

A dimensionless number for understanding the evolutionary dynamics of antigenically variable RNA viruses

Katia Koelle^{1,2,*}, Oliver Ratmann^{1,3}, David A. Rasmussen¹,
Virginia Pasour¹ and Jonathan Mattingly⁴

¹*Department of Biology, Duke University, PO Box 90338, Durham, NC 27708, USA*

²*Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA*

³*Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, UK*

⁴*Department of Mathematics, Duke University, PO Box 90230, Durham, NC 27708, USA*

Antigenically variable RNA viruses are significant contributors to the burden of infectious disease worldwide. One reason for their ubiquity is their ability to escape herd immunity through rapid antigenic evolution and thereby to reinfect previously infected hosts. However, the ways in which these viruses evolve antigenically are highly diverse. Some have only limited diversity in the long-run, with every emergence of a new antigenic variant coupled with a replacement of the older variant. Other viruses rapidly accumulate antigenic diversity over time. Others still exhibit dynamics that can be considered evolutionary intermediates between these two extremes. Here, we present a theoretical framework that aims to understand these differences in evolutionary patterns by considering a virus's epidemiological dynamics in a given host population. Our framework, based on a dimensionless number, probabilistically anticipates patterns of viral antigenic diversification and thereby quantifies a virus's evolutionary potential. It is therefore similar in spirit to the basic reproduction number, the well-known dimensionless number which quantifies a pathogen's reproductive potential. We further outline how our theoretical framework can be applied to empirical viral systems, using influenza A/H3N2 as a case study. We end with predictions of our framework and work that remains to be done to further integrate viral evolutionary dynamics with disease ecology.

Keywords: phylodynamics; infectious disease dynamics; viral evolution; immune escape; antigenic evolution

1. INTRODUCTION

Dramatic increases in the amount of viral sequence data have enabled researchers to describe the evolutionary dynamics of specific viral pathogens in ever greater detail over the last several decades. What we are still largely lacking, however, is a synthetic understanding of the diversity of viral evolutionary patterns that are empirically observed [1]. Here, we seek to address this gap by developing a theoretical framework that relates ecological factors to the dynamics of viral antigenic evolution. We specifically focus on antigenically variable RNA viruses, whose surface proteins can exhibit starkly different evolutionary dynamics, as is evidenced by the topological differences in their phylogenies (e.g. figure 1*a–c*). An understanding of the causes of these evolutionary differences is necessary to anticipate how these viruses, and the ones that will emerge in the future, will evolve. Moreover, this understanding could lead to novel control strategies that are not simply reactive, but instead aim to shape the evolutionary dynamics of viral pathogens to allow for more effective control.

At the heart of our theoretical framework is a dimensionless number that links herd immunity-driven epidemiological dynamics to an expected rate of antigenic diversification. We base our framework on a dimensionless number because these numbers (e.g. the Reynolds number [2] and the basic reproduction number R_0 [3]) have been shown to be useful in understanding and predicting a system's dynamics, regardless of the specific scale at which the system operates. The dimensionless number κ that we propose here for viral evolutionary dynamics is similar to both of these numbers in that its value depends on the properties of two players (the fluid and the pipe in the case of the Reynolds number, and the host and the virus in the case of R_0 and κ). It is further similar to R_0 in that its value depends on epidemiological parameters and, ideally, in that it could be used to inform control strategies. The critical difference between R_0 and κ is that R_0 quantifies a pathogen's reproductive potential, while κ quantifies a virus's antigenic diversification potential.

Our approach is novel for two reasons. First, although classifications of viral phylogenies based on ecological factors such as the degree of immune selection have previously been introduced [4], these have been qualitative in nature, with notable exceptions [5]. Second, previous analyses considering either within-host [6] or population-level [4,5] evolutionary dynamics have classified viral

* Author for correspondence (katia.koelle@duke.edu).

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rspb.2011.0435> or via <http://rspb.royalsocietypublishing.org>.

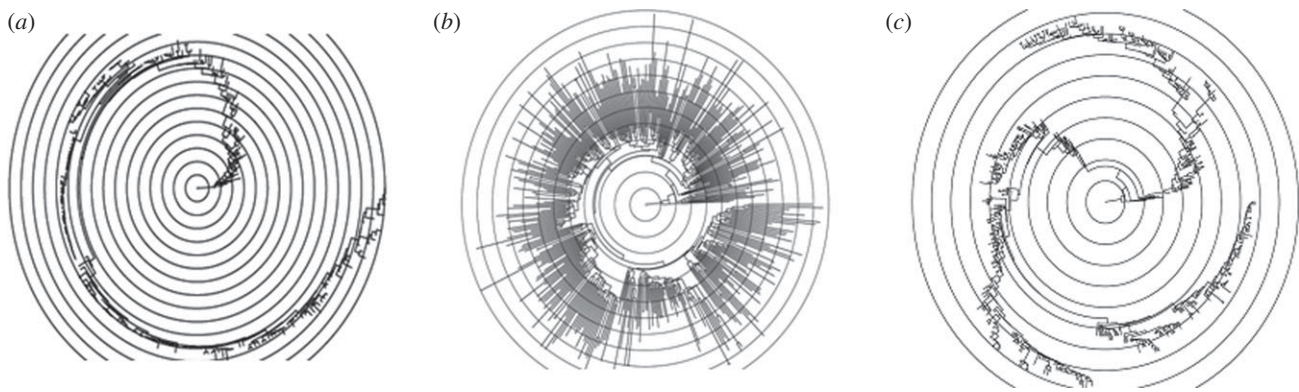


Figure 1. Phylogenetic trees of several antigenically variable RNA viruses. (a) Influenza A/H3N2's HA phylogeny in humans, inferred from sequences isolated between 1968 and 2003. (b) HIV's gp120 phylogeny (C2-V5 region), inferred from sequences isolated in the US between 1981 and 2007. (c) Influenza B's HA phylogeny, inferred from sequences isolated between 1973 and 2008. In all subplots, concentric circles measure time.

phylogenies into one of the two general types: 'cactus-like' (or 'ladder-like') phylogenies, with low levels of genetic diversity circulating at any point in time and rapid lineage turnover, and 'acacia-like' phylogenies, with growth in genetic diversity over time. Examples of the former include influenza A/H3N2's haemagglutinin (HA) (figure 1a) [7,8] and norovirus's capsid protein [9,10], while examples of the latter include HIV's gp120 protein (figure 1b) and the HA of influenza A/H5N1 circulating among avian hosts. In contrast to these previous analyses, our κ framework considers a spectrum between these diversification patterns. This allows us to quantitatively consider phylogenetic patterns that do not readily fall into one of these two extremes, such as the evolutionary dynamics of influenza B's HA (figure 1c).

We organize this paper by first developing our theoretical framework in §2 and then by illustrating the ways in which this framework can be applied to empirical host-virus systems in §3. In §4, we conclude with a discussion of the limitations of our current framework and ways forward towards a better understanding of viral evolutionary dynamics.

2. THE κ FRAMEWORK

In §2a, we first describe the structure of the epidemiological model that underlies our theoretical framework and develop our dimensionless number κ from this model. In §2b, we then show how κ can be used to probabilistically describe long-term patterns of antigenic evolution.

(a) *The epidemiological model*

We assume that the dynamics of antigenically variable RNA viruses are governed by a status-based multi-strain model [11] of the form:

$$\frac{dS_i}{dt} = \mu N - \beta \frac{S_i}{N} \sum_j \sigma_{ij} I_j - \mu S_i \quad (2.1)$$

and

$$\frac{dI_i}{dt} = \beta \frac{S_i}{N} I_i - (\mu + \nu) I_i - h_W(t - t_i) I_i \quad (2.2)$$

where subscript i denotes a single antigenic variant, μ is the birth rate and death rate, β is the transmission rate, ν is the recovery rate and σ_{ij} is the degree of cross-immunity

between antigenic variants i and j . The term $h_W(t - t_i)$ represents a *per capita antigenic emergence rate*: the rate at which an individual infected with antigenic variant i changes into being infected with a new and yet-unseen antigenic variant [12]. While most evolutionary multi-strain models assume that this rate is constant and equivalent to the mutation rate (or a factor thereof), we here more generally allow h_W to be a function of an antigenic variant's age a , given by the difference between the current time t and the time of the variant's emergence t_i . Specifically, we parametrize h_W using a Weibull hazard rate with shape parameter k_W and scale parameter λ_W :

$$h_W(a) = \frac{k_W}{\lambda_W} \left(\frac{a}{\lambda_W} \right)^{k_W - 1} \quad (2.3)$$

With $k_W = 1$, h_W is a constant. With $k_W > 1$, h_W increases with a variant's age. This can happen if phenotypically neutral mutations accruing over time increase the probability that additional mutations result in significant antigenic change (e.g. [13]) or if a certain number of 'permissive' mutations are necessary for phenotypic change to occur (e.g. [14]). The scale parameter λ_W can be thought of as quantifying the extent of phenotypic robustness to genetic change, with higher values of λ_W being indicative of higher robustness. As written, equations (2.1)–(2.3) explicitly consider immune selection on viral antigenic phenotypes, rather than on individual genotypes. We use this approach because it yields a computationally simpler model, while remaining consistent with a range of previously published multi-strain model formulations and a range of empirical observations (e.g. [15]). Along with equations (2.1)–(2.3), we assume that a single antigenic variant comprises genotypes that evolve neutrally.

To derive our dimensionless number κ , we consider a simplified, stochastic version of this epidemiological model. We start with a single endemic antigenic variant X circulating in a host population, giving rise to new variants ('offspring variants') over time. We assume that the degree of cross-immunity between variant X and an offspring variant is given by σ , that there is no cross-immunity between offspring variants, and that offspring variants do not give rise to new antigenic variants. The epidemiological dynamics of variant X are governed by classic single-strain susceptible-infected-recovered (SIR) dynamics until an offspring variant emerges that is not

stochastically lost from the population. We call this first successful variant Y . As a result of variant Y 's frequency-dependent selective advantage, and the competition between variants X and Y for susceptible hosts, the invasion of variant Y results in a decrease in the number of individuals infected with variant X , until, under a wide range of model parameter values, variant X is lost from the host population. Once this happens, the only variants remaining in the population are variant Y and any secondary successful offspring variants that may have been generated by variant X prior to its exclusion. The dimensionless number κ we propose is given by the expected number of these secondary variants ('excess variants'). Setting the time of emergence of variant X to $t_X = 0$, κ is given by:

$$\kappa = \int_0^\infty g_Y(t_Y) \left[\int_{t_Y}^\infty u(t) h_W(t) I_X(t) dt \right] dt_Y, \tag{2.4}$$

where $I_X(t)$ is the number of individuals infected with variant X at time t , $u(t)$ is the probability that a variant given rise to at time t is not stochastically lost from the population and $g_Y(t_Y)$ is the probability density function for the time to variant Y 's emergence.

A conceptually similar quantity was outlined by Sasaki & Haraguchi [6] when considering viral dynamics within chronically infected hosts. Two critical differences, however, exist between the formulation of their dimensionless number and the one we propose here. First, their within-host model assumes that a variant will be driven to increasingly low numbers and ultimately to extinction from a persistent, variant-specific B-cell response. In contrast, an antigenic variant in our population-level model reaches a quasi-stationary endemic steady state when alone in the host population, as long as $R_0 > 1$. Variant X will therefore always generate at least one new successful antigenic variant, variant Y , and κ quantifies the expected number of *excess* variants that variant X generates over its lifetime. Second, Sasaki & Haraguchi's within-host model defines a variant genetically, such that the production of new variants within a host occurs at a constant mutation rate. As described above, our model defines a variant antigenically.

To evaluate κ for a given set of epidemiological parameters, we can derive analytical expressions to approximate $I_X(t)$, $u(t)$ and $g_Y(t_Y)$. To derive $I_X(t)$, we assume that the selective advantage s of variant Y is constant and given by its initial selective advantage at the time of its emergence t_Y . This selective advantage is appropriately defined using variants Y 's and X 's net reproductive rates. These can be evaluated at time t_Y by analytically solving equations (2.1)–(2.2) for equilibrium when only one variant (variant X) is present in the population:

$$s = \frac{R_0(S_Y/N)}{R_0(S_X/N)} - 1 = \frac{R_0}{1 + \sigma(R_0 - 1)} - 1, \tag{2.5}$$

where S_Y and S_X are the number of hosts susceptible to variants Y and X , respectively, and $R_0 = \beta/(\mu + \nu)$. The relative reproductive rate of variant Y in units of calendar time is therefore given by $s(\mu + \nu)$, where $1/(\mu + \nu)$ is an individual's expected duration of infection. Further assuming that the total number of infected individuals ($I_X + I_Y$) stays constant at the equilibrium number of infected individuals for a single-strain SIR system ($I = \mu N(R_0 - 1)/\beta$), variant Y 's population dynamics following its emergence

at time t_Y can be written deterministically according to the logistic equation:

$$\frac{dI_Y}{dt} = s(\mu + \nu) \left(1 - \frac{I_Y}{I} \right) I_Y. \tag{2.6}$$

The analytical solution to this equation is:

$$I_Y(t) = \frac{C e^{s(\mu + \nu)(t - t_Y)}}{[1 + (C/I) e^{s(\mu + \nu)(t - t_Y)}]}, \tag{2.7}$$

where, with $I_Y(t_Y) = 1$, $C = I/(I - 1)$. Our final analytical approximation to $I_X(t)$ is therefore given by the difference between I and $I_Y(t)$:

$$I_X(t) = I - I_Y(t) = \frac{I(I - 1)}{I - 1 + e^{s(\mu + \nu)(t - t_Y)}}. \tag{2.8}$$

In figure 2a, the dynamics of variants X and Y from a representative stochastic simulation are shown alongside the analytical expression for $I_X(t)$ given by equation (2.8).

To derive $u(t)$, we rely on the well-known result that the probability of stochastic loss of an invading virus is given by the inverse of its net reproductive rate [16]. Assuming that the number of individuals susceptible to infection at time t by an offspring variant that has not yet emerged can be approximated by the number of individuals susceptible to this variant prior to variant Y 's emergence; u becomes

$$u = (1 - \sigma) \left(1 - \frac{1}{R_0} \right), \tag{2.9}$$

where we have dropped the notation of time dependence. In figure 2b, our constant approximation of u is shown alongside the dynamics of $u(t)$ computed from the epidemiological simulation shown in figure 2a.

Finally, we derive an analytical approximation to $g_Y(t)$. This density function can be written as:

$$g_Y(t) = I f(t) (1 - F(t))^{I-1}. \tag{2.10}$$

where $f(t)$ is the probability density function for the time it takes a single individual infected with variant X to generate an antigenic variant that is not stochastically lost, and $F(t)$ is its cumulative distribution function. $f(t)$ is given by:

$$f(t) = u h_W(t) e^{-\int_{s=0}^t u h_W(s) ds}, \tag{2.11}$$

where u is the probability given by equation (2.9). Substituting equation (2.3) into equation (2.11) and writing $F(t)$ in terms of $f(t)$, equation (2.10) simplifies to:

$$g_Y(t) = I \frac{u k_W}{\lambda_W} \left(\frac{t}{\lambda_W} \right)^{k_W - 1} e^{-u I (t/\lambda_W)^{k_W}}. \tag{2.12}$$

Taking expectations, we have the mean time to the generation of variant Y :

$$E(t_Y) = \lambda_W (u I)^{-1/k_W} \Gamma(1 + 1/k_W). \tag{2.13}$$

In figure 2c,d, we compare our approximation for $g_Y(t)$ given by equation (2.12) to $g_Y(t)$ distributions from stochastic simulations of the simplified epidemiological model.

By substituting expressions (2.8), (2.9) and (2.12) into equation (2.4), the value of κ can now be computed through numerical integration once all seven epidemiological model parameters (N , μ , ν , R_0 , σ , k_W and λ_W)

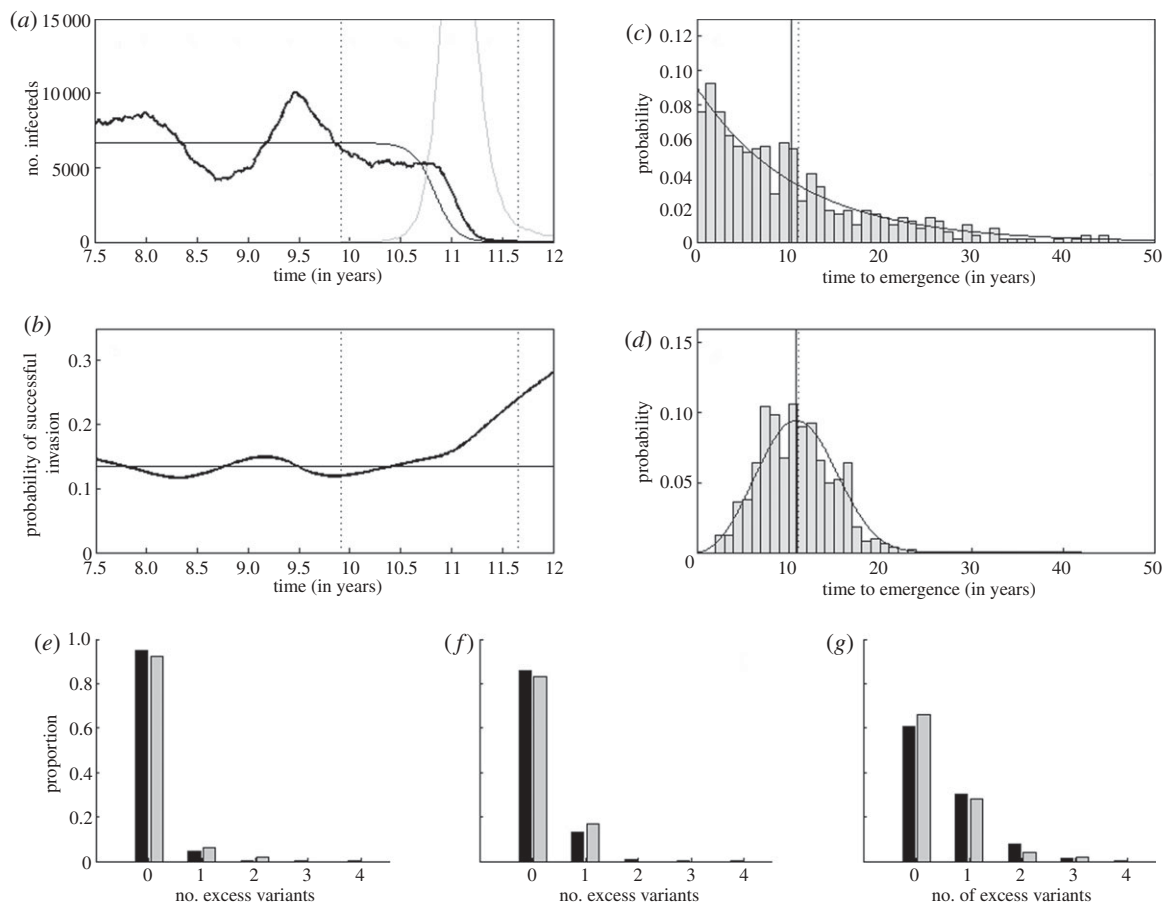


Figure 2. A comparison between the analytical expressions used to calculate κ and stochastic simulations of the simplified epidemiological model. (a) A representative simulation showing the dynamics of variant X (bold black) and variant Y (grey), alongside the analytical expression for $I_X(t)$ (black), given by equation (2.8). (b) For the simulation shown in (a), the probability that a new variant, emerging at time t , successfully invades ($u(t)$, bold black) alongside its assumed analytical value given by equation (2.9) (black). $u(t)$ is calculated from the simulation as $1 - 1/(R_0(S(t)/N))$, where $S(t)$ is the number of individuals susceptible, at time t , to an offspring variant that has not yet emerged. Simulation parameters are $N = 18$ million, $R_0 = 10$, $\nu = 1/6 \text{ d}^{-1}$, $\mu = 1/40 \text{ yr}^{-1}$ and $\sigma = 0.85$. In both (a) and (b), the first vertical line shows the time of variant Y 's emergence and the second vertical line shows the time of variant X 's exclusion. (c) The distribution of times to the emergence of variant Y from 500 simulations of the simplified epidemiological model (bars), alongside the probability density function, $g_Y(t)$. Simulation parameters are as in (a,b), with additional parameters $k_W = 1$ and $\lambda_W = 10\,000$. Vertical solid black line shows the mean of the 500 simulated times to emergence. Vertical dotted line shows the analytical expected time to emergence (equation (2.13)). (d) Same as in (c), only with $k_W = 3$ and $\lambda_W = 120$. (e–g) Distributions for the number of excess variants, from 100 simulations of the simplified epidemiological model (grey bars), alongside Poisson distributions with mean κ (black bars). In all three subplots, simulation parameters were $N = 100$ million, $R_0 = 10$, $\nu = 1/10 \text{ d}^{-1}$, $\mu = 1/70 \text{ yr}^{-1}$, $\sigma = 0.9$ and $\lambda_W = 2000$. (e) Number of excess variants from 100 simulations, each with $k_W = 3.34$, yielding $\kappa = 0.05$. The mean number of excess variants, computed from the simulation results, is 0.09. (f) Number of excess variants from 100 simulations, each with $k_W = 2.01$, yielding $\kappa = 0.15$. The mean number of excess variants, computed from the simulation results, is 0.17. (g) Number of excess variants from 100 simulations, each with $k_W = 1.476$, yielding $\kappa = 0.50$. The mean number of excess variants, computed from the simulation results, is 0.42. The number of excess variants for a given simulation was calculated by subtracting one from the number of variants co-circulating in the population 5 years after the extinction of variant X . We used a likelihood-ratio test to determine whether the null hypothesis (that the number of excess variants comes from a Poisson distribution with mean κ) could be rejected at the 95% confidence level in any of the three cases (e–g) above. In none of the three cases could this null hypothesis be rejected.

have been specified. As long as I and the product uI are large, an approximation to this double-integral expression yields:

$$\kappa \approx Iu \frac{\log^{k_W}(I)}{(\lambda_W s(\mu + \nu))^{k_W}}. \quad (2.14)$$

The derivation of this approximation is detailed in the electronic supplemental material. It is clear from equation (2.14) that the epidemiological model parameters affect the final value of κ directly and/or indirectly through I , s and u . The sensitivity of κ to each of these parameters is illustrated

in the electronic supplementary material, figure S1. κ increases with population size N , with the duration of infection ($1/\nu$), with the degree of cross-immunity σ and decreases with host lifespan ($1/\mu$). Interestingly, the effect of the basic reproduction number R_0 on κ depends on the values of the other parameters. Finally, κ increases with a decrease in the antigenic emergence rate's scale parameter λ_W and with a decrease in its shape parameter k_W , most dramatically as k_W approaches one.

To determine whether κ is a good approximation to the mean number of excess variants generated by variant X prior to extinction, we simulated the simplified

epidemiological model under three different κ parametrizations (figure 2e–g). These simulations show that κ , despite the numerous approximations made in its derivation, closely matches the mean number of excess variants computed from simulations. Furthermore, these simulations indicate that the number of excess variants generated by a single endemic variant can be considered a Poisson random variable with mean κ .

(b) The link between κ and long-term patterns of viral antigenic evolution

To relate patterns of viral antigenic evolution to κ , we first define some useful terminology. Let a generation be defined as the mean time between an endemic variant’s emergence time and the emergence time of this variant’s first successful antigenic offspring. In terms of the simplified model, a generation is given by $E(t_Y)$. Furthermore, let n_k be defined as the number of antigenic variants co-circulating in a host population in generation k . κ can be used to project the distribution of n_k over successive generations and thereby to describe long-term patterns of viral antigenic evolution under three assumptions: (i) that the time between the emergence of the offspring variant(s) and the extinction of the variant that gave rise to it/them is negligible when compared with the generation time, (ii) that successful antigenic variants immediately equilibrate to their endemic levels, given by I , and (iii) that the number of excess variants generated by a single endemic variant is a Poisson random variable with mean κ , which we have already seen to be a reasonable assumption (figure 2e–g).

Given these assumptions, we have a forward model for simulating antigenic evolution. Starting with only a single endemic antigenic variant ($n_0 = 1$), $n_1 = 1 + i_0$, where i_0 is a random variable drawn from a Poisson distribution with mean κ . Continuing, the next generation then has $n_2 = n_1 + i_1$, where i_1 is the sum of a collection of n_1 independent random variables all drawn from a Poisson distribution with mean κ . Continuing further, this process results in a synchronized accumulation of antigenic diversity over viral generations. Figure 3a illustrates a stochastic realization of these viral diversification dynamics for $\kappa = 0.2$. Given our model’s specification that viral genotypes belonging to a single antigenic phenotype evolve neutrally, we can sample a representative phylogeny arising from these diversification dynamics using coalescent theory (figure 3b).

The benefit of κ being dimensionless now becomes evident. Although the epidemiological parameters and generation times of two host–virus systems may differ from one another, these systems will be governed by the same diversification rates if they have the same κ value. Epidemiologically distinct host–virus systems may therefore exhibit what is known as ‘dynamic similarity’, with similar patterns of antigenic evolution arising from a quantitative agreement of their dimensionless numbers.

From the structure of our forward model, we can also compute the probability mass function for the number of co-circulating variants in a given generation using only κ :

$$g_k(n_k; \kappa) = \sum_{m=1}^{n_k} [g_{k-1}(m; \kappa) l(n_k - m; \kappa m)], \quad (2.15)$$

where $l(i; x)$ is the probability that there are exactly i variants from a Poisson distribution with mean x , $g_0(n_0 = 1; \kappa) = 1$ and $g_0(n_0; \kappa) = 0$ for all $n_0 \neq 1$.

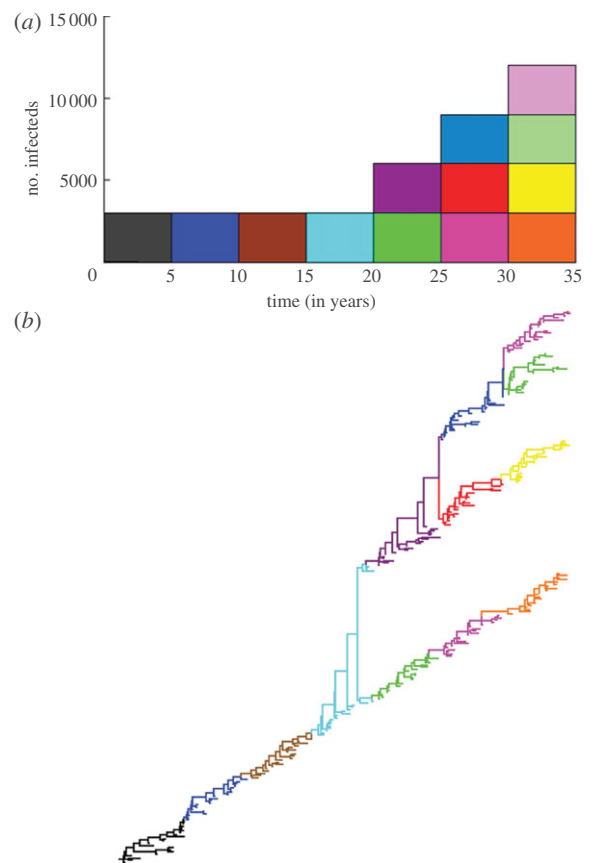


Figure 3. Long-term patterns of antigenic evolution and their dependence on κ . (a) A single stochastic realization of long-term antigenic evolution, given a κ value of 0.2, an equilibrium number of $I = 3000$ infected individuals and a generation time of 5 years. Different colours represent different antigenic variants. Over the first three generations, none of the endemic variants generate any excess variants. In the fourth generation, a single excess variant is generated. In the fifth generation, one of the two endemic variants generates no excess variants while the other generates one excess variant. In the sixth generation, two of the three endemic variants generate no excess variants while the third generates one excess variant. (b) A representative phylogeny arising from the diversification dynamics shown in (a). The phylogeny was constructed using a neutral coalescent model to generate the branching structure within clusters while the ancestral relationships among clusters and their emergence and death times were determined by the dynamics shown in (a). The topology represents one realization of the phylogenetic patterns possible under the given epidemiological dynamics. Colours of the branches correspond to the variant designations shown in (a).

It is clear from this expression and from the structure of the forward model that a large number of generations may pass in systems with small κ before a substantial amount of viral diversity accrues. However, the mean of n_k is given by $(1 + \kappa)^k$, and its standard deviation grows similarly with the number of generations k . The number of co-circulating variants will therefore always end up growing exponentially. According to this model, all viral populations will therefore eventually have many co-circulating variants, and the level of antigenic diversity will not saturate.

The relationship between κ and the number of co-circulating antigenic variants in a given generation rests on the three assumptions listed above. To determine our framework’s ability to anticipate patterns of antigenic

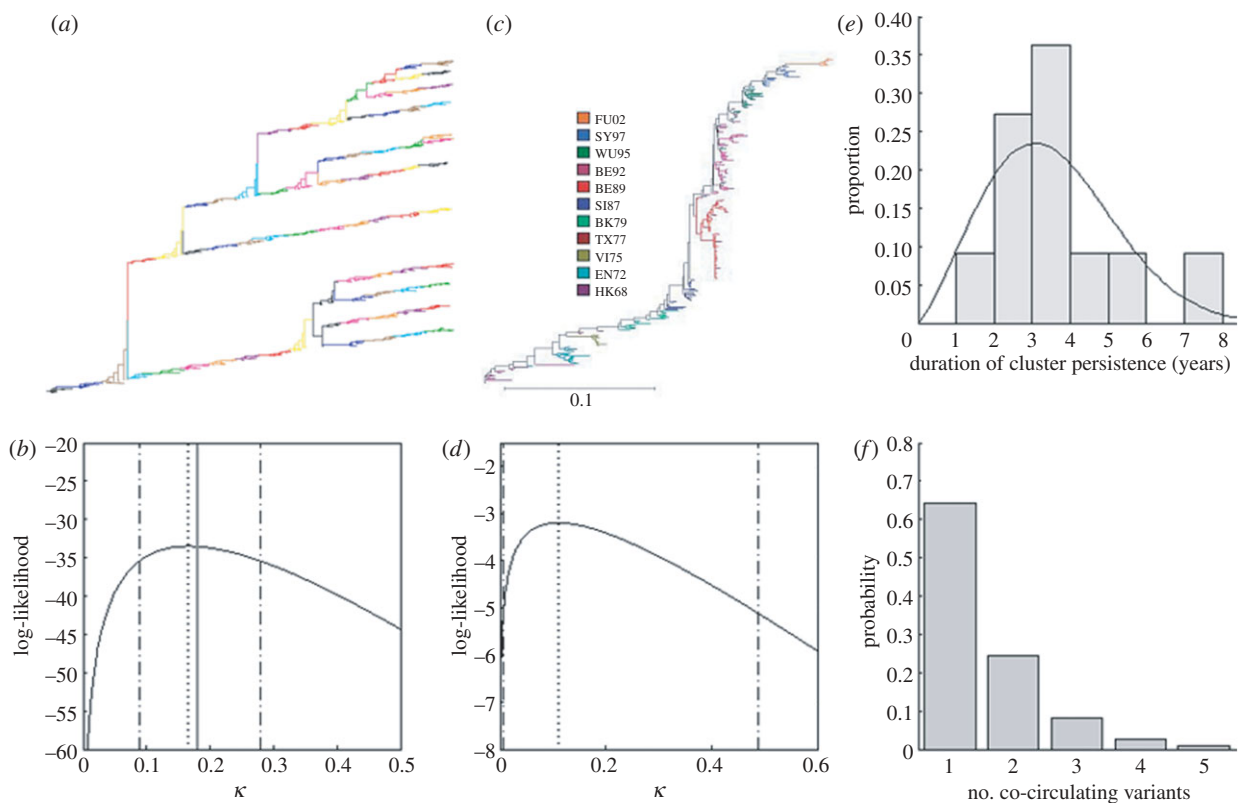


Figure 4. Applying the κ framework to the empirical host–virus systems. (a) An antigenically-typed phylogeny that was simulated using the stochastic forward model described in §2b. The model was simulated for 16 generations using a κ value of 0.18. (b) Log-likelihood values for a range of κ values, calculated from the likelihood expression provided in the text. Solid vertical line indicates the true value of $\kappa = 0.18$, dotted line is the maximum-likelihood estimate of κ and the dashed lines are the 95% CI. (c) An antigenically-typed phylogeny of influenza A/H3N2 from sequences isolated between 1968 and 2003. Reproduced from Koelle *et al.* [13]. (d) Log-likelihood values for a range of κ values calculated from the distribution of excess variants deduced from the phylogeny in (c). (e) The duration of time each antigenic cluster persisted, defined as the amount of time between when a cluster emerges and when it generates its first successful offspring variant. The mean cluster persistence time is 3.51 years. Black curve shows the $g_V(t)$ for the best estimate of k_W and λ_W , calculated as follows: I was first computed from estimates of the annual attack rate, N and v , u was then computed using equation (2.9), given estimates of R_0 and σ . Using these values of I and u , and for given values of k_W and λ_W the probability of observing each of the estimated cluster persistence times can be calculated using equation (2.12). The maximum-likelihood estimates of k_W and λ_W were found by searching (k_W and λ_W) parameter space for the maximum product of these probabilities, yielding $k_W = 2.26$ and $\lambda_W = 1162$. (f) The probability of observing a given number of co-circulating antigenic variants in 2011. These probabilities were computed using equation (2.15), knowing that there was only one circulating variant in 2003 (FU02) and that a generation lasts approximately 3.5 years, and assuming that $\kappa = 0.11$.

evolution in light of these simplifications, we compared antigenic diversity levels predicted by equation (2.15) against those observed from full simulations of equations (2.1)–(2.3), indicating good qualitative agreement (electronic supplementary material, figure S2). These simulations also show that a range of viral diversification patterns (cactus-like, acacia-like and intermediate phylogenetic tree shapes) can be reproduced under this framework. In the electronic supplementary material, we also further detail how the κ framework can be modified for a broader set of strain interactions.

3. APPLYING THE κ FRAMEWORK TO EMPIRICAL HOST–VIRUS SYSTEMS

Here, we detail two ways in which the κ framework can be used to understand and predict the evolutionary dynamics of viruses in empirical systems, using influenza A/H3N2 circulating in humans as a case study.

(a) Estimating κ from a phylogeny with antigenic classification of viral isolates

The forward model of antigenic evolution assumes that the number of excess variants generated by an endemic viral variant prior to its extinction is Poisson-distributed with mean κ . We can use this assumption to compute the probability of observing a given number of excess variants arising from a single circulating antigenic variant, and multiply these probabilities to compute the overall likelihood of the dimensionless number κ from the distribution of observed excess variants. Figure 4a,b provides an example, with a phylogeny simulated as in figure 3. The antigenically typed phylogeny shows that 60 out of the 72 endemic variants yielded zero excess variants, while the remaining 12 yielded a single excess variant. The likelihood function is therefore given by:

$$L(\kappa) = \left(\frac{\kappa^0 e^{-\kappa}}{0!} \right)^{60} \left(\frac{\kappa^1 e^{-\kappa}}{1!} \right)^{12}.$$

Evaluating this function for κ values between 0 and 0.5 clearly shows that our maximum-likelihood estimate of $\kappa = 0.167$ falls close to the true value of $\kappa = 0.18$ that was used in the simulation of the phylogeny.

Applying this same approach to influenza A/H3N2 in humans, we see that eight antigenic variants between 1968 and 2003 yielded zero excess variants, and only a single variant, SI87, yielded one excess variant, BE92 (figure 4c). (SI87's first successful offspring variant BE89 was lost from the population within a few years.) Our maximum-likelihood estimate of κ from this distribution of excess variants is 0.11 (figure 4d). Given a mean generation time of approximately 3.5 years (figure 4e), we can now use this estimate to probabilistically predict the number of co-circulating antigenic variants in three viral generations, bringing us roughly to today (figure 4f). Our framework predicts with a probability of 64 per cent that only a single antigenic variant will be circulating today. This is indeed the case, with antigenic variant Perth/09 currently circulating, having evolved from variant Brisbane/07 [17].

(b) Estimating κ directly from epidemiological parameters

To review, the seven fundamental parameters that are needed to estimate κ are N , ν , μ , R_0 (or β), σ and the antigenic emergence rate's parameters k_W and λ_W . Host-population size N , however, only plays a role in determining the equilibrium number of infected individuals I , and host lifespan similarly only plays this role if $(\mu + \nu) \approx \nu$. An estimate of I , along with estimates of ν , R_0 , σ , k_W and λ_W , are therefore only required to compute κ . Some of these parameters are well-known or routinely estimated. For example, for influenza A/H3N2, $1/\nu$ is roughly 3 days, R_0 is roughly 2 [18], σ is roughly 0.75 [19,20] and I can be estimated by assuming an annual attack rate of 6 per cent and a population size of six billion. This leaves k_W and λ_W , which can be estimated together from the distribution of cluster persistence times and estimates of I , R_0 and σ (figure 4e). Given these parameter estimates, we compute $\kappa = 0.50$, at the very top end of the 95% confidence interval (CI) using the method discussed above (figure 4d). Clearly, this value of κ is an overestimate, and we return to this disparity below.

4. DISCUSSION

Here, we proposed a dimensionless number κ for understanding and predicting the evolutionary dynamics of antigenically variable viruses. We derived this number from our knowledge of multi-strain epidemiological dynamics and the frequency-dependent selection pressures that arise from these dynamics. The dynamics depend only on a handful of parameters, many of which are known or can be estimated in a straightforward way from empirical data.

The κ framework presented here makes several predictions. One set of predictions is provided by the expression of κ , given by equation (2.4), approximated by equation (2.14), and depicted graphically in electronic supplementary material, figure S1: with all else equal, we expect a virus to exhibit more rapid antigenic diversification when circulating in a larger host population, when circulating in a shorter-lived host population and when

resulting in longer durations of infection. We also expect viruses with lower cross-immunity between antigenic variants to have lower rates of diversification. Finally, we expect the rate of antigenic diversification to strongly depend on the antigenic emergence rate's k_W parameter, controlling how the rate of antigenic emergence depends on a variant's age. As the effect of 'ageing' diminishes, k_W approaches one, yielding higher κ values. This relationship can therefore explain why epidemiological models that do not include this effect (such that $k_W = 1$) result in explosive genetic and antigenic diversity in the absence of other factors such as generalized immunity [21], whereas models that include this ageing effect can reproduce cactus-like evolutionary dynamics without invoking other factors [12,13,22].

In our application to influenza, we outlined two approaches that can be used to estimate κ . The first was based on an antigenically typed phylogeny and yielded reasonable κ estimates. The second was based on independent parameter estimates; in this case our calculation of κ was a considerable overestimate. This could have been owing to several reasons. First, our framework assumed a homogeneous, well-mixed host population and thereby ignored geographic subdivision and networks of interaction, which have clearly been shown to play a role in shaping viral phylogenies [23–25] and impacting antigenic diversity [26]. Second, we assumed no seasonality or other factors that could lead to significant variation in the number of infected individuals over time. These factors are likely to reduce the degree of antigenic diversity by increasing the probability of stochastic loss. Finally, viral strains may differ phenotypically in ways unrelated to their antigenicity, and these differences may also have impacted influenza A/H3N2's HA phylogeny. Future work should therefore focus on extending this minimal κ framework to improve our ability to predict patterns of antigenic diversification from independent parameter estimates.

Despite these limitations, the κ framework presented here was shown to reasonably anticipate the viral evolutionary dynamics of influenza A/H3N2's HA when κ was estimated from a phylogeny. Our estimate of κ , however, assumed a status-based multi-strain model, which has recently been criticized for its biological implausibility. Although this may seem like a major limitation, alternative multi-strain models (e.g. [27]) can easily be accommodated by modifying subcomponents of κ (electronic supplementary material).

From the framework we presented here, we can start to understand why viral phylogenies from certain host–virus systems differ topologically from those in other systems, or why they may have remarkable evolutionary similarities. Through this improvement in our understanding, we hope that κ may result in more effective disease prediction and control. For endemic viruses, the dependence of κ on modifiable epidemiological parameters already points us towards control strategies that may drive a given virus into an evolutionary regime that can be more effectively controlled. For example, the lengthening of livestock lifespans is expected to significantly lower the rates of antigenic diversification. In conjunction with vaccines that would be more effective and long-lasting, this could therefore lead to dramatic reductions in infection levels. For emerging viral zoonoses, κ could be estimated by combining what is

known about a naive host population (e.g. its population size) with what is known about the virus from its circulation in a reservoir population (e.g. its duration of infection). This could lead to predictions of how a virus will evolve in a host population prior to its establishment and help us focus our efforts on the viruses that would be difficult to control antigenically if establishment occurred. In sum, the minimal κ framework presented here, open to many extensions and modifications, could prove increasingly useful for understanding and controlling the evolutionary dynamics of antigenically variable RNA viruses.

We thank Greg Sawicki and the Koelle research group for helpful discussions. This work was funded by the James S. McDonnell Foundation and grant NSF-EF-08-27416 to K.K. O.R. was further supported by the Sir Henry Wellcome Fellowship WT092311MF; J.C.M. through grant DMS-0616710; D.R. through an NSF pre-doctoral fellowship; and K.K. by the RAPIDD programme of the Science and Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health.

REFERENCES

- Holmes, E. C. & Grenfell, B. T. 2009 Discovering the phylodynamics of RNA viruses. *PLoS Comput. Biol.* **5**, e1000505. (doi:10.1371/journal.pcbi.1000505)
- Rott, N. 1990 Note on the history of the Reynolds number. *Ann. Rev. Fluid Mech.* **22**, 1–11. (doi:10.1146/annurev.fl.22.010190.000245)
- Anderson, R. M. & May, R. M. 1991 *Infectious diseases of humans: dynamics and control*. Oxford, UK: Oxford University Press.
- Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A. & Holmes, E. D. C. 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332. (doi:10.1126/science.1090727)
- Andreasen, V. & Sasaki, A. 2006 Shaping the phylogenetic tree of influenza by cross-immunity. *Theoret. Popul. Biol.* **70**, 164–173. (doi:10.1016/j.tpb.2006.04.004)
- Sasaki, A. & Haraguchi, Y. 2000 Antigenic drift of viruses within a host: a finite site model with demographic stochasticity. *J. Mol. Evol.* **51**, 245–255.
- Earn, D. J. D., Dushoff, J. & Levin, S. A. 2002 Ecology and evolution of the flu. *Trends Ecol. Evol.* **17**, 334–340. (doi:10.1016/S0169-5347(02)02502-8)
- Fitch, W. M., Bush, R. M., Bender, C. A. & Cox, N. J. 1997 Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl Acad. Sci. USA* **94**, 7712–7718. (doi:10.1073/pnas.94.15.7712)
- Lindesmith, L. C., Donaldson, E. F., LoBue, A. D., Cannon, J. L., Zheng, D.-P., Vinje, J. & Baric, R. S. 2008 Mechanisms of GII.4 norovirus persistence in human populations. *PLoS Med.* **5**, 269–290. (doi:10.1371/journal.pmed.0050031)
- Siebenga, J. J., Vennema, H., Renckens, B., de Bruin, E., van der Veer, B., Siezen, R. J. & Koopmans, M. 2007 Epochal evolution of GII.4 norovirus capsid proteins from 1995 to 2006. *J. Virol.* **81**, 9932–9941. (doi:10.1128/JVI.00674-07)
- Gog, J. R. & Grenfell, B. T. 2002 Dynamics and selection of many-strain pathogens. *Proc. Natl Acad. Sci. USA* **99**, 17 209–17 214. (doi:10.1073/pnas.252512799)
- Koelle, K., Khatri, P., Kamradt, M. & Kepler, T. B. 2010 A two-tiered model for simulating the ecological and evolutionary dynamics of rapidly evolving viruses, with an application to influenza. *J. R. Soc. Interface* **7**, 1257–1274. (doi:10.1098/rsif.2010.0007)
- Koelle, K., Cobey, S., Grenfell, B. & Pascual, M. 2006 Epochal evolution shapes the phylodynamics of influenza A (H3N2) in humans. *Science* **314**, 1898–1903. (doi:10.1126/science.1132745)
- Bloom, J. D., Gong, L. I. & Baltimore, D. 2010 Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328**, 1272–1275. (doi:10.1126/science.1187816)
- Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E. & Fouchier, R. A. M. 2004 Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371–376. (doi:10.1126/science.1097211)
- Keeling, M. J. & Rohani, P. 2008 *Modeling infectious diseases in humans and animals*. Princeton, NJ: Princeton University Press.
- Yang, J. R. *et al.* 2010 A new antigenic variant of human influenza A (H3N2) virus isolated from airport and community surveillance in Taiwan in early 2009. *Virus Res.* **151**, 33–38. (doi:10.1016/j.virusres.2010.03.011)
- Truscott, J. *et al.* 2009 Quantifying the transmissibility of human influenza and its seasonal variation in temperate regions. *PLoS Curr. Influenza* e9548. (doi:10.1371/currents.RRN1125)
- Gill, P. W. & Murphy, A. M. 1977 Naturally acquired immunity to influenza type A: a further prospective study. *Med. J. Aust.* **2**, 761–765.
- Meiklejohn, G., Eickhoff, T. C., Graves, P. & Josephine, I. 1978 Antigenic drift and efficacy of influenza virus vaccines, 1976–1977. *J. Infect. Dis.* **138**, 618–624. (doi:10.1093/infdis/138.5.618)
- Ferguson, N. M., Galvani, A. P. & Bush, R. M. 2003 Ecological and immunological determinants of influenza evolution. *Nature* **422**, 428–433. (doi:10.1038/nature01509)
- Koelle, K., Kamradt, M. & Pascual, M. 2009 Understanding the dynamics of rapidly evolving pathogens through modeling the tempo of antigenic change: influenza as a case study. *Epidemics* **1**, 129–137. (doi:10.1016/j.epidem.2009.05.003)
- Holmes, E. C. 2008 Evolutionary history and phylogeography of human viruses. *Ann. Rev. Microbiol.* **62**, 307–328. (doi:10.1146/annurev.micro.62.081307.162912)
- Pybus, O. G. & Rambaut, A. 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550. (doi:10.1038/nrg2583)
- Lewis, F. *et al.* 2008 Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* **5**, 0392–0402.
- Buckee, C. O., Koelle, K., Mustard, M. J. & Gupta, S. 2004 The effects of host contact network structure on pathogen diversity and strain structure. *Proc. Natl Acad. Sci. USA* **101**, 10 839–10 844. (doi:10.1073/pnas.0402000101)
- Andreasen, V., Lin, J. & Levin, S. A. 1997 The dynamics of co-circulating influenza strains conferring partial cross-immunity. *J. Math. Biol.* **35**, 825–842. (doi:10.1007/s002850050079)