Research

Author for correspondence:
Miran Park
e-mail: miran.park@ucla.edu

Multiple scales of selection influence the evolutionary emergence of novel pathogens

Miran Park[1], Claude Loverdo[1], Sebastian J. Schreiber[2] and James O. Lloyd-Smith[1,3]

[1]Department of Ecology and Evolutionary Biology, University of California, 610 Charles E. Young Dr. South, Los Angeles, CA 90095, USA
[2]Department of Evolution and Ecology and Graduate Group in Applied Mathematics, University of California, Davis, CA 95616, USA
[3]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

When pathogens encounter a novel environment, such as a new host species or treatment with an antimicrobial drug, their fitness may be reduced so that adaptation is necessary to avoid extinction. Evolutionary emergence is the process by which new pathogen strains arise in response to such selective pressures. Theoretical studies over the last decade have clarified some determinants of emergence risk, but have neglected the influence of fitness on evolutionary rates and have not accounted for the multiple scales at which pathogens must compete successfully. We present a cross-scale theory for evolutionary emergence, which embeds a mechanistic model of within-host selection into a stochastic model for emergence at the population scale. We explore how fitness landscapes at within-host and between-host scales can interact to influence the probability that a pathogen lineage will emerge successfully. Results show that positive correlations between fitnesses across scales can greatly facilitate emergence, while cross-scale conflicts in selection can lead to evolutionary dead ends. The local genotype space of the initial strain of a pathogen can have disproportionate influence on emergence probability. Our cross-scale model represents a step towards integrating laboratory experiments with field surveillance data to create a rational framework to assess emergence risk.

1. Introduction

Emerging infectious diseases impose major health and economic burdens worldwide, and arise through a range of ecological and evolutionary mechanisms [1–3]. A recurring theme in many emergence events is that a pathogen lineage is exposed to a novel environment (e.g. a new host species or an antimicrobial drug) in which its fitness is reduced. When the initial pathogen genotype has fitness below the replacement level, the pathogen lineage will go extinct unless it adapts quickly enough to improve its fitness and successfully invade this new environment (e.g. new host species) or escape a lethal selection pressure (e.g. drug or vaccine) [4]. Adaptation can occur at several evolutionary stages and through different mechanisms, and key mutations may occur in the reservoir or the novel environment [5]. Here we focus on evolution in the novel environment, and we call this process evolutionary emergence. There is growing evidence that such adaptation has played an important role in host jumps of viruses such as influenza and severe acute respiratory syndrome coronavirus (SARS-CoV) [5,6]. Studying the evolutionary dynamics of this process, and linking theory to current empirical efforts that characterize the basic determinants of viral fitness, is an important frontier in understanding conditions that favour pathogen emergence. Developing theoretical tools allows us to assess possible emergence threats and what ecological and evolutionary mechanisms facilitate emergence. The acute need for such progress is evident from the recent controversy surrounding the

reports that just a few mutations are sufficient to enable airborne transmission of highly pathogenic H5N1 avian influenza virus among mammals [7–9].

Empirical research on pathogen evolution is defining the dimensions of the problem of evolutionary emergence. Notable steps have been taken towards mapping the fitness landscapes associated with pathogen emergence events, by measuring the fitness (or a proxy for fitness) of pathogen genotypes and effects of pertinent mutations [10]. Two studies have mapped the fitness landscapes associated with development of drug resistance in *Escherichia coli* and *Plasmodium falciparum* genes, by phenotyping all intermediate genotypes bearing some subset of the resistance mutations [11,12]. Another recent study has extended this comprehensive approach to a viral host jump, studying capsid protein mutations in canine parvovirus [13]. A related strategy, taken by the H5N1 influenza studies cited above [7,8] and across the literature for other emerging viruses such as SARS-CoV [5], is to characterize several traits associated with fitness for a more limited set of genotypes that comprise a putative pathway to emergence. A powerful complementary approach has tracked viral evolution *in vivo* by measuring changes in genotype frequencies in the course of experimental infection and transmission studies [7,8,14–16]. Ultimately, the aim is to connect these various experimental approaches to genotype frequencies detected in field surveillance, either before [17] or after [18,19] an emergence event occurs.

A conspicuous pattern arising from empirical studies is that measures of pathogen fitness (or fitness components) can be taken at different biological scales. For instance, recent studies of H5N1 influenza report cell receptor binding, viral titres in different tissues, *in vivo* replication kinetics, airborne transmission efficiency, and time to host death for a range of viral genotypes [7,8]. These diverse empirical measures of fitness support the need to distinguish within-host fitness, describing the pathogen's ability to grow within infected individuals, from between-host fitness, or transmissibility. Given a set of pathogen genotypes, we must define separate fitness landscapes corresponding to within-host and between-host fitness, i.e. each genotype has a fitness at both scales. This aligns with current research in other domains of infectious disease dynamics [7,8,16], and opens the possibility of conflicts, correlations or other interactions among selective forces acting at multiple scales, which can profoundly influence evolutionary outcomes [20–22].

The fitnesses of particular pathogen genotypes at within-host and between-host scales are not always positively correlated. Higher pathogen loads often lead to higher rates of transmission [23–25], in which case there may be positive correlation between fitnesses at the two scales. However, as explored in the extensive literature on virulence evolution, various costs can cause total transmissibility to decline if the pathogen load gets too high [26,27]. Different pathogen life histories and tissue tropisms may also influence the relationship between fitnesses across scales. For bloodborne pathogens, we would expect a positive correlation between pathogen load and infectiousness; indeed this is observed for HIV-1 set-point viral load, although a concomitant effect on the duration of infection causes total transmissibility to peak at intermediate viral load [23]. Thus, the correlation between within-host fitness (as reflected by viral load) and between-host fitness can be positive or negative. Similarly, pathogens that infect numerous tissue types, or that involve intermediate hosts or environmental stages in transmission, can exhibit complex relationships between fitnesses at the two scales. A well-known and relevant example is the tissue tropism of influenza virus, where higher binding affinity for different conformations of sialic acid on epithelial cells leads viruses to target the upper or lower respiratory tract. A viral mutation that increases affinity for the α-2,3 conformation might increase within-host replication while decreasing transmissibility by moving the infection deeper into the lung [28]. Such tissue tropism is thought to be a crucial determinant of host adaptation for influenza [7,8], so it is possible that cross-scale conflicts in selection play an important role in evolutionary emergence. We now know that circulating strains of H5N1 avian influenza are within a few mutations of genotypes that transmit much more efficiently among mammals [7,17,29]. Many mammals (human and otherwise) have been infected with H5N1 influenza—why have not these transmissible genotypes arisen, given that they certainly would confer a fitness benefit to the virus in mammal populations? One possible explanation is that these transmissible genotypes (or intermediate genotypes en route to them) are less fit at the within-host scale so they might not rise to high enough frequencies within hosts to realize their transmission advantage.

Translating our growing empirical knowledge of pathogen phenotypes into an improved understanding of emergence risks will require analytical methods to integrate the key mechanisms across scales. Theoretical study of pathogen emergence has previously focused on evolutionary invasion at the host population scale: an introduced pathogen exhibits weak transmission in the novel host environment, and must mutate to higher transmissibility before the transmission chain dies out. Stochastic models such as multi-type branching processes have been used to compute the probability of emergence for simple genotype spaces and corresponding (between-host) fitness landscapes [4,30]. These studies have yielded important insights, showing that even when initial transmissibility is too low to start an epidemic, higher values of transmissibility (bringing the pathogen closer to the threshold for sustained spread) lead to greatly increased probability of evolutionary emergence [30]. Subsequent work has explored the influence of epidemiological complexities [31–33], but key elements of the evolutionary dynamics have not yet been addressed. Crucially, the model parameters describing evolutionary change of the pathogen have been assumed not to depend on the fitnesses of the genotypes involved. Within-host fitness, and the consequent action of within-host selection, has not been included. André & Day [34] contributed the valuable extension of considering selective sweeps during the course of an individual's infection, but similar to the previous work the rate of fixation of new mutants was assumed not to depend on the strength of selection within hosts. These omissions separate the current theory from the empirical evidence, which largely focuses on within-host fitness [5], and overlook the fact that selection acts most immediately within a host, as pathogen genotypes compete with each other for target cells or other resources or to escape the immune system [35–37].

We present a theoretical framework to study how the evolutionary emergence of pathogens is influenced by selection at within-host and between-host scales. Our aim is to create a tractable cross-scale model from which analytical

insights and biological intuition can be derived. We represent the between-host scale using multi-type branching processes as in previous models [30,34]. However, instead of assuming equal rates of mutation between all pairs of genotypes, we introduce a sub-model for within-host selection based on population genetic theory. In particular, we follow the approach used in recent analyses of mutational trajectories in empirical fitness landscapes [11,12,38] and apply the strong selection, weak mutation (SSWM) limit to derive a compact representation of adaptive evolution [39]. Using this framework, we analyse how fitness landscapes at within-host and between-host scales can interact to influence the probability that a pathogen lineage will emerge. Here we focus on the mechanisms involved in host jumps of pathogens, because our model describes invasion of a pathogen into a large susceptible population; later we discuss how this model could be applied to other emergence situations such as developing resistance to an antimicrobial drug. At the within-host scale, selection acts on relative fitnesses of adjacent genotypes, with strong selection leading to rapid fixation of new beneficial mutants. At the between-host scale, we consider a stochastic transmission framework that depends on the absolute fitness of neighbouring genotypes, where individuals are infected with a particular genotype. We explore two scenarios of simple genotype spaces, illustrating basic principles of multi-scale selection in this context, and exploring the potential for emergence to be prevented by evolutionary conflicts across scales. We hope that this cross-scale mechanistic model begins to bridge the gap between the growing body of empirical data from laboratory experiments and pathogen sequencing studies, and large-scale public health questions about emergence risk. We conclude by discussing necessary extensions and possible links to empirical studies.

2. A cross-scale model of evolutionary emergence

(a) Defining the system

Studying the evolutionary dynamics of pathogen populations at multiple scales can lead to substantial complexity, so it is necessary to make simplifying assumptions. Following earlier work [30,34], we assume that each infected host has a single pathogen genotype at any point in time, and we characterize the host individual by this type. Parameters are marked with a subscript or superscript i corresponding to the pathogen genotype in question. We analyse evolutionary dynamics on a defined genotype space, which consists of a set of pathogen genotypes and the pathways of mutation that connect them. A mutation is broadly defined as a change at a specific locus in the genome giving rise to a new genotype; this can include point mutations, insertions, deletions or other mechanisms of genetic change. Each pathogen genotype has two measures of fitness associated with it, corresponding to the within-host and between-host scales; these define two fitness landscapes over the genotype space. At the between-host scale, the fitness of the pathogen corresponds to its ability to transmit through the population. At the within-host scale, the fitness of the pathogen describes how well it replicates within a host. For our analyses, we create case studies of fitness landscapes, and explore how they interact to drive pathogen evolution.

The between-host fitness of genotype i is given by the reproductive number, $R_0^{(i)}$, which is the average number of secondary infections caused by a type i host in a completely susceptible population. For our evolutionary emergence problem, we consider a pathogen that is initially maladapted to the novel environment, i.e. genotype 1 has $R_0^{(1)} < 1$. Such a pathogen causes short chains of transmission but goes extinct with certainty if it does not evolve. Through mutation and selection, which we treat as within-host processes, new genotypes can arise and fix in some host individuals. Eventually, the pathogen lineage may reach an 'emergence genotype' with $R_0^{(i)} > 1$, which has a non-zero chance of successfully invading the new host population.

For our numerical work, we consider simple scenarios for which the initial and intermediate genotypes always have $R_0^{(i)} < 1$, and there is only one emergence genotype. We calculate the probability that the emergence genotype arises and successfully invades the host population, $P(\text{emergence})$, using techniques described below. Calculating $P(\text{emergence})$ allows us to compare interactions between different fitness landscapes, lending an understanding of general trends that arise as selection acts across scales.

(b) Between-host transmission dynamics

Building on existing literature in evolutionary emergence [34], we use a continuous-time multitype branching process to model the stochastic dynamics of transmission, recovery and genotype change at the population scale. The model tracks the population dynamics of infected individuals, which are classified according to the pathogen genotype of their infection. We assume a well-mixed homogeneous population in which the number of susceptibles is large enough that it is not significantly depleted by the limited number of cases that occur before pathogen emergence.

Each infected host of type i infects other host individuals at a constant rate b_i, giving rise to an additional infected host of the same type, and ceases to be infectious (through recovery or death) at a rate d_i. The reproductive number for type i is $R_0^{(i)} = b_i/d_i$. Within-host evolutionary processes cause the dominant genotype to change from type i to type j at a rate $m_{i,j}$ during the course of an individual's infection, where $m_{i,j} = 0$ for a genotype j that is more than one mutational step away from genotype i. During a small time interval of length Δt, these events occur with approximate probabilities $b_i \Delta t$, $d_i \Delta t$ and $m_{i,j} \Delta t$, respectively.

(c) Within-host evolutionary dynamics

Previous models of evolutionary emergence assume that substitution rates do not depend on the fitnesses of the genotypes involved. Here, we replace this assumption with a mechanistic model for within-host evolution, which we embed within the branching process framework used for population-scale dynamics. To represent the key population genetic mechanisms in a compact manner, we use the SSWM paradigm [39].

In the SSWM limit, strong selection means that only beneficial mutations are considered, and mutation rates are sufficiently low that simultaneous mutation events can be neglected. The simplicity of the SSWM limit arises because beneficial mutations go to fixation much faster than new mutations arise, so at any point in time the population is essentially fixed for some genotype. This fixed genotype

can only be displaced by pathogen genotypes with higher within-host fitness.

The SSWM assumption allows changes in the infectious genotype within the host to be modelled as a continuous time Markov chain [39]. We begin by defining the absolute within-host fitness of a particular genotype i as w_i. The relationship between the absolute within-host fitness of genotype i and that of a different genotype j is $w_j = (1 + s_{i,j})w_i$, where $s_{i,j}$ is the selection coefficient of the genotype j invading a system with genotype i at its equilibrium. If the current genotype within the host is type i and a substitution occurs, the probability that type j fixes next is given by $s_{i,j}/\sum_{k \in M_i} s_{i,k}$, where M_i is the set of genotypes that are a single mutational step away from genotype i. The waiting time before the next jump occurs is dependent on the size of the virus population, N, and the mutation rate, μ, and is exponentially distributed with mean proportional to $1/(N\mu \sum_{k \in M_i} s_{i,k})$ [39]. Therefore, in the SSWM limit, we can express the substitution rate for each genotype j

$$m_{i,j} \propto N\mu s_{i,j}. \tag{2.1}$$

The population size N and mutation rate are assumed to be constant; in appendix A, we present a derivation of equation (2.1) from a model of within-host viral dynamics which leads to an alternative interpretation of these quantities when SSWM is applied to viruses. From this derivation, we are able to make intuitive connections between our model and traditional ideas in population genetics, broadly supporting the use of the SSWM framework for within-host evolution. Calculating $m_{i,j}$ from equation (2.1) requires a constant of proportionality; for our numerical calculations, we set this constant to 0.4 following the original assumption by Gillespie (who interpreted it as a measure of the strength of selection) [39]. While this choice is arbitrary, it does not affect the qualitative results, as it affects all substitution rates equally and the timescales of these processes are otherwise arbitrary.

The SSWM model for within-host evolution means that a higher relative fitness of a neighbouring genotype leads to a faster rate of substitution, so in general each step through genotype space has a different speed at which it occurs. The biological basis for this effect derives from the probability of fixation of a new genotype when it first arises within a host. A greater fitness advantage for the new genotype leads to a higher likelihood that it will fix after it arises. Consequently, even if all neighbouring genotypes arise at the same rate, the rate of substitution is faster when the relative fitness difference is large.

We first consider a simple, sequentially connected chain of genotypes, where for each genotype, there exists only one 'neighbouring' genotype that is more fit. If the Lth genotype is the emergence genotype with $R_0^{(L)} > 1$, then we can derive an approximation for the probability of emergence, combining elements of arguments from Iwasa et al. [4] and André & Day [34], and using branching process theory

$$P(\text{emergence})_{\text{seq.}} \approx \left(\frac{1}{1 - R_0^{(1)}}\right)\left(\frac{m_{1,2}}{m_{1,2} + d_1}\right) \cdots \left(\frac{1}{1 - R_0^{(L-1)}}\right)$$
$$\times \left(\frac{m_{L-1,L}}{m_{L-1,L} + d_{L-1}}\right)\left(1 - \frac{1}{R_0^{(L)}}\right). \tag{2.2}$$

This expression breaks down into three biologically meaningful factors. Each factor of $1/(1 - R_0^{(i)})$ is the expected number of infections in a subcritical chain of transmission initiated by a type-i individual in the absence of evolution. The factors $m_{i,j}/(m_{i,j} + d_i)$ give the probability of the fixed genotype changing from type i to type j before recovery or death of a host infected with type i. The final factor, $1 - 1/(R_0^{(L)})$, is the probability that the emergence genotype will successfully invade the host population if it arises in a single host individual.

We can extend this approximation to the more general case of an arbitrary genotype space. To estimate the probability of emergence starting with one infected individual of type 1, let $L - 1$ be the minimal number of mutational steps from the initial genotype to an emergence genotype. The probability of emergence will be proportional to μ^{L-1} as longer paths to emergence add terms of order μ^L or higher (though note that factors of μ are implicit in $m_{i,j}$). Let $\mathcal{P}$ be the set of mutational pathways of length L, each spanning genotype i_1 to an emergence genotype i_L. Then the probability of emergence can be approximated as

$$P(\text{emergence}) \approx \sum_{(i_1, \ldots, i_L) \in \mathcal{P}} \left(1 - \frac{1}{R_0^{(i_L)}}\right)$$
$$\times \prod_{k=1}^{L-1} \frac{1}{1 - R_0^{(i_k)}} \frac{m_{i_k, i_{k+1}}}{(m_{i_k, i_{k+1}} + d_{i_k})}. \tag{2.3}$$

Each term within the summation corresponds to a particular mutational pathway, and matches the approximation shown in equation (2.2). The low mutation rate assumption allows us to neglect outcomes where more than one virus lineage reaches emergence. In the analyses presented below, we illustrate that the approximation works well through most of the parameter range considered.

(d) Calculating the probability of emergence

The branching process model gives us a framework to calculate the probability of emergence, $P(\text{emergence})$. For our models, there are two ways to compute the emergence probabilities from the embedded discrete-time branching process: numerically using standard methods [40] or using the exact solutions we derive in appendix B. While both approaches yield the same results, we present only results based on the exact solutions. To gain intuition into the determinants of emergence, we also present a simple approximation for the probability of emergence in the limit of low initial between-host fitness (low $R_0^{(1)}$) and low mutation rates.

3. Effects of cross-scale selection on pathogen emergence

We analyse two scenarios to explore the possible influence of multiple scales of selection on evolutionary emergence. In the first scenario, we consider a simple genotype space, and a basic set of qualitatively distinct fitness landscapes, to understand the fundamentals of how fitness landscapes at the two scales interact to produce evolutionary outcomes. In the second scenario, we extend these fitness landscapes to consider multiple competing pathways of pathogen evolution, creating the potential for conflict across scales.
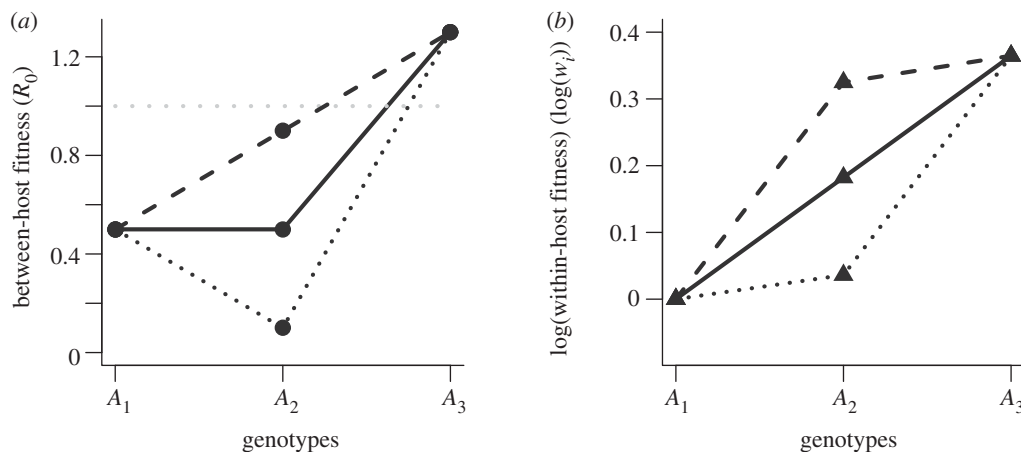
Figure 1. Fitness landscapes at between-host and within-host scales. Example landscapes are shown for a simple genotype space, with three genotypes connected sequentially. For this illustrative example, we consider landscapes where the fitnesses of initial (A_1) and emergence (A_3) genotypes are fixed, and only the fitness of the intermediate genotype (A_2) varies. (a) At the between-host scale, the intermediate genotype may be more fit (yielding an 'uphill' landscape (dashed line)), of equal fitness ('jackpot' (solid line)) or less fit ('valley' (dotted line)) than the initial genotype but only the emergence genotype A_3 has $R_0 > 1$. (b) At the within-host scale, under the SSWM model, the gain in relative fitness determines the rate of substitution between neighbouring genotypes. The 'equal-rate' landscape arises when relative fitness gains are equal for each step, the 'fast$-$slow' landscape when the first step yields greater fitness gain than the second, and the 'slow$-$fast' landscape in the reverse case. Note that we plot the logarithm of the absolute within-host fitnesses w_i, as noted in the text, because $w_j = (1 + s_{i,j})w_i$. We have normalized the fitnesses such that $w_1 = 1$. ($R_0^{(i)}$ (uphill) = [0.5,0.9,1.3], $R_0^{(i)}$ (jackpot) = [0.5,0.5,1.3], $R_0^{(i)}$ (valley) = [0.5,0.1,1.3], w_i (equal $-$ rate, solid, line) = [1,1.2,1.44], w_i (fast $-$ slow, dashed line) = [1,1.38,1.44], w_i (slow $-$ fast, dotted line) = [1,1.04,1.44].

(a) Scenario 1: exploring interactions between scales of selection in a simple genotype space

We consider a simple genotype space, with three genotypes sequentially connected in a chain, and explore how selection at different scales impacts disease emergence (figure 1). To distinguish fitness landscapes in our scenarios from the general theoretical results presented above, we refer to these particular genotypes by a capital letter and numerical subscript i, e.g. genotype A_1.

At the between-host scale, we consider three fitness landscapes (figure 1a). We explore scenarios where the initial and emergence genotypes have fixed fitnesses, and explore the landscapes arising from differing fitnesses of the intermediate genotype. In the 'jackpot' landscape, between-host fitness does not change until the pathogen reaches the emergence genotype (and thus hits the jackpot) ($R_0^{(1)} = R_0^{(2)} < R_0^{(3)}$). In the 'uphill' landscape, the fitness increases with each step through genotype space ($R_0^{(1)} < R_0^{(2)} < R_0^{(3)}$). We arbitrarily choose fitnesses that increase linearly for this example. In the 'valley' landscape, the fitness of the intermediate genotype is lower than the fitness of the initial genotype, so the pathogen must traverse a valley of lower fitness to reach emergence ($R_0^{(1)} > R_0^{(2)} \ll R_0^{(3)}$) (figure 1a). For simplicity, in all of our examples, we assume rates of recovery or death (d_i) are equal across genotypes. Variation in the recovery or death rates d_i leads to qualitatively similar results, though the probabilities of emergence increase more rapidly with $R_0^{(1)} = R_0^{(2)}$ because the rising reproductive numbers correspond to longer infectious periods $1/d_i$, allowing more time for substitution events to occur [34].

At the within-host scale, only pathways with increasing fitness are relevant under the SSWM framework, so we consider three cases that span the qualitative range of possible fitness landscapes, given that we fix the fitnesses of the initial and emergence genotypes (figure 1b). We define the 'equal-rate' landscape as the case that has equal gains in relative fitness when moving from the initial to the intermediate genotype, and from the intermediate to the emergence genotype. Under the SSWM model for within-host evolution, this yields equal substitution rates for the two mutational steps ($m_{1,2} = m_{2,3}$). We note that the equal-rate landscape under SSWM corresponds to previous models that have assumed equal substitution rates and no back-mutations. The 'fast$-$slow' landscape has a greater fitness gain from the initial to the intermediate genotype than from the intermediate to the terminal genotype; thus the substitution rate for the first substitution is faster than the second ($m_{1,2} > m_{2,3}$). The 'slow$-$fast' landscape is the opposite case, with the substitution rate for the first substitution slower than the second ($m_{1,2} < m_{2,3}$). We assume a symmetry of fitnesses between the fast$-$slow and slow$-$fast landscapes, for ease of comparison: $m_{1,2}$ in the fast$-$slow case is equal to $m_{2,3}$ in the slow$-$fast case and vice-versa. To depict the within-host fitness landscapes, we plot the logarithm of the absolute fitnesses w_i. This emphasizes the multiplicative relationships that define relative fitnesses which drive the SSWM framework. For example, the equal-rate landscape is linear in log-scaled absolute fitness (figure 1b).

The approximation (equation (2.2)) shows that the probability of emergence is maximized when $m_{1,2} = m_{2,3}$. This is because, when the fitnesses of the initial and emergence genotypes are fixed, the product of substitution rates is maximized when the rates are equal (and hence when the relative fitnesses for each genotype transition are equal). This in turn maximizes the overall probability of emergence, since faster substitution means less chance that the competing recovery rates d_i will prevail. This outcome can also be explained through Jensen's inequality [41], because the logarithm of the product of terms $m_{i,j}/(m_{i,j} + d_i)$ in equation (2.2) is concave down as a function of $m_{i,j}$. Thus, we expect anything other than the equal-rate case to have lower probability of emergence, because variation in the $m_{i,j}$'s decreases the value of this product. We test this prediction
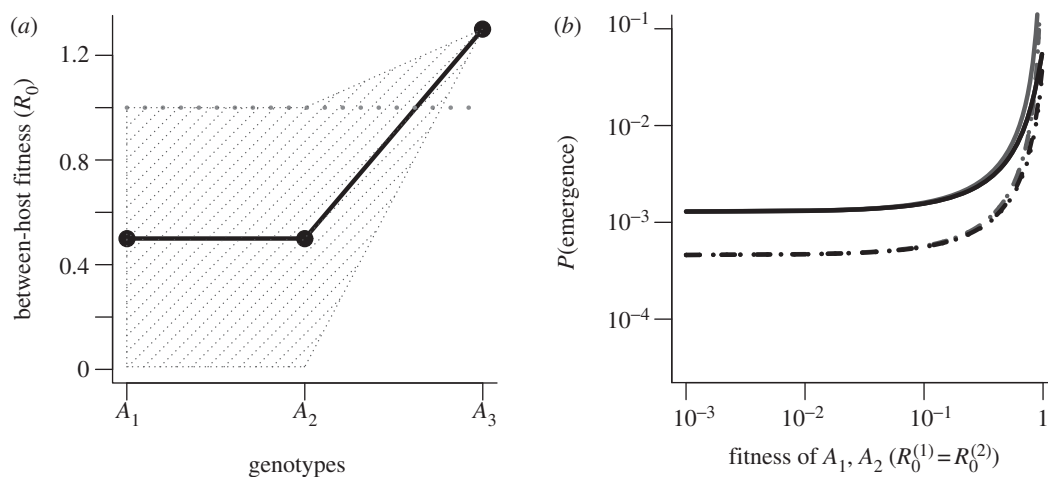
Figure 2. Example of the interaction between selection at different scales. (*a*) We consider a set of jackpot landscapes at the between-host scale, with a range of fitness values for the initial (A_1) and intermediate (A_2) genotypes (indicated by the shaded region). (*b*) The probability of emergence as a function of the between-host fitness of the A_1 and A_2 genotypes, showing the interaction with the three within-host landscapes. The equal-rate (solid line) landscape gives the highest probability of emergence because it gives the fastest substitutions overall. Results for fast–slow (dashed line) and slow–fast (dotted line) landscapes overlap almost exactly. Black lines show exact solutions; grey lines show the approximation from equation (2.2), which fits well until $R_0^{(1)}$ and $R_0^{(2)}$ approach 1 ($d_1 = d_2 = d_3 = 1$; $b_3 = 1.3$, b_1 and b_2 adjusted to yield desired $R_0^{(i)}$ values; $N = 10^6$, $\mu = 10^{-6}$, constant of proportionality = 0.4).

and illustrate the interplay between fitness landscapes at different scales, by considering how the probability of emergence for a jackpot between-host landscape is affected by different within-host fitness landscapes (figure 2*a*). The equal-rate scenario has the highest probability of emergence, as we predicted; we also see that the approximation (equation (2.2)) is quantitatively accurate through most of the parameter range considered (figure 2*b*). The fast–slow and slow–fast cases have virtually identical probabilities of emergence, given the jackpot between-host landscape and our assumption of symmetry between the fast–slow and the slow–fast landscapes.

We can explore the different qualitative interactions across scales by varying the intermediate values for both within-host and between-host fitness landscapes, fixing the initial and emergence fitnesses at both scales. Figure 3*a* shows how the probability of emergence varies across a range of possible intermediate values, spanning from valley to uphill landscapes for the between-host scale, and from slow–fast to fast–slow landscapes at the within-host scale. The probability of emergence increases going from a valley to uphill between-host landscape (i.e. along the horizontal axis), as expected intuitively and known from earlier studies [4,30,34]. Considering different within-host landscapes, we see that the probability of emergence is maximal close to the equal-rate case, as predicted from the approximation, but deviations from this pattern arise from interactions between the fitness landscapes at each scale. For clarity, we focus on the three within-host landscapes shown in figure 1*b*, and track how the probability of emergence varies as the between-host fitness of the intermediate state increases (figure 3*b*; shown as slices of the plot in figure 3*a*). When the intermediate between-host fitness is greater than the initial fitness ($R_0^{(2)} > R_0^{(1)}$), it is more advantageous for the pathogen to mutate immediately and gain the between-host fitness advantage so the fast–slow scenario is more favourable for emergence. When the intermediate fitness is below the initial fitness ($R_0^{(2)} < R_0^{(1)}$), it is more advantageous for the pathogen to spend less time in the intermediate state, so the slow–fast

scenario is more favourable for emergence (figure 3*b*). Based on these arguments, we would expect the curves for the slow–fast and fast–slow cases to cross at $R_0^{(2)} = 0.5$, the fitness of the initial genotype. However, the crossing point is shifted slightly in favour of the fast–slow scenario, reflecting an additional evolutionary benefit to spending more time in the A_2 genotype. All else equal, it is beneficial to spend more time in the A_2 genotype than the A_1 genotype, because all new cases infected by an A_2-infected individual are born into the A_2 genotype (and have a chance of mutating directly to the A_3 genotype) and thus have a head-start towards emergence. (This effect also causes the slight inequality between emergence probabilities for the slow–fast and fast–slow landscapes in figure 2*b*.) This scenario illustrates how selection can interact across scales in non-obvious ways, as the geometry of the within-host fitness landscape can shift between-host outcomes and change the expected probabilities of emergence.

(b) Scenario 2: alternative pathways illustrate the potential for conflict across scales

To explore the potential for conflicts in selection pressure across scales to influence pathogen emergence, we extend our analysis to a more complex scenario where two neighbouring mutations are available to the initial genotype B_0: one that leads to a pathway of decreasing between-host fitness and eventual extinction (B_1', B_2'), and one that leads to a pathway of increasing between-host fitness and possible emergence (B_1, B_2). We assume that these pathways have linearly decreasing or linearly increasing between-host fitness values, respectively (figure 4*a*). As a first exploration of interactions across scales, we consider simple within-host scenarios by fixing the extinction pathway (B_1', B_2') to have a particular equal-rate landscape and exploring the space of possible equal-rate landscapes for the emergence pathway (B_1, B_2) (figure 4*b*). This creates a potential conflict at the two scales for some pathways (i.e. when the within-host landscape for the emergence pathway (B_1, B_2) is relatively flat) as
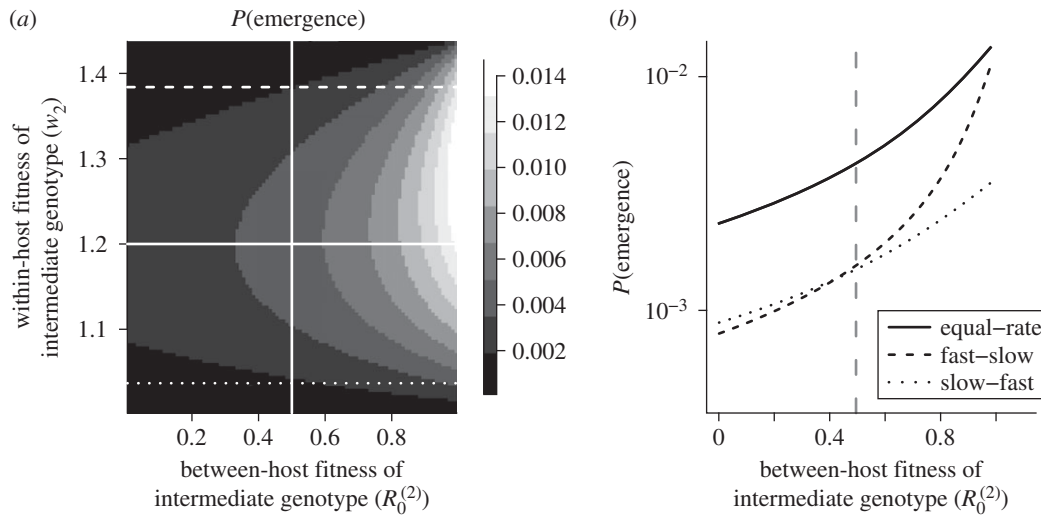
Figure 3. Cross-scale interactions for all combinations of landscapes in a three-genotype chain. (a) Heat-map showing the probability of emergence as the intermediate genotype is varied in fitness at both the within-host and between-host scales. Fitnesses of the initial and emergence genotypes are fixed, as in figure 1. The vertical solid line marks the between-host jackpot scenario and the horizontal solid line marks the within-host equal-rate scenario. The dotted and dashed lines represent the slow−fast (bottom) and fast−slow (top) from figure 1b, respectively. (b) Probability of emergence as a function of the between-host fitness of the intermediate genotype, $R_0^{(2)}$, for three within-host landscapes (solid line, equal-rate; dashed line, fast−slow; dotted line, slow−fast), corresponding to the horizontal lines in (a). The grey vertical shows where $R_0^{(2)} = 0.5$, i.e. the same value as $R_0^{(1)}$.
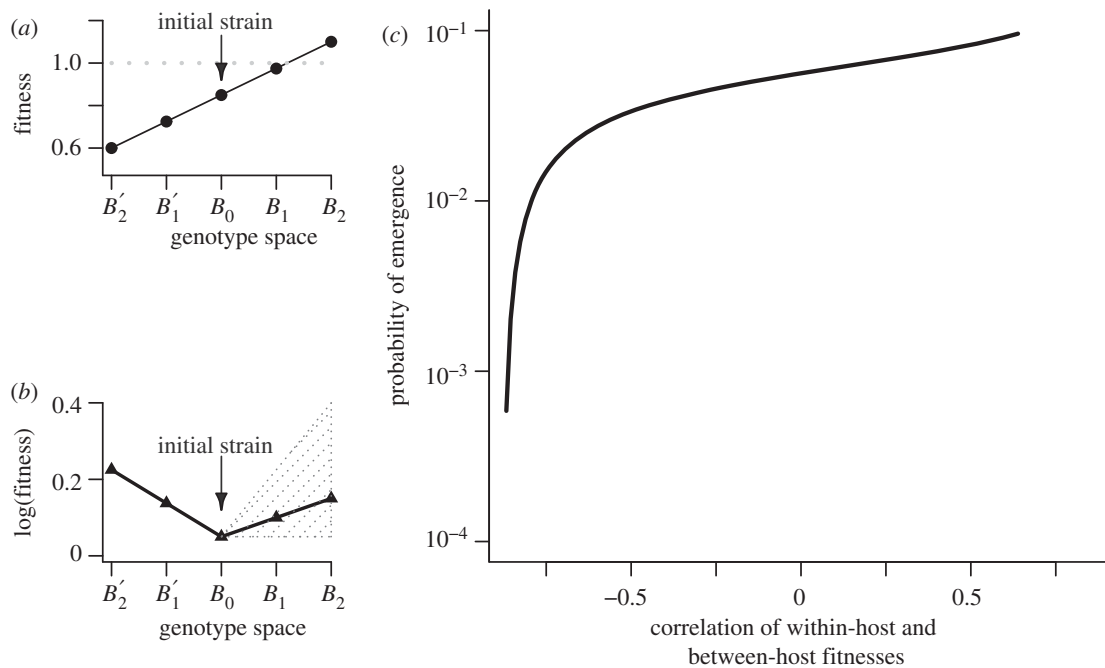

Figure 4. Alternative pathways can lead to conflict between scales. We assume the initial case is infected with genotype B_0, and two evolutionary pathways are available to the pathogen population. (a) At the between-host scale, fitness increases along the B_1, B_2 pathway, leading to possible emergence; fitness decreases along the B'_1, B'_2 pathway, leading to certain extinction. (b) At the within-host scale, we assume equal-rate landscapes on both pathways, but vary the slope of the B_1, B_2 pathway (as indicated by the shaded area). Depending on the slope of the B_1, B_2 pathway at the within-host scale, substitution rates towards B_1 and B_2 can be higher or lower than towards B'_1 and B'_2. (c) The correlation between fitnesses at within-host and between-host scales is a strong determinant of the probability of emergence. Positive correlations favour emergence, while negative correlations (which corresponds to lower slopes of the within-host B_1, B_2 landscape) cause the pathogen lineage to be drawn towards lower between-host fitness and extinction. We use Pearson's correlation coefficient to measure the linear dependence between the fitnesses at the two scales. ($R_0[B'_2, B'_1, B_0, B_1, B_2] = [0.2, 0.45, 0.7, 0.95, 1.2], d_i[B'_2, B'_1, B_0, B_1, B_2] = [0.1, 0.1, 0.1, 0.1, 0.1, 0.1], w_i[B'_2, B'_1, B_0, B_1, B_2]$ were drawn uniformly from the ranges [1.44, 1.2, 1, 1.01 − 2.24, 1.02 − 2.25], within-host parameters N and μ as in previous figures).

within-host selection favours the pathway that leads to extinction at the between-host scale. We summarize this effect with the Pearson's correlation coefficient between the fitness values at the within-host scale (w_i) and the fitness values at the between-host scale ($R_0^{(i)}$) for each genotype. When within-host fitness is negatively correlated with

between-host fitness (i.e. when (B_1, B_2) is flat), the probability of emergence is low. The emergence probability drops drastically as the negative correlation becomes stronger, as the lineage almost always evolves into the extinction pathway; in effect, the lineage is lured into an evolutionary dead end. When within-host fitness is positively correlated with
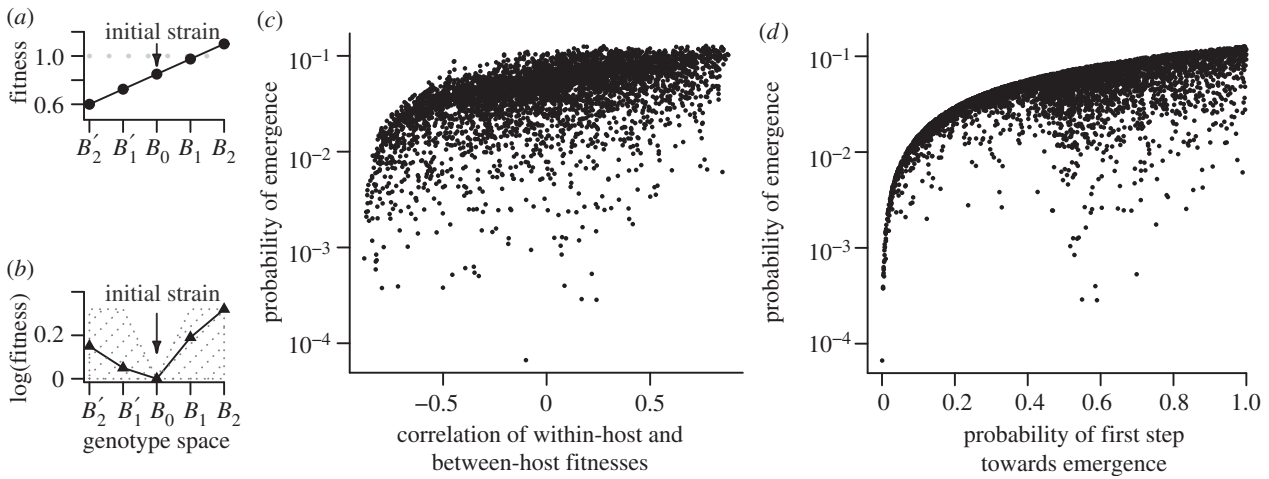
Figure 5. Exploration of alternative pathways for general within-host landscapes. This figure generalizes figure 4, such that (*a*) the between-host landscape is identical, but (*b*) the within-host landscapes sample a much broader set of possibilities. Within-host fitness of genotypes B'_2, B'_1, B_0, B_1 and B_2 were selected randomly, such that $w_i[B'_2, B'_1, B_0, B_1, B_2]$ were drawn uniformly from the ranges $[1.02 - 2.25, 1.01 - 2.24, 1, 1.01 - 2.24, 1.02 - 2.25]$. (*c*) The probability of emergence shows a positive association with the correlation between fitnesses across scales, though with considerable scatter. (*d*) The probability of emergence is more strongly associated with the probability that the first substitution event is towards genotype B_1 (i.e. towards possible emergence). Points for plots in (*c*) and (*d*) show the results of 5000 simulated within-host landscapes. (All parameters as in figure 4.)

between-host fitness, then the probability of emergence is higher as the lineage almost always evolves along the emergence pathway (figure 4*c*).

To explore the generality of these insights, we examine a much broader set of scenarios by assigning random values to the within-host fitnesses of all genotypes (B'_1, B'_2, B_1, B_2) (figure 5*a,b*). The positive association between the probability of emergence and the correlation of fitnesses across scales is maintained (figure 5*c*). There is significant scatter in the relationship, because the correlation is influenced by the fitnesses of genotypes B_2 and B'_2, which may have minimal influence on the probability of emergence depending on the fitnesses of B_1 and B'_1. Thus, having a high correlation between the two fitnesses at the two scales does not necessarily mean the lineage will be drawn towards the emergence pathway by within-host selection. Additionally, because correlation describes the linear dependence between the fitnesses at both scales, it becomes a less appropriate measure given the nonlinearity of the within-host fitness values. To clarify this relationship, we plot the probability of emergence versus the probability that the first mutational step is to genotype B_1, and therefore the emergence pathway (figure 5*d*). This shows a strong positive relationship with less scatter, indicating that the probability of emergence is influenced powerfully by which evolutionary pathway is taken by the pathogen population, and hence by the within-host fitnesses of the mutational neighbours of the introduced strain. The residual scatter comes from randomly generated landscapes for genotypes B_1 and B_2 that correspond to the fast−slow scenario. A high within-host fitness w_1 leads to a high probability of stepping towards emergence, but then the fitness w_2 is only marginally higher, so the substitution rate $m_{1,2}$ is slow, and there is a high likelihood that the lineage never reaches emergence. The strong influence of the first mutational step partially results from the absence of back-mutations (a consequence of strong selection), which means if the first mutational step is towards the extinction pathway (which may be favourable at the within-host scale despite its cost at the between-host scale) the pathogen is unable to reach emergence.

4 Discussion

We have presented a cross-scale model of evolutionary emergence of pathogens, drawing on population genetic theory to embed a mechanistic model for within-host selection into a branching process model for population-scale emergence. Our results show that within-host fitness plays an important role in evolutionary emergence and that interactions between selection pressures at the within-host and between-host scales can have a substantial effect on the probability of emergence. A growing number of studies are mapping the structure of within-host fitness landscapes for pathogens [42–44], making it clear that within-host selection plays a non-trivial role in real-world emergence scenarios. At the same time, empirical research has started to measure fitness at multiple scales [7,8] and track cross-scale evolution [15,16,45] for pathogens linked to emergence events. Improving our understanding of pathogen emergence in novel environments requires integration of evolutionary and ecological phenomena across scales [5]. Our model provides a framework to begin this integration, offering the potential of coupling phenotypic data from experimental studies to pathogen genotypes detected in field surveillance.

The most important results of our analysis are the qualitative insights about the relative risk of different emergence pathways. Our simulations illustrate two key points about the interactions between selection at the within-host and between-host scales when multiple evolutionary trajectories are available. First, and most broadly, positive correlations between the fitnesses across scales increase the likelihood of emergence. Because within-host selection drives movement through genotype space, this conclusion is consistent with theory showing that positive correlations between fitness and dispersal patterns increase the establishment likelihood of invasive species [46]. This also echoes themes from prior theoretical studies of the influence of cross-scale selection on the evolution of virulence, which have emphasized the importance of conflicts in selection and the consequences for optimal virulence and coexistence of strains with different strategies [47–49]. Second, the local neighbourhood of the

initial genotype in the within-host fitness landscape has a dominant effect on emergence probabilities, because the first mutational step determines what evolutionary trajectories are accessible, reflecting empirical results in bacteriophage experiments [50]. The importance of the local neighbourhood of the initial genotype is especially pertinent when selection is strong, so that the probability of back-mutation is negligible. Both of these effects are masked when emergence dynamics are studied at a single scale.

Our analysis shows that selection acts differently at the within-host and between-host scales in the evolutionary emergence scenarios we are considering. At the within-host scale, under assumptions of SSWM, evolutionary change is driven by the relative fitness of neighbouring genotypes (compared with the current fixed genotype), and the effects of selection are manifested chiefly in the duration that a given genotype is fixed. At the between-host scale, the absolute fitness of the current genotype ($R_0^{(i)}$) is the crucial measure, as it determines whether the transmission chain continues or goes extinct. These differences stem from basic population dynamic properties of the emergence problem, which apply to many emerging infections, such as weakly transmitting zoonoses [3]. Because $R_0^{(i)} < 1$ for unadapted genotypes, the between-host process is in an invasion regime, and competition for susceptible hosts is negligible. We have assumed that all genotypes are viable at the within-host scale in order to focus our attention on population-scale emergence, and because the emerging pathogens of greatest concern are those that are already able to infect the novel host. However, it is important to recognize that the pathogen can also undergo evolutionary invasion or escape within the host, and that these can be cross-scale problems involving within-cell processes [4,37,45,51].

We have used the SSWM framework to incorporate mechanistic evolutionary principles at the within-host scale. The SSWM model has been a favoured approach to analysing evolutionary trajectories in empirically derived fitness landscapes [11,12,38]. Some aspects of the SSWM framework are very well suited to modelling pathogen emergence, such as the stochastic nature of mutation and fixation and the strong selection pressures experienced by pathogens in novel environments [11,45]. However, other aspects of the SSWM model are poor approximations to many pathogen emergence problems. For instance, the assumption of weak mutation (and consequently, a single genotype within each host) does not match the high mutation rates of RNA viruses and the tremendous diversity that can result [52,53]. Quasi-species theory may provide a more accurate portrayal of pathogens with high mutation rates, where selection acts on a 'cloud' of mutants rather than any individual genotype [54,55]. The assumption of constant population size inherent to SSWM is also a strong simplification, because pathogen loads can vary markedly throughout an infection (and between infections). An important future aim is to integrate the within-host population dynamics of the pathogen, which will influence the relative strength of selection versus drift. A particularly important application is to study evolutionary change during transmission bottlenecks, which can be extremely narrow [56,57] so drift can act strongly. André & Day [34] presented an elegant model showing how this effect can interact with within-host substitutions to influence evolutionary emergence, but further work is needed to model both evolutionary processes in the context of explicit within-host genetic diversity. These strong assumptions of the SSWM model should be borne in mind when interpreting

our results, as well as those in earlier studies applying the SSWM framework to pathogen emergence problems. Indeed, we have shown that previous models assuming equal substitution rates for all genotypes, and no back-mutations, are equivalent to the SSWM model for within-host evolution with an equal-rate fitness landscape. Therefore, the caveats outlined above apply equally to these earlier studies, with the added caution that the equal-rate landscape tends to give an upper bound for probabilities of emergence.

The SSWM framework has previously been applied to extracellular parasites and bacteria, as well as viruses [11,12,38]. Because viruses have a distinctive life-history involving reproduction within host cells, we have explored the applicability of the SSWM framework to viruses by deriving the substitution rate under SSWM assumptions from a basic model of within-host viral dynamics (see appendix A). This derivation reveals additional assumptions that are implicit in using SSWM to represent viral evolution. Namely, we assumed that all within-host fitness differences among genotypes arise from replication rates (not cell infection or within-host clearance), that viruses reproduce by budding at a constant rate and that mutations in offspring virions of a given host cell occur independently [51]. Our result also gives new perspectives into the population size component of the SSWM substitution rate. First, the derivation shows that the relevant population size is the equilibrium abundance of infected target cells, not viral particles. Second, this population size will vary as a function of within-host viral fitness, and will not remain constant for all genotypes as assumed under the classical SSWM formulation. Further investigation of how within-host dynamics lead to shifts in viral genotypes is an important avenue to developing improved cross-scale models.

Recent empirical studies have increasingly reported measures of viral fitness or tracked viral evolutionary dynamics across biological scales. These show how our work could be applied, and also guide priorities for ongoing theory development. As a first example, we consider the recent studies describing mutations that enable H5N1 influenza to transmit among mammals [7,8]. This work shows a predominantly positive correlation between fitness measures across scales, indicating that some mammal-transmissible genotypes may be favoured at the within-host scale [7]. This amplifies concerns that these genotypes could emerge in naturally circulating virus populations, though we emphasize that there is no evidence that the higher-fitness genotypes would have $R_0 > 1$ in humans, since experiments were performed in ferrets under laboratory conditions. It is also possible that other nearby genotypes (as yet uncharacterized) may have higher within-host fitness, leading to evolutionary dead ends as illustrated in figures 4 and 5. Nevertheless, our study contributes new insights to the assessment of risk from these H5N1 influenza genotypes by providing a theoretical framework in which to qualitatively assess and compare the risk of emergence of particular genotypes that could arise through mutation, given fitness measures at within-host and between-host scales. The model also presents a complementary approach to other modelling analyses that have focused on the within-host dynamics of emergence [17]. Conversely, consideration of these influenza studies reveals complexities in current data that our model does not address. Future work will need to relate temporal changes in viral titre to within-host fitness (and hence selection), and consider the

potentially crucial influence of different tissue compartments within a host [28].

Similar opportunities are evident when we consider recent studies of viruses that have emerged across species boundaries, such as canine parvovirus [13,58,59] and SARS-CoV [60,61]. Extensive laboratory work on SARS-CoV, motivated by genotypes detected in field samples, has identified adaptive mutations that improve cell receptor binding in humans; these were found in viruses transmitted between humans, but not in civet isolates [61]. Tracking the spread of such mutations in the early stages of human-to-human transmission would provide a unique opportunity to reconstruct an evolutionary emergence event, if the data can be linked [60]. Beyond-consensus sequencing studies have mapped out changes in genotype frequencies within hosts and through transmission chains, giving a window into cross-scale dynamics [15,16], and showing preliminary evidence of how viral diversity is influenced by transmission bottlenecks. Such datasets will allow us to test the validity of cross-scale evolutionary models, and refine our understanding of pertinent mechanisms. A recent study of HIV-1 highlights the unexpected insights than can arise from considering sequence data across scales. Investigating the phenomenon that HIV-1 exhibits faster substitution rates within hosts than between hosts, it concludes that the probable mechanism is that viruses closely related to the infecting strain are preferentially transmitted following storage in long-lived CD4$^+$ T cells [62]. Such a finding demonstrates the potential importance of considering specific (and sometimes idiosyncratic) biological factors when addressing questions about particular host–pathogen systems, and shows the power of cross-scale data to advance our understanding of pathogen evolution.

Accurate quantitative prediction of emergence probabilities is probably a distant goal, but mechanistic models help us better understand the relative risk of different pathogen genotypes, and assess which pathogens may be closer to emergence. As a simple example, if two viral strains are each shown to be two mutations from an emergence genotype with $R_0 > 1$, but the within-host fitness landscape is smoothly uphill for one trajectory and rugged for the other, then the strain with a smooth evolutionary path is the greater risk. Our theoretical results show us how relationships between fitnesses at multiple scales influence emergence, providing an integrative lens through which to view accumulating data on emerging pathogens. These data are arising from a broad array of approaches, from empirical mapping of fitness landscapes to deep-sequencing studies of evolutionary dynamics, and from *in vitro* and *in vivo* experiments to global field surveillance. All of these approaches can yield insight on evolutionary dynamics of pathogens at within-host or between-host scales. We applaud the recent trend towards characterizing transmissibility and inter-host evolution, since this has been a crucial data gap [5]; however, our results show that within-host fitness must be measured in parallel to arrive at a holistic picture of emergence risk. As the complexity and abundance of empirical work on emerging pathogens (or pathogens that threaten to emerge) continue to grow, the need for theoretical frameworks to analyse the resulting data and draw integrative conclusions will be even greater. The model introduced here represents a foundation for such an integrative cross-scale theory.

Appendix A. Deriving the strong selection, weak mutation (SSWM) transition probabilities from a model of viral dynamics

For a population of viruses with a single strain or genotype, let V denotes the density of the virus, U the density of uninfected target cells and I the density of infected target cells. The uninfected target cell population has a net growth rate $f(U)$. In the absence of infection, assume that dU/dt has a unique positive, stable equilibrium at U^*, i.e. $f(U^*) = 0$ and $f'(U^*) < 0$ (the virus-free equilibrium). Free viral particles encounter and infect the uninfected target cells at a rate aU. Infected cells produce new viral particles via budding at a rate β, and infected cells die at a rate δ. Viral particles are cleared from the host at a rate c. With the exception of $f(U)$, all rates are per capita. Provided the viral population is sufficiently large we can describe the viral and host cell dynamics by a mean field equation

$$\left.\begin{aligned} \frac{dU}{dt} &= f(U) - aUV, \\ \frac{dI}{dt} &= aUV - \delta I \\ \frac{dV}{dt} &= \beta I - cV - aUV. \end{aligned}\right\} \quad (A\,1)$$

and

When the virus initially infects the host at low numbers, we approximate the establishment probability within the host by assuming that $U = U^*$ is constant, and I and V are determined by a continuous time branching process with transitions

$$\left.\begin{aligned} (I, V) &\to (I + 1, V - 1) \text{ with rate } aUV, \\ (I, V) &\to (I, V + 1) \text{ with rate } \beta I \\ (I, V) &\to (I, V - 1) \text{ with rate } cV. \end{aligned}\right\} \quad (A\,2)$$

and

Then a single viral particle infects a cell with probability

$$q = \frac{aU^*}{c + aU^*} \quad (A\,3)$$

and does not infect a cell (i.e. is cleared by the host) with probability $1 - q$. In the event that a viral particle infects a cell, it gives rise to a number of offspring viruses that is geometrically distributed with mean $n = \beta/\delta$. Thus, the reproductive number for a virus, which is the expected number of viruses produced after one full cycle of cell infection, is qn. The generating function for the complete offspring distribution is given by [51]

$$g(x) = 1 - q + \frac{q}{1 + n(1 - x)}. \quad (A\,4)$$

From basic branching process theory, the ultimate probability of extinction for the viral lineage is given by the non-zero solution to $g(e) = e$. We can then solve for the

probability of establishment for a single viral particle, which is $q - 1/n$ [51].

If the viral population establishes, under appropriate assumptions, the quasi-stationary distribution is concentrated on the attractor of the system [63]. For simplicity, let us assume that this attractor is an equilibrium. We denote this equilibrium with established virus by the superscript †. The probability that a virus infects a cell is denoted by $q^\dagger$, and is found by substituting $U^\dagger$ for U^* in equation (4.3). Because the system is at equilibrium, $q^\dagger n = 1$, giving

$$\frac{aU^\dagger}{c + aU^\dagger} n = 1. \tag{A 5}$$

From this, we can solve for $U^\dagger$ and for the non-trivial equilibrium of equation (4.1)

$$\left.\begin{array}{l} U^\dagger = \dfrac{c}{a(n-1)}, \\[2mm] I^\dagger = \dfrac{f(U^\dagger)}{\delta} \\[2mm] \text{and} \qquad V^\dagger = \dfrac{f(U^\dagger)}{aU^\dagger}. \end{array}\right\} \tag{A 6}$$

Now consider two competing viral genotypes, genotype 1 and genotype 2. Assume that there is no superinfection, i.e. each cell has only one fixed viral genotype, and that infection, clearance and infected cell death rates for the two viral types are the same (i.e. $a_1 = a_2 = a$, $c_1 = c_2 = c$, and $\delta_1 = \delta_2 = \delta$). Then the viral dynamics are given by

$$\left.\begin{array}{l} \dfrac{\mathrm{d}U}{\mathrm{d}t} = f(U) - aU(V_1 + V_2), \\[2mm] \dfrac{\mathrm{d}I_i}{\mathrm{d}t} = aUV_i - \delta I_i \\[2mm] \text{and} \qquad \dfrac{\mathrm{d}V_i}{\mathrm{d}t} = \beta_i I_i - cV_i - aUV_i. \end{array}\right\} \tag{A 7}$$

Define the equilibrium abundance of uninfected cells when only viral type i is present as

$$U_i^\dagger = \frac{c}{a(n_i - 1)} \tag{A 8}$$

where $n_i = \beta_i/\delta$. With this deterministic model, linearization of the boundary equilibrium suggests that the viral population with the lower $U_i^\dagger$ invades and displaces the other viral type. This aligns with classical theory of ecological competition for a single limiting resource. Without loss of generality, assume that $U_1^\dagger > U_2^\dagger$, which occurs when $n_2 > n_1$. Note that, because $a_1 = a_2$ and $c_1 = c_2$, the cell infection probabilities $q_1 = q_2 = q$ for a given uninfected cell density U. Therefore, the n_i are proportional to the reproductive numbers for each viral type, and the condition $n_2 > n_1$ means that type 2 has higher fitness.

If viral type 1 has already established in the host and is at equilibrium abundance $U_1^\dagger$, we can approximate the invasion dynamics of the other genotype with a continuous time branching process with transitions

$$\left.\begin{array}{l} (I_2, V_2) \rightarrow (I_2 + 1, V_2 - 1) \text{ with rate } aU_1^\dagger V_2, \\ (I_2, V_2) \rightarrow (I_2, V_2 + 1) \text{ with rate } \beta_2 I_2, \\ \text{and} \quad (I_2, V_2) \rightarrow (I_2, V_2 - 1) \text{ with rate } cV_2. \end{array}\right\} \tag{A 9}$$

Following from the probability of establishment of viral type 2 being $q_2 - 1/n_2$ and using equation (4.5), the

probability of establishment of viral type 2 is:

$$\frac{aU_1^\dagger}{c + aU_1^\dagger} - \frac{1}{n_2} = \frac{1}{n_1} - \frac{1}{n_2}, \tag{A 10}$$

where $aU_1^\dagger/(c + aU_1^\dagger)$ corresponds to q for viral type 2 invading over viral type 1 at equilibrium. This probability is always positive given our assumption of $U_1^\dagger > U_2^\dagger$.

When viral type 1 is at equilibrium, mutant viruses are produced at a rate $\nu\beta_1 I_1^\dagger$ where ν is the probability that a given offspring virion will bear a mutation at the locus that converts type 1 to type 2. Altogether this gives the rate of substitution (in which viral genotype 2 displaces viral genotype 1) as

$$m_{1,2} = \nu\beta_1 I_1^\dagger \left(\frac{1}{n_1} - \frac{1}{n_2}\right). \tag{A 11}$$

We equate the establishment probability $(1/n_1 - 1/n_2)$ with the selection coefficient $s_{1,2}$ from the main text. This choice is consistent with Haldane's proof that a beneficial allele with selection coefficient s sweeps to fixation with a probability directly proportional to s [64], and is equivalent with Gillespie's definition of s in the weak selection limit used in his original SSWM derivation. We can then compare quantities from the derivation above with the classical SSWM formulation shown in the main text: $\nu\beta_1$ in our result is equivalent to the mutation rate μ, and $I_1^\dagger$ in our result is equivalent to the population size N. Thus, the expression $\nu\beta_1 I_1^\dagger(1/n_1 - 1/n_2)$ derived here corresponds closely to the $m_{i,j}$ expression in the main text (equation (2.1)).

We have shown conditions under which the viral dynamics model reduces to a form closely analogous to the SSWM formulation for substitution rates. Yet this derivation cannot be interpreted as an exact derivation of the SSWM model, as there are some subtle inconsistencies. The measure of population size, $I_1^\dagger$ corresponds to the infected host cell population, not the total number of viral particles. Significantly, the quantity $I_1^\dagger$ is a function of within-host viral fitness, leading to the potential for viral fitness to impact substitution rates through the rate at which new mutants are generated. This demographic impact of higher fitness is neglected in the classical SSWM formulation. Overall, though, the derivation shows that a close analogue of the SSWM framework can be derived from a simple model of viral dynamics, subject to similar assumptions about the strength of selection and mutation processes. Further work to explore how within-host pathogen dynamics can link to simple models of genotype substitution would be valuable, given the widespread usage of SSWM in the empirical literature.

Appendix B. Exact solution for the probability of emergence in a sequentially connected landscape

To write down an exact solution for the extinction probabilities, we consider the 'embedded' discrete-time branching process where one unit of time corresponds to an update of the continuous time branching process. The extinction probabilities for the embedded process and the original continuous time process are equivalent. The ith component of the generating map $G : [0, 1]^k \rightarrow [0, 1]^k$ for the embedded

branching process is given by a power series

$$G_i(z_1, \ldots, z_k) = \sum_{n_1, \ldots, n_k} p_i(n_1, \ldots, n_k) z_1^{n_1} z_2^{n_2} \cdots z_k^{n_k},$$

where $p_i(n_1, \ldots, n_k)$ is the probability that an individual of type i has n_1 offspring of type 1, n_2 offspring of type 2, etc.

For our model, an update of an individual host infected with genotype i leads to death with probability $d_i/(b_i + d_i + \sum_{j \in M_i} m_{i,j})$, leads to a birth with probability $b_i/(b_i + d_i + \sum_{j \in M_i} m_{i,j})$, or leads to a substitution event in which genotype i is replaced by genotype j with probability $m_{i,j}/(b_i + d_i + \sum_{j \in M_i} m_{i,j})$ for each genotype j in M_i. Hence, the ith component of the generating map is

$$G_i(z_1, \ldots, z_k) = \frac{1}{b_i + d_i + \sum_{j \in M_i} m_{i,j}} \left(b_i z_i^2 + d_i + \sum_{j \in M_i} m_{i,j} z_j \right).$$

Let e_i be the extinction probability of the process given the initial condition of one individual infected with genotype i. Provided that the branching process is supercritical (i.e. there is a positive probability of emergence), the vector of extinction probabilities $e = (e_1, \ldots, e_k)$ is the unique solution to $G(e) = e$ in $[0,1)^k$.

For the landscapes considered in the text, we can solve for e explicitly in an inductive fashion. Consider the case of a sequential landscape for which $M_1 = \{2\}$, $M_2 = \{3\}, \ldots$ $M_{k-1} = \{k\}$, and $M_k = \emptyset$. For $1 \le i < k$, let $m_i = m_{i,i+1}$. As in the text, we assume that $b_k/d_k > 1$, i.e. type k is an emergence genotype. To solve for the extinction probabilities, we proceed inductively from genotype k back to genotype 1. The

extinction probability e_k is the unique solution $e_k = z_k$ to

$$\frac{b_k z_k^2 + d_k}{b_k + d_k} = z_k \quad \text{for } z_k \in (0,1)$$

which is given by

$$e_k = \frac{d_k}{b_k}.$$

Proceeding inductively, suppose that we have solved for e_{i+1}. The ith component of the generating map G_i only depends on z_i and z_{i+1}. Slightly abusing notation, e_i is the unique solution $e_i = z_i$ to $G_i(z_i, e_{i+1}) = z_i$ with $z_i \in (0,1)$. Equivalently, e_i is the unique solution to

$$\frac{b_i z_i^2 + d_i + m_i e_{i+1}}{b_i + d_i + m_i} = z_i \quad \text{for } z_i \in (0,1)$$

which is given by

$$e_i = \frac{-\sqrt{-4 b_i m_i e_{i+1} + m_i^2 + 2(d_i + b_i) m_i + d_i^2 - 2 b_i d_i + b_i^2} + m_i + d_i + b_i}{2 b_i}.$$

A similar induction method can be used to find the exact solutions to the landscapes with two linear paths emanating from genotype i. More generally, it is possible to write down an exact solution for landscapes whenever the underlying directed graph has no directed cycles. This result will be presented in a future study.

References

1. Morens DM, Folkers GK, Fauci AS. 2004 The challenge of emerging and re-emerging infectious diseases. *Nature* **430**, 242–249. (doi:10.1038/nature02759)

2. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak D. 2008 Global trends in emerging infectious diseases. *Nature* **451**, 990–993. (doi:10.1038/nature06536)

3. Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JRC, Dobson AP, Hudson PJ, Grenfell BT. 2009 Epidemic dynamics at the human–animal interface. *Science* **326**, 1362–1367. (doi:10.1126/science.1177345)

4. Iwasa Y, Michor F, Nowak MA. 2004 Evolutionary dynamics of invasion and escape. *J. Theor. Biol.* **226**, 205–214. (doi:10.1016/j.jtbi.2003.08.014)

5. Pepin KM, Lass S, Pulliam JRC, Read AF, Lloyd-Smith JO. 2010 Identifying genetic markers of adaptation for surveillance of viral host jumps. *Nat. Rev. Microbiol.* **8**, 802–813. (doi:10.1038/nrmicro2440)

6. Parrish CR, Holmes EC, Morens DM, Park E-C, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P. 2008 Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol. Mol. Biol. Rev.* **72**, 457–470. (doi:10.1128/MMBR.00004-08)

7. Imai M et al. 2012 Experimental adaptation of an influenza H5 HA confers respiratory droplet

transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* **486**, 420–428. (doi:10.1038/nature10831)

8. Herfst S et al. 2012 Airborne transmission of influenza A/H5N1 virus between ferrets. *Science* **336**, 1534–1541. (doi:10.1126/science.1213362)

9. Fauci AS, Collins FS. 2012 Benefits and risks of influenza research: lessons learned. *Science* **336**, 1522–1523. (doi:10.1126/science.1224305)

10. Poelwijk FJ, Kivient DJ, Weinreich DM, Tans SJ. 2007 Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386. (doi:10.1038/nature05451)

11. Weinreich DM, Delaney NF, DePristo MA, Hartl DL 2006 Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114. (doi:10.1126/science.1123539)

12. Lozovsky ER, Chookajorn T, Brown KM, Imwong M, Shaw PJ, Kamchonwongpaisan S, Neafsey DE, Weinreich DM, Hartl DL 2009 Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proc. Natl Acad. Sci. USA* **106**, 12 024–12 030. (doi:10.1073/pnas.0905922106)

13. Stucker KM, Pagan I, Cifuente JO, Kaelber JT, Lillie TD, Hafenstein S, Holmes EC, Parrish CR. 2012 The role of evolutionary intermediates in the host adaptation of canine parvovirus. *J. Virol.* **86**, 1514–1521. (doi:10.1128/JVI.06222-11)

14. Hoelzer K, Murcia PR, Baillie GJ, Wood JL, Metzger SM, Osterrieder N, Dubovi EJ, Holmes EC, Parrish CR. 2010 Intrahost evolutionary dynamics of canine influenza virus in naive and partially immune dogs. *J. Virol.* **84**, 5329–5335. (doi:10.1128/JVI.02469-09)

15. Murcia PR et al. 2010 Intra- and interhost evolutionary dynamics of equine influenza virus. *J. Virol.* **84**, 6943–6954. (doi:10.1128/JVI.00112-10)

16. Murcia PR et al. 2012 Evolution of an Eurasian avian-like influenza virus in naive and vaccinated pigs. *PLoS Pathog.* **8**, e1002730. (doi:10.1371/journal.ppat.1002730)

17. Russell CA et al. 2012 The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. *Science* **336**, 1541–1547. (doi:10.1126/science.1222526)

18. The Chinese-SARS Molecular Epidemiology Consortium. 2004 Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303**, 1666–1669. (doi:10.1126/science.1092002)

19. Reperant LA, Kuiken T, Grenfell BT, Osterhaus ADME, Dobson AP. 2012 Linking influenza virus tissue tropism to population-level reproductive fitness. *PLoS ONE* **7**, e43115. (doi:10.1371/journal.pone.0043115)

20. Williams GC. 1992 *Natural selection: domains, levels and challenges*, vol. 4. Oxford series in ecology and evolution. New York, NY: Oxford University Press.

21. Keller L. 1999 *Levels of selection in evolution*. Princeton, NJ: Princeton University Press.

22. Okasha S. 2006 *Evolution and the levels of selection*. Oxford, UK: Clarendon Press.

23. Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP 2007. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc. Natl Acad. Sci. USA* **104**, 17 441–17 446. (doi:10.1073/pnas.0708559104)

24. Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F, Meehan MO, Lutalo T, Gray RH. 2000 Viral load and heterosexual transmission of human immunodeficiency virus type 1. *N. Engl. J. Med.* **342**, 921–929. (doi:10.1056/NEJM200003303421303)

25. Antia R, Levin BR, May RM. 1994 Within-host population dynamics and the evolution and maintenance of microparasite virulence. *Am. Nat.* **144**, 457–472. (doi:10.1086/285686)

26. Ganusov VV, Antia R. 2003 Trade-offs and the evolution of virulence of microparasites: do details matter? *Theor. Popul. Biol.* **64**, 211–220. (doi:10.1016/S0040-5809(03)00063-7)

27. Porco TC, Lloyd-Smith JO, Gross KL, Galvani AP. 2005. The effect of treatment on pathogen virulence. *J. Theor. Biol.* **233**, 91–102. (doi:10.1016/j.jtbi.2004.09.009)

28. Reperant LA, Kuiken T, Grenfell BT, Osterhaus ADME, Dobson AP. 2012 Linking influenza virus tissue tropism to population-level reproductive fitness. *PLoS ONE* **7**, e43115. (doi:10.1371/journal.pone.0043115)

29. Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, Grenfell BT. 2006 Host species barrier to influenza virus infections. *Science* **312**, 394–397. (doi:10.1126/science.1122818)

30. Antia R, Regoes RR, Koella JC, Bergstrom CT. 2003 The role of evolution in the emergence of infectious diseases. *Nature* **426**, 658–661. (doi:10.1038/nature02104)

31. Yates A, Antia R, Regoes RR. 2006 How do pathogen evolution and host heterogeneity interact in disease emergence? *Proc. R. Soc. B* **273**, 3075–3083. (doi:10.1098/rspb.2006.3681)

32. Arinaminpathy N, McLean AR. 2009 Evolution and emergence of novel human infections. *Proc. R. Soc. B* **276**, 3937–3943. (doi:10.1098/rspb.2009.1059)

33. Alexander HK, Day T. 2010 Risk factors for the evolutionary emergence of pathogens. *J. R. Soc. Interface* **7**, 1455–1474. (doi:10.1098/rsif.2010.0123)

34. André J-B, Day T. 2005 The effect of disease life history on the evolutionary emergence of novel pathogens. *Proc. R. Soc. B* **272**, 1949–1956. (doi:10.1098/rspn.2005.3170)

35. Amaku M, Burattini M, Coutinho F, Massad E. 2010 Modeling the dynamics of viral evolution considering competition within individual hosts and at population level: the effects of treatment. *Bull. Math. Biol.* **72**, 1294–1314. (doi:10.1007/s11538-009-9495-6)

36. Iwasa Y, Michor F, Nowak MA. 2005 Virus evolution within patients increases pathogenicity. *J. Theor. Biol.* **232**, 17–26. (doi:10.1016/j.jtbi.2004.07.016)

37. Mullins JI, Rolland M, Allen TM. 2008 Viral evolution and escape during primary human immunodeficiency virus-1 infection: implications for vaccine design. *Curr. Opin. HIV AIDS* **3**, 60–66. (doi:10.1097/COH.0b013e3282f233d9)

38. da Silva J, Coetzer M, Nedellec R, Pastore C, Mosier DE. 2010 Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. *Genetics* **185**, 293–303. (doi:10.1534/genetics.109.112458)

39. Gillespie JH. 1984 Molecular evolution over the mutational landscape. *Evolution* **38**, 1116–1129.

40. Harris TE. 1963 *The theory of branching processes*. Berlin, Germany: Springer.

41. Jensen JLWV. 1906 Sur les fonctions convexes et les inégalités entre les valeurs. *Acta Mathematica* **30**, 175–193.

42. Elena SF, Sanjuán R. 2007 Virus evolution: insights from an experimental approach. *Annu. Rev. Ecol. Evol. Syst.* **38**, 27–52. (doi:10.1146/annurev.ecolsys.38.091206.095637)

43. Fernàndez G, Clotet B, Martínez MA. 2006 Fitness landscape of human immunodeficiency virus type 1 protease quasispecies. *J. Virol.* **81**, 2485–2496. (doi:10.1128/JVI.01594-06)

44. Burch CL, Chao L. 1999 Evolution by small steps and rugged landscapes in the RNA virus ϕ 6. *Genetics* **151**, 921–927.

45. Park AW, Daly JM, Lewis NS, Smith DJ, Wood JL, Grenfell BT. 2009 Quantifying the impact of immune escape on transmission dynamics of influenza. *Science* **326**, 726–728. (doi:10.1126/science.1175980)

46. Schreiber SJ, Lloyd-Smith JO. 2009 Invasion dynamics in spatially heterogeneous environments. *Am. Nat.* **174**, 490–505. (doi:10.1086/605405)

47. Gilchrist MA, Coombs D 2006 Evolution of virulence: interdependence, constraints, and selection using nested models. *Theor. Popul. Biol.* **69**, 145–153. (doi:10.1016/j.tpb.2005.07.002)

48. Coombs D, Gilchrist MA, Ball CL. 2007 Evaluating the importance of within- and between-host selection pressures on the evolution of chronic pathogens. *Theor. Popul. Biol.* **72**, 576–591. (doi:10.1016/j.tpb.2007.08.005)

49. Mideo N, Alizon S, Day T. 2008 Linking within- and between-host dynamics in the evolutionary epidemiology of infectious diseases. *Trends Ecol. Evol.* **23**, 511–517. (doi:10.1016/j.tree.2008.05.009)

50. Burch CL, Chao L. 2000 Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature* **406**, 625–628. (doi:10.1038/35020564)

51. Loverdo C, Park M, Schreiber SJ, Lloyd-Smith JO. 2012 Influence of viral replication mechanisms on within-host evolutionary dynamics. *Evolution* **66**, 3462–3471. (doi:10.1111/j.1558-5646.2012.01687.x)

52. Wright CF, Morelli MJ, Thbaud G, Knowles NJ, Herzyk P, Paton DJ, Haydon DT, King DP. 2010 Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J. Virol.* **85**, 2266–2275. (doi:10.1128/JVI.01396-10)

53. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010 Viral mutation rates. *J. Virol.* **84**, 9733–9748. (doi:10.1128/JVI.00694-10)

54. Eigen M. 1993 Viral quasispecies. *Sci. Am.* **269**, 42–49. (doi:10.1038/scientificamerican0793-42)

55. Más A, López-Galíndez C, Cacho I, Gómez J, Martínez MA. 2010 Unfinished stories on viral quasispecies and Darwinian views of evolution. *J. Mol. Biol.* **397**, 865–877. (doi:10.1016/j.jmb.2010.02.005)

56. Keele BF et al. 2008 Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl Acad. Sci. USA* **105**, 7552–7557. (doi:10.1073/pnas.0802203105)

57. Wang GP, Sherrill-Mix SA, Chang K-M, Quince C, Bushman FD. 2010 Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. *J. Virol.* **84**, 6218–6228. (doi:10.1128/JVI.02271-09)

58. Kaelber JT, Demogines A, Harbison CE, Allison AB, Goodman LB, Ortega AN, Sawyer SL, Parrish CR. 2012 Evolutionary reconstructions of the transferrin receptor of Caniforms supports canine parvovirus being a re-emerged and not a novel pathogen in dogs. *PLoS Pathog.* **8**, e1002666. (doi:10.1371/journal.ppat.1002666)

59. Allison AB et al. 2012 Role of multiple hosts in the cross-species transmission and emergence of a pandemic parvovirus. *J. Virol.* **86**, 865–872. (doi:10.1128/JVI.06187-11)

60. Zhao G-P. 2007 SARS molecular epidemiology: a Chinese fairy tale of controlling an emerging zoonotic disease in the genomics era. *Phil. Trans. R. Soc. B* **352**, 1063–1081. (doi:10.1098/rstb.2007.2034)

61. Sheahan T, Rockx B, Donaldson E, Corti D, Baric R. 2008 Pathways of cross-species transmission of synthetically reconstructed zoonotic severe acute respiratory syndrome coronavirus. *J. Virol.* **82**, 8721–8732. (doi:10.1128/JVI.00818-08)

62. Lythgoe KA, Fraser C 2012 New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proc. R. Soc. B* **279**, 3367–3375. (doi:10.1098/rspb.2012.0595)

63. Faure M, Schreiber SJ. 2012 Quasi-stationary distributions for randomly perturbed dynamical systems. (http://arxiv.org/abs/1101.3420)

64. Haldane JBS. 1927 The mathematical theory of natural and artificial selection. V. Selection and mutation. *Proc. Camb. Philos. Soc.* **23**, 838–844. (doi:10.1017/S0305004100015644)