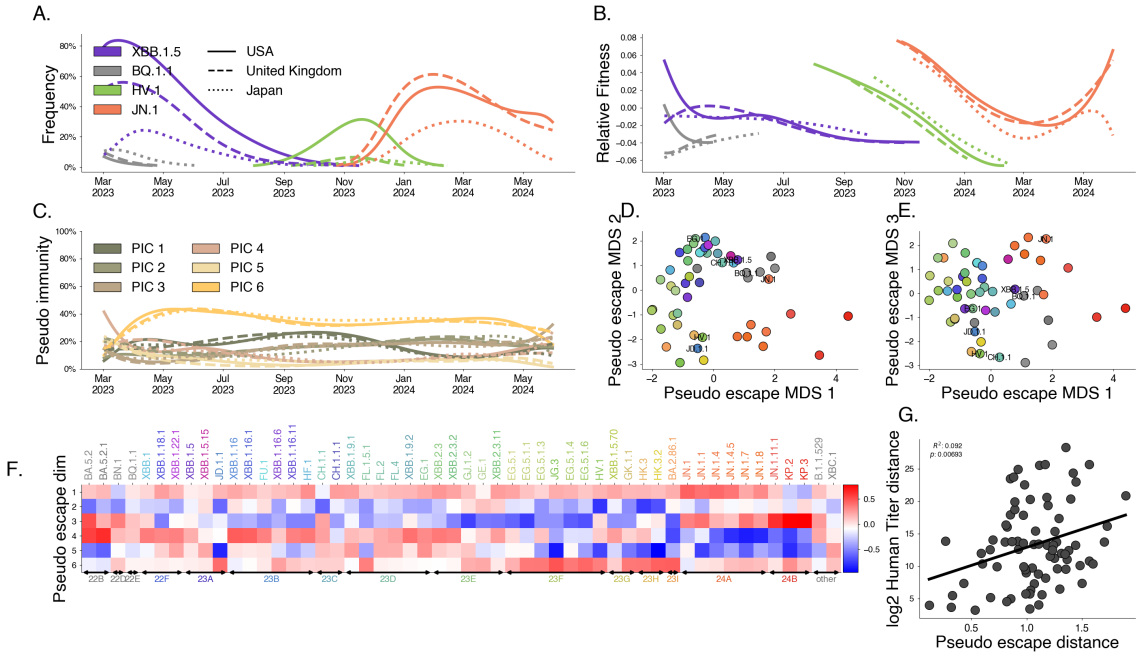

Fig. S14. Latent factor model with $D = 6$ pseudo immune dimensions.

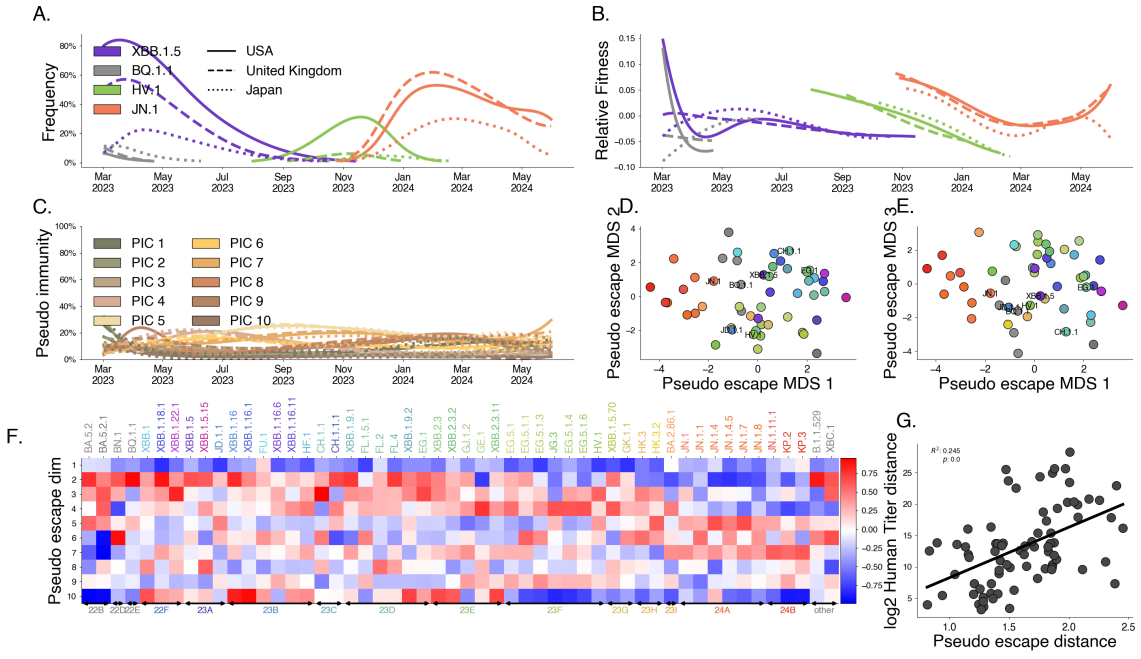

Fig. S15. Latent factor model with $D = 10$ pseudo immune dimensions.

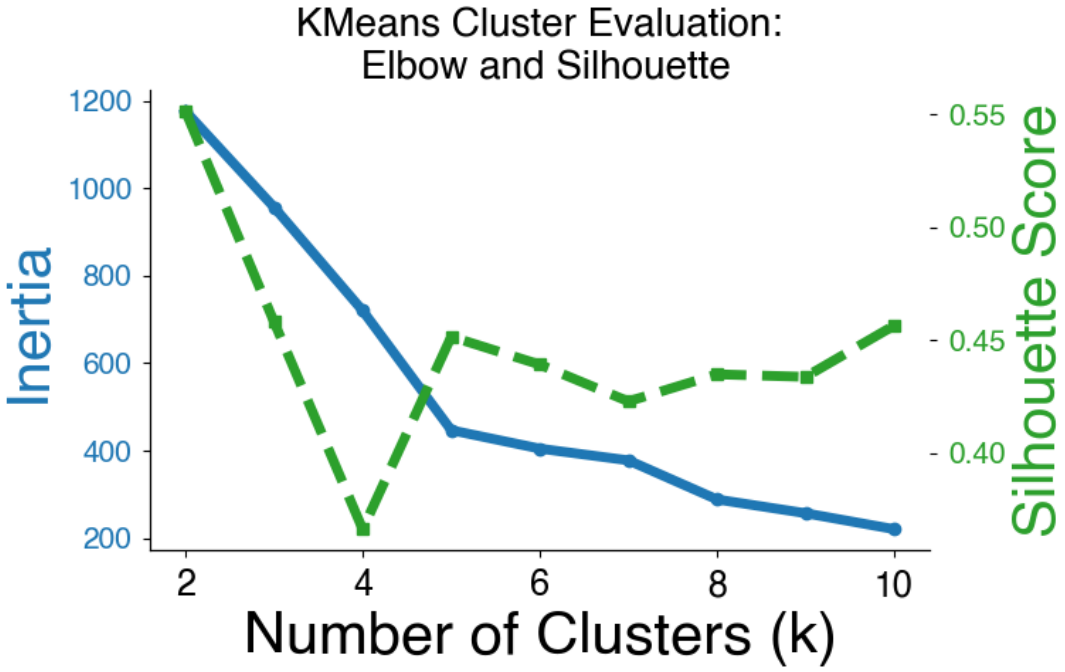

Fig. S16. Choosing the number of immune clusters ($k = 5$) using inertia and silhouette score. This analysis uses only titer data from Jian et al. [7].

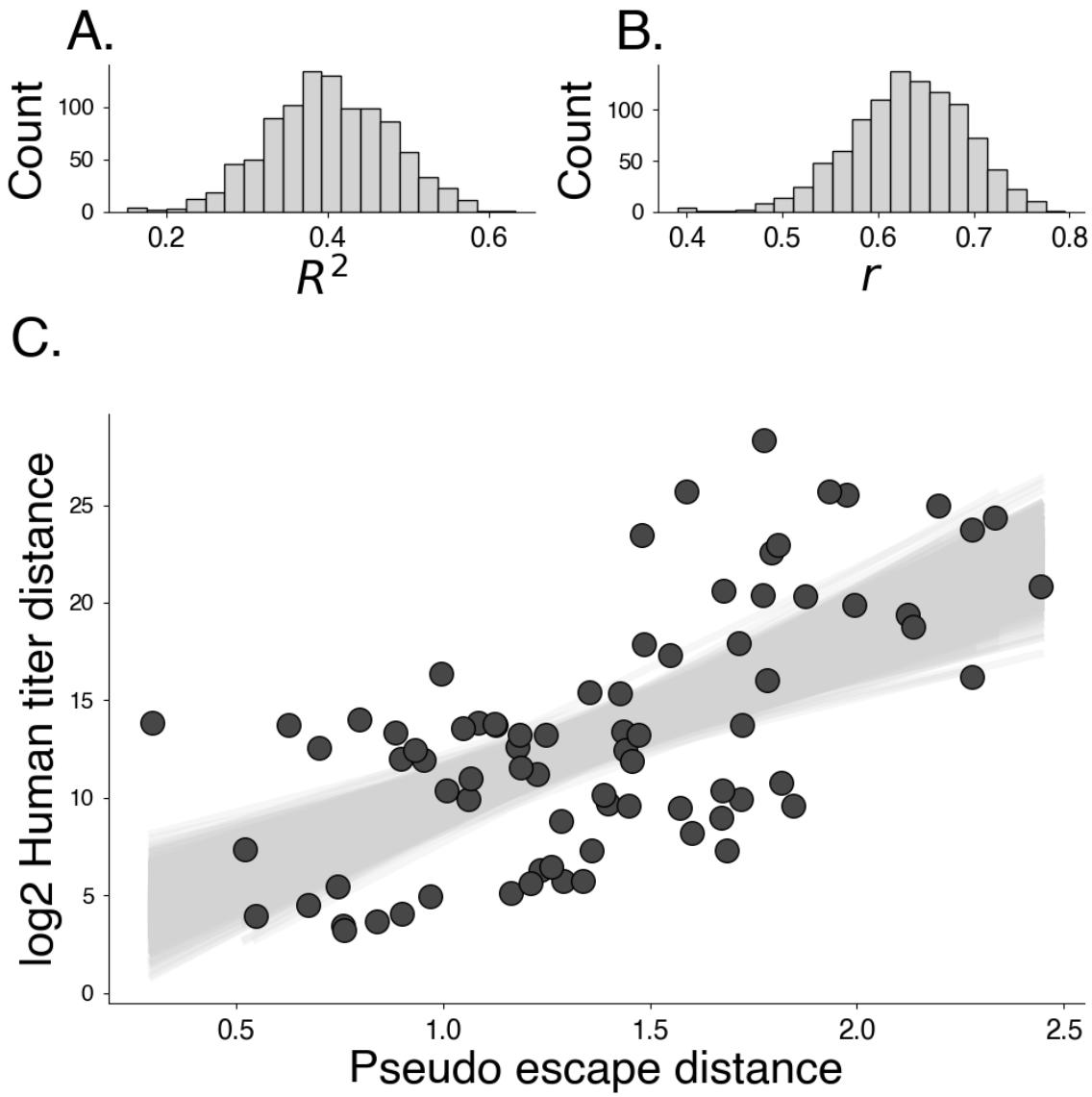

Fig. S17. Bootstrapping pseudo-immune distance and human titer distance analysis ($N_{\mathbf{replicate}} = 1,000$).

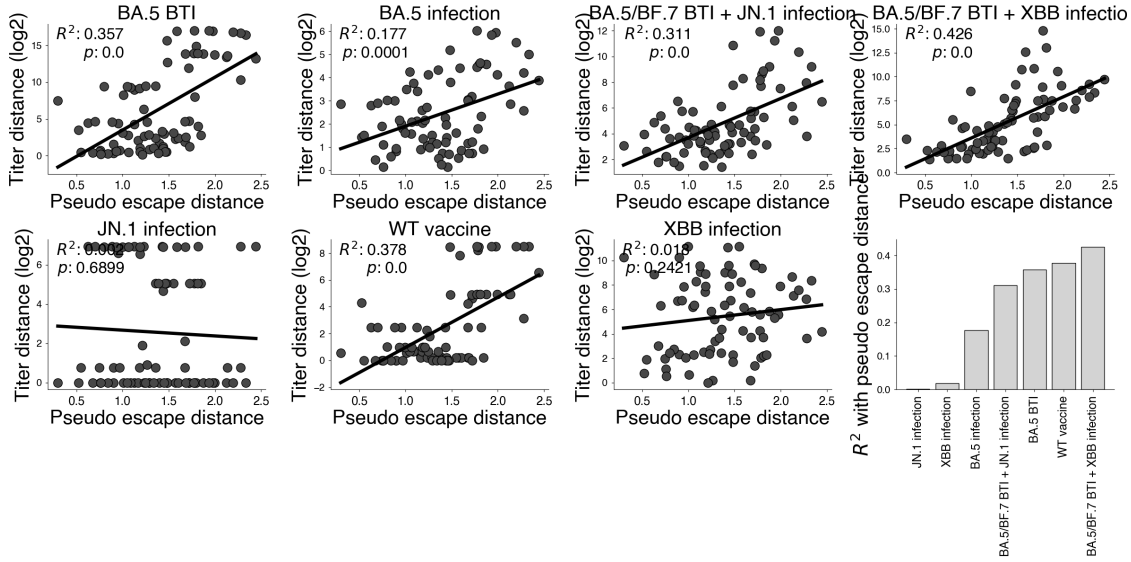

Fig. S18. Comparing pseudo-escape distance and titer distance between exposure groups.

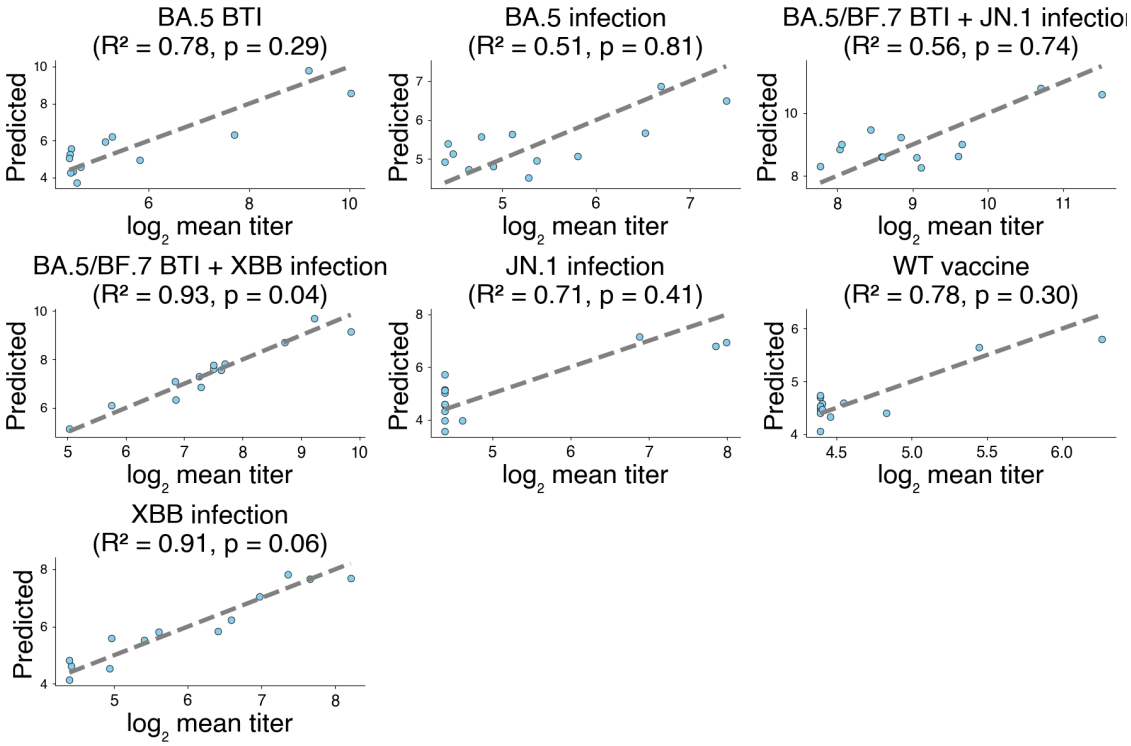

Fig. S19. Titers predicted with pseudo-escape values and observed titers within infection cohorts.

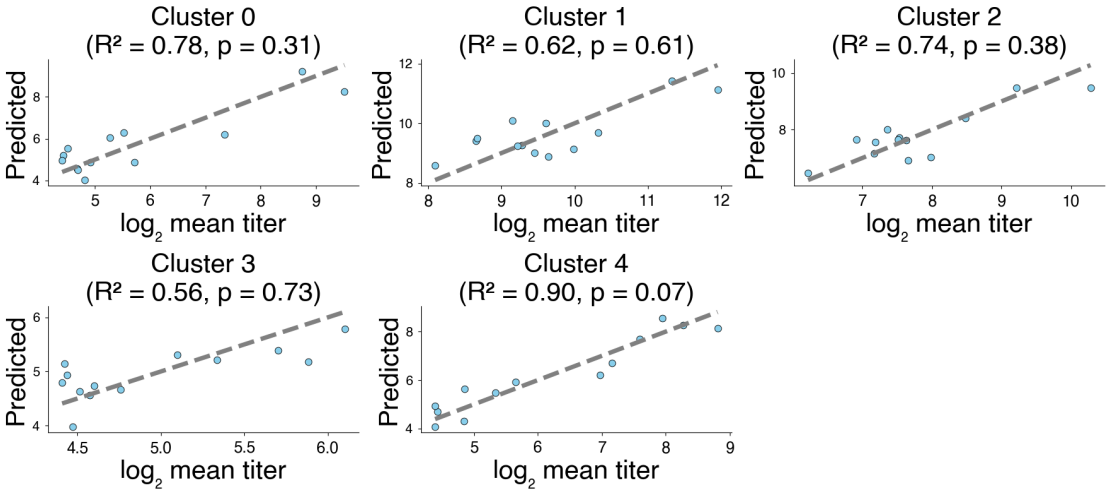

Fig. S20. Titers predicted with pseudo-escape values and observed titers within immune clusters.