The dynamics of viral adaptation : theoretical and experimental approaches

Martin Guillemet

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Sciences de l'Évolution et de la Biodiversité

École doctorale GAIA

Unité de recherche UMR 5175 CEFE

The dynamics of viral adaptation: Theoretical and experimental approaches

Présentée par Martin GUILLEMET

Le 14 septembre 2023

Sous la direction de Sylvain Gandon

Devant le jury composé de

Ophélie RONCE, Directrice de recherche, CNRS Montpellier	Présidente du jury
Hildegard Uecker, Chargée de Recherche, Max Planck Institute Plön	Rapporteure
Roland Regös, Maître de Conférences, ETH Zurich	Rapporteur
Olivier Tenaillon, Directeur de Recherche, INSERM Paris	Examinateur
Sylvain Gandon, Directeur de Recherche, CNRS Montpellier	Directeur de Thèse

UNIVERSITÉ DE MONTPELLIER

Acknowledgments / Remerciements

I would like to thank the reviewers Hildergard Uecker and Roland Regös for accepting to review my thesis and to participate in my jury.

Je remercie également Ophélie Ronce et Olivier Tenaillon d'avoir accepté de participer à ce jury de thèse.

Merci à Stéphanie Bedhomme, Nicolas Bierne et Olivier Tenaillon (encore une fois!) d'avoir constitué mon comité de suivi de thèse, merci pour ce regard apporté, ces remarques et tous les encouragements pour la suite.

Merci énormément à Guillaume Martin pour la collaboration ayant donné lieu au premier chapitre ainsi qu'à un appendice de cette thèse. Même si le prix à payer est que Sylvain m'appelle parfois Guillaume.

Un grand merci à Sonja Lehtinen et l'équipe à l'ETH de Zürich, pour m'avoir proposé de venir en post-doc et donc permis de finir ce manuscrit sereinement sans paniquer pour la suite!

Evidemment un grand merci à Sylvain, d'avoir toujours été bienveillant et disponible, pour tous les "c'est pas mal" pour me dire gentiment que certaines idées étaient pourries mais aussi pour l'enthousiasme à chaque fois que j'avais de nouveaux résultats. Tu as mis la barre très haute en terme de supervision, et je souhaiterai à tous les thésards de t'avoir comme directeur.

Merci à tous les permanents de l'équipe EEE: Rémi, Sébastien et Thierry, pour l'atmosphère de l'équipe et les repas du midi, pour m'avoir supporté pendant que je parlais du Tour de France, de Roland Garros ou l'autre dernière compétition sportive..

Merci à Amélie et Marina de passage en postdoc qui auront égayé l'équipe, et participé à dissiper la concentration des thésards avec les longues conversations sur le pas de notre porte..

Pour les autres doctorants de l'équipe, merci à Alicia et Valentin pour ces après-midi à regarder les matchs en discutant de la nouvelle série ou du dernier pokemon,

vive le bureau 204!

Evidemment un grand merci à Wakinyan, le réel pilier du bureau 204. A toutes ces après-midi où on aura commencé à regarder un truc de stats en finissant sur de sombres pages Wikipedia à se poser des questions sur la jeunesse d'un politique, ou comment présenter un homard, ou à contempler la pensée de Bergson...

Merci à Erwan d'être venu en stage, d'avoir réchauffé le bureau avec les grosses simulations pendant la canicule mais surtout pour les conseils de vélo!

Merci bien sûr à Fanny, la boss du CEFE mais aussi et avant tout des escargots, toujours là pour discuter quand ça va pas, ou pour mettre du Jul quand ça va bien, merci pour tout.

Merci à Lilian pour avoir fait vivre le CEFE et les apéros, même si tu m'as battu de 2 secondes au trail j'oublierai jamais qui est le vrai maître des apéros et de l'imitation de Céline Dion.

Merci à Adrien de l'ISEM évidemment, pour toutes les discussions à Montmaur, le p'tit gin, le p'tit rhum, mais aussi pour les conseils sur la thèse et les entretiens de postdoc, sans oublier les photos d'étoiles au lac de Pises même si j'ai failli finir ecrasé par un sanglier.

Merci à tous les non permanents du CEFE, stagiaires, thésards et postdoc, et particulièrement aux comités d'animation successifs, pour l'ambiance et évidemment pour avoir enflammé les soirées au bois de Montmaur (sauf bien sûr pendant les couvre-feux non surtout pas).

Merci à tous les étudiants que j'ai eu la chance d'avoir en cours à l'Université. J'ai vraiment adoré toutes ces heures. Parfois quand je m'ennuie, je me dis que je donnerai bien un petit cours de modèle linéaire, juste comme ça.

Un grand merci à tous les profs que j'ai pu avoir jusqu'ici, en particulier aux profs de prépa à Malherbe et surtout à Olivier Perraud sans qui je ne serai surement jamais arrivé jusque là.

Un énorme merci à toute l'équipe de Lyon, pour tous ces week-ends, soirées, que

ce soit à Laval'dance ou bien Morvan 1, Morvan 2 ou encore Morvan 3 et bien sûr pour les Brouettes, le palet et la grenouille.

Merci à l'équipe du lycée à Dumont, en particuler à Sarah et PE, d'avoir été et d'être toujours là.

Merci à toute ma famille de m'avoir supporté et encouragé jusque là. Un immense merci à mon frère Adrien et à Rhiannon qui auront corrigé in extremis une partie de ce manuscrit dans le sprint final, quand j'étais moi-même en compote.

Merci à toute la famille de Manon, en particulier aux pépé mémés pour toutes ces invitations et l'accueil à Clarensac.

Et principalement merci à Manon, d'être à mes côtés depuis la prépa et de m'avoir soutenu tout ce temps, et pour tous les moments à venir.

Et finalement, merci à la Truite pour tous les miaou.

Contents

This document presents the results of my PhD at the CEFE (Centre d'Ecologie Fonctionnelle et Evolutive) in Montpellier under the direction of Sylvain Gandon. This PhD started on September 1st 2020 and the present document was finalized in July of 2023. The grant funding this PhD is from the Ministère de la Recherche et de l'Enseignement Supérieur and was awarded by the Ecole Normale Supérieure de Lyon.

Introduction

In this thesis we conducted several projects under the overarching theme of the dynamics of viral adaptation. Viruses can mutate and adapt in very short time scales which has direct impacts for example on the way viral diseases like the Human Immunodeficiency Virus (HIV) are treated, or vaccination strategies are deployed. The pandemic of Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) harshly reminded us of the threat posed by emerging or re-emerging pathogen. In this context, it is then crucial that we understand the mechanisms driving viral adaptation to design effective prophylactic, therapeutic or non pharmaceutical interventions which would limit unwanted consequences of viral evolution.

Evolutionary dynamics is moulded by the selective pressures imposed by the environment. For pathogens, it is the availability of susceptible hosts that constitutes the environment. For this reason, we begin this introduction with a section on epidemiological dynamics before addressing more precisely the question of viral adaptation in three distinct sections, corresponding to the three chapters of the thesis. First, we present the adaptation of viruses through changes in life-history traits such as transmission rate or virulence, in a homogeneous host population. Second, we introduce resistance in the host population, and discuss the effect on pathogen emergence and subsequent evolution. Third, we discuss the coevolutionary dynamics of viral population with their hosts.

In this thesis we use both theoretical and experimental approaches. With theory we try to disentangle the effects of different evolutionary mechanisms and provide qualitative and quantitative predictions for the outcome of viral adaptation. We use experiments to validate some of these predictions and to try to uncover new biological processes.

Epidemiology

SARS-CoV-2: a zoonotic disease

The SARS-CoV-2 pandemic that has swept through human populations since 2019 illustrates how the interplay between epidemiological and evolutionary dynamics can affect the viral spread. Before the onset of the pandemic, we can speculate that this virus was circulating in several other mammal species and "jumped" from its original animal reservoir host to humans in a process known as a zoonotic spillover. This process has been estimated to be at the origin of 60 to 75% of human emerging infectious diseases (Taylor, Latham, and Woolhouse, 2001; Woolhouse and Gowtage-Sequeria, 2005). The spillover of a pathogen from another species to humans requires two main elements. First the contact between the pathogen and a human host. Human population growth, alongside massive deforestation and consumption of animal derived products, has led to an ever-increasing rate of contact between humans, animals, and their pathogens (Ellwanger and Chies, 2021). Pathogens have spilled over to humans from a variety of wild animal hosts such as rats, bats and camels, but also from domesticated animals like poultry and pigs. Secondly, a pathogen in contact with a human host must be able to successfully infect this host for any kind of outbreak to happen. This requires specific molecular features, and in particular for intracellular pathogens like viruses, the recognition of a human receptor to allow entry into a human cell before any replication can happen. For example, the spike protein of SARS-CoV-2 binds the ACE2 receptor which can be found on the surface of human lung cells, while the crucial amino acids for this binding have not been found in one of the closest known relative of SARS-CoV-2, the virus RaTG13 infecting the horseshoe bat *Rholophus affinis* (Andersen et al., 2020).

Epidemiological dynamics

Viruses can jump between different host species through zoonotic spillovers, but not all spillovers will result in pandemics of the scale of Covid-19. Many "dead end" spillovers (which are by nature difficult to detect) may result in a pathogen jump to a new species in which the pathogen cannot replicate significantly. To understand how epidemic outbreaks can result in such different scenarios, we can use the following toy SIR (Susceptible-Infected-Recovered) model in continuous time:

$$\begin{aligned}\dot{S}(t) &= b - S(t)(d + \beta I(t)) \\ \dot{I}(t) &= \beta S(t)I(t) - (d + \alpha + \gamma)I(t) \\ \dot{R}(t) &= \gamma I(t) - dR(t)\end{aligned} \tag{1}$$

In this model, susceptible hosts S enter at a constant rate b and die at a per capita rate d. Infected cells I infect susceptible cells upon contact with a transmission rate β. Infected hosts suffer an additional mortality of α (which we will call the virulence), and recover at rate γ, becoming resistant to further infection. In this model the equilibrium density of susceptible cells in the absence of disease is $S(0) = b/d$. One can make the approximation that when a pathogen is introduced, for example after a zoonotic spillover, the density of susceptible cells is $S(0)$. The epidemic will grow if the time derivative of the density of infected cells $\dot{I} > 0$. This results in the following condition:

$$R_0 = \frac{\beta}{d + \alpha + \gamma} S(0) > 1 \tag{2}$$

where R_0 is called the basic reproduction number. This number can be interpreted as the number of secondary infections that will be caused, on average, by a single infected host in a population of otherwise susceptible hosts (Anderson and May, 1992). Indeed, if an infected host infects, on average, less than one other host, the number of infected hosts will decrease over time and the epidemic will become extinct. On the other hand, if this number is greater than 1, the number of infected hosts will increase and the infection will spread in the host population. Examining the expression for R_0 gives intuition as to why certain viruses will cause massive pandemics and others only small and contained outbreaks. To successfully spread to other hosts, the pathogen must easily be transmitted from infected to susceptible hosts with rate β. The duration of infectiousness also needs to be sufficiently long. Indeed if infected hosts die or recover quickly, there may not be enough time to cause secondary infections. This period of time, which in the simple SIR model is the lifespan of an infected host, is found in the expression for R_0 and is equal to $\frac{1}{d+\alpha+\gamma}$. Finally the basic reproduction number depends on the density of susceptible hosts, which can be thought of as the resource available to the pathogen. If the pathogen successfully emerges, then the system will eventually converge towards the following endemic equilibrium:

$$\begin{aligned}S^\star &= \frac{d + \alpha + \gamma}{\beta} \\ I^\star &= \frac{b\beta - d(d + \alpha + \gamma)}{\beta(d + \alpha + \gamma)}\end{aligned} \tag{3}$$

Disease emergence and stochasticity

The question of emergence of a pathogen is inherently linked to small populations of pathogens, down to just a single infected host. Indeed with small numbers, the approximation that a pathogen will escape initial extinction and create an epidemic if $R_0 > 1$ does not hold. In fact, what this deterministic result provides is a lower bound on the basic reproduction number for the emergence of an epidemic. Let us consider a single infected host with $R_0 = 2$ in a population of otherwise susceptible hosts. On average, such a host would infect two other hosts. Yet, the introduction of this pathogen is not certain to lead to a major epidemic. Pathogen emergence in the simple context of our SIR model can be modeled with a one dimensional birth-death branching process, and the probability of emergence (Diekmann, Heesterbeek, and Britton, 2013) after the introduction of I_0 infected host with $R_0 > 1$ is:

$$p_e^n = 1 - \left(\frac{1}{R_0}\right)^{I_0} \tag{4}$$

Thus the higher the R_0, the likelier the pathogen is to escape initial extinction. The number of introduced infected hosts, which we can relate to the number of events of zoonotic spillover, also increases this probability.

Adaptation

The rest of this introduction is divided into three sections, corresponding to the three chapters of this thesis. The objectives of each chapter are presented at the end of the corresponding section.

1. Pathogen adaptation in an homogeneous host population

Fitness and the dynamics of a mutant pathogen

If a newly introduced pathogen escapes initial extinction, it will eventually reach a high enough density so that it can be well described by deterministic demographic trajectories. In this framework, the value of R_0 is not well suited as it describes the number of secondary infections produced by an infected host, but with no indication on the speed of such infections. To study demographic trajectories in continuous time, the more appropriate measure is the malthusian fitness or growth rate. In the epidemiological model (1) the fitness of infected hosts is:

$$r(t) = \beta S(t) - (d + \alpha + \gamma) \tag{5}$$

When the malthusian fitness is positive, the number of infected hosts increases. Malthusian fitness provides critical information to study adaptation. Let us now consider the following system where there are two strains of a pathogen (I_w for *wild-type* and I_m for *mutant*) with different values of life-history traits, and thus fitness:

$$\begin{aligned}
\dot{S}(t) &= b - S(t)(d + \beta_w I_w(t) + \beta_m I_m(t)) \\
\dot{I}_w(t) &= \beta_w S(t) I_w(t) - (d + \alpha_w + \gamma_w) I_w(t) \\
\dot{I}_m(t) &= \beta_m S(t) I_m(t) - (d + \alpha_m + \gamma_m) I_m(t) \\
\dot{R}(t) &= \gamma_w I_w(t) + \gamma_m I_m(t) - dR(t)
\end{aligned} \tag{6}$$

Although not immediately apparent, one can check that it is their difference in fitness which dictates which of these strains outgrows the other. If we study $p_m(t) = I_m(t)/(I_w(t) + I_m(t))$ we get:

$$\dot{p}_m(t) = \underbrace{p_m(t)\big(1 - p_m(t)\big)}_{\text{variance}} \underbrace{\big(r_m(t) - r_w(t)\big)}_{\text{selection coefficient}} \tag{7}$$

We see that a mutant will invade a population if its fitness is higher than the fitness of the resident wild-type strain, and the speed is controlled by the fitness difference as well as the frequency of the mutant at that time.

Long term adaptation

We have seen with equation (7) the dynamics of the frequency of a mutant pathogen. Now we can ask the question: can a certain mutant m invade a population where there is a resident strain w? And more generally, can we find a strain that cannot be invaded by any other? Addressing this question requires the introduction of the term Evolutionary Stable Strategy (ESS): a strategy which, in a given environment, cannot be displaced by a new emerging strategy. A naive method to find the ESS is developed in the Teaching Appendix C, where we build a Pairwise Invasibility Plot, which explores in a given space of life-history traits whether a mutant strain will outcompete a resident strain. To find the ESS analytically in our simple epidemiological model we can use the maximisation of R_0 criteria (Anderson and May, 1982): the strain with the highest possible R_0 will not be invaded by any other strain, and thus be at the ESS. If a pathogen strain spreads more between hosts, then it will be favoured by natural selection. The speed of pathogen spread can also be quantified by the malthusian fitness. Consider an endemic equilibrium for a resident pathogen strain w where the density respectively of infected and susceptible hosts are $I_w^\star$ and $S_w^\star$. This strain will be at the ESS if at its endemic equilibrium, any other mutant strain m has a negative fitness upon introduction. This invasion fitness for pathogen m is:

$$r_m^\star = \beta_m S_w^\star - (d + \alpha_m + \gamma_m) \tag{8}$$

Indeed, if this invasion fitness is negative, the mutant strain m will not be able to grow and thus will not outcompete the resident ESS strain. From the above expression, it is easy to check that consistent with the idea of R_0 maximisation, the invasion fitness being negative is equivalent to the condition that the R_0 of strain m is lower than the R_0 of the ESS strain.

R_0 maximisation has been tightly linked to the development of the adaptive theory of virulence, which probably stemmed from Theobald Smith's "Law of declining virulence" (Méthot, 2012). For a long time, virulence has been seen as a transient side effect of disease emergence, based on the idea that a virulent pathogen will greatly deplete its resource: susceptible hosts.

Referred to as the "conventional wisdom" by (May and Anderson, 1983), this

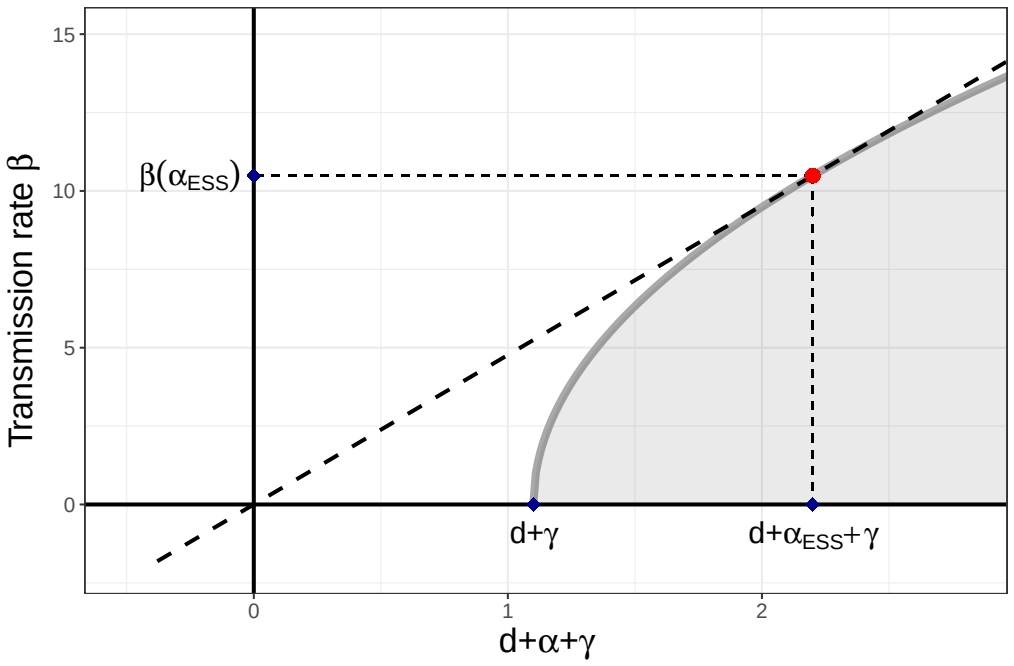


Figure 1: The trade-off curve and finding the ESS. Maximal transmission rate β is dependent on the virulence and we show it as a function of $d + \alpha + \gamma$ which corresponds to the rate at which the infection ends. All phenotypes in the shaded area under the curve are accessible, but through natural selection they will be cleared and strains with higher transmission rate and lower virulence will be selected, until they reach the boundary of the trade-off curve. The ESS can be determined using the tangent to the trade-off curve that passes through the origin. Indeed, R_0 is maximised when the ratio $\beta(\alpha)/(d + \alpha + \gamma)$, which corresponds to the slope of the dashed line, is maximised.

classical hypothesis that pathogens would tend to avirulence has since been challenged, perhaps most notably with the trade-off hypothesis (Anderson and May, 1982) reviewed in (Alizon et al., 2009). This hypothesis states that virulence is an unavoidable consequence of parasite transmission as the latter must come at a cost. For instance a given value of transmission rate requires a minimum value of virulence based on underlying mechanisms. One such mechanism would be that both transmission rate and virulence are dependent on the within-host reproduction of the pathogen. A better reproduction would lead, for a virus, to an increased amount of viral particles which would lead to an increased viral shedding and a better spread between hosts. However an enhanced reproduction within the host would certainly impose a cost on this host, linked to the virulence.

An interesting metaphor for the transmission-virulence trade-off is proposed by (Bonneaud and Longdon, 2020). To travel between islands, humans would need to

build boats. To build such boats they would need to cut down trees to get wood, and in doing so they would deplete the resources from their present island. Thus they would need to make the most boats while cutting down the minimum number of trees to be able to visit more new islands. In the same way, pathogens need to optimize the use of their hosts resources to spread. A "trade-off curve" is often used to model this hypothesis where the transmission rate is now $\beta(\alpha)$. It corresponds to a parametric curve linking directly virulence with the maximum transmission rate it allows. If this curve is convex (positive second derivative) or linear, then no ESS can be achieved and a more transmissible and virulent strain on the curve will always be favoured. A more biologically meaningful curve is a concave (negative second derivative) one, where there is a saturation in the increase of transmission rate with virulence. In this case, a single ESS exists where R_0 is maximised and we show in Figure 1 how the corresponding computation can be visualised.

Yet maximising R_0 to obtain the ESS is not a generality and there are cases when this rule does not apply. In fact, evolution will only maximise R_0 in very simple environments and does not apply when more complex environmental feedback loops are introduced (e.g. density-dependent mortality, host heterogeneity, spatial structure etc...)(Lion and Metz, 2018). Besides in simple environments such as our SIR model, the ESS will indeed be a strategy where R_0 is maximised, yet this only applies at equilibrium. We show in the Teaching Appendix that transiently, other strains can outcompete the ESS, particularly after the introduction of the pathogen. In this case, the density of susceptible hosts is higher than at equilibrium, and a strain with a higher transmission rate than the ESS would actually be favoured.

Fisher's fundamental theorem of adaptation

The interplay between fitness and natural selection was described by Fisher in 1930 with the "Fundamental theorem of natural selection". This theorem states that "The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time" (Fisher, 1999). Indeed we can transform the previous expression (7) to obtain the dynamics of mean fitness:

$$\dot{\overline{r}}(t) = \underbrace{\sum_i p_i(t)\big(r_i - \overline{r}\big)^2}_{\text{Variance in fitness}} \tag{9}$$

where $p_i(t)$ and $r_i(t)$ are respectively the frequency and fitness of strain i. Similarly to the dynamics of the mutant frequency, the mean fitness of the pathogen population will increase with a speed dictated by the heterogeneity in genotypes,

and the difference in fitness of these genotypes. In this epidemiological context, it means that the pathogen population will be enriched in strains that are more transmissible, yet less virulent and with a lower rate of recovery. In the above expression we dropped the dependence on time, thus assuming a constant population of susceptible hosts $S(t)$. This is necessary to squarely fit to this formulation of Fisher's fundamental theorem. Therefore it should be noted that this formulation can be misleading and actually describes the rate of increase in fitness from natural selection only.

Dynamics of life-history traits

Besides mean fitness it is possible to derive directly the dynamics of the mean trait. If only one trait X is under selection and the environment is constant, the dynamics of the mean trait in case of perfect heritability can be expressed as:

$$\dot{\overline{X}}(t) = \mathrm{Cov}\big(X(t), r_X(t)\big) \tag{10}$$

where $cov(X, r_X(t))$ denotes the covariance in the population between trait X and the fitness. For example, if transmission rate β is the only varying trait in the population, we can write the dynamics of the mean transmission rate as:

$$\begin{aligned} \dot{\overline{\beta}}(t) &= \mathrm{Cov}\big(\beta(t), r_\beta\big) \\ &= \mathrm{Cov}\big(\beta(t), \beta(t)S - (d+\alpha)\big) \\ &= S\,\mathrm{Var}\big(\beta(t)\big) \end{aligned} \tag{11}$$

with $\mathrm{Var}(\beta)$ the variance in transmission rate in the population. The mean transmission rate is thus governed by the trait variance, and as the density of hosts S linearly scales the effect of transmission rate on fitness, it also linearly scales the intensity of selection on this trait.

Using this approach, it is also possible to study jointly the dynamics of two traits under selection in the population. Contrary to the equilibrium approach of R_0 maximisation, we can write the dynamics of both mean transmission rate and virulence as:

$$\begin{pmatrix} \dot{\overline{\beta}}(t) \\ \dot{\overline{\alpha}}(t) \end{pmatrix} = \underbrace{\begin{pmatrix} \mathrm{Var}\big(\beta(t)\big) & \mathrm{Cov}\big(\beta(t), \alpha(t)\big) \\ \mathrm{Cov}\big(\beta(t), \alpha(t)\big) & \mathrm{Var}\big(\alpha(t)\big) \end{pmatrix}}_{\mathbf{G}} \cdot \begin{pmatrix} S(t) \\ -1 \end{pmatrix} \tag{12}$$

where $\mathbf{G}$ is the variance-covariance matrix of the traits.

Impact of environment change on fitness

Adaptation and the change in mean fitness is not only due to natural selection, but also to change in environment (Price, 1972). In environment e, the change in mean fitness $\overline{r}$ between two points in time is:

$$\Delta\overline{r} = \overline{r}'|e' - \overline{r}|e \tag{13}$$

where the prime denotes the fitness and environment at the next time point. To better understand this formula it is useful to rewrite it as (Frank and Slatkin, 1992):

$$\begin{aligned}\Delta\overline{r} &= (\overline{r}'|e - \overline{r}|e) + (\overline{r}'|e' - \overline{r}'|e) \\ &= \Delta r_{ns} + \Delta r_{ec}\end{aligned} \tag{14}$$

where Δr_{ns} is the effect of natural selection only, while Δr_{ec} describes the change in mean fitness due to environmental change, these are all the biotic or abiotic factors that could affect fitness. Thus Fisher's theorem explains Δr_{ns} only, which explains its apparent lack of generality. Environmental change include many factors, with potentially frequency-dependent processes, that can be dependent or not on the adaptation of the organism of interest.

We can derive the change in mean fitness according to natural selection and environmental change in our simple epidemiological model. First we must write the mean fitness using (5) as a function of the mean life-history traits it includes, which also depend on time:

$$\overline{r}(t) = \overline{\beta}(t)S(t) - (d + \overline{\alpha}(t) + \overline{\gamma}(t)) \tag{15}$$

From which we can write:

$$\dot{\overline{r}}(t) = \underbrace{\overline{\beta}(t)\dot{S}(t)}_{\Delta r_{ec}} + \underbrace{\dot{\overline{\beta}}(t)S(t) - \dot{\overline{\alpha}}(t) - \dot{\overline{\gamma}}(t)}_{\Delta r_{ns}} \tag{16}$$

We now have an additional term influencing the dynamics of variance in our simple SIR model. While Δr_{ns} is always positive, i.e. natural selection always increases mean fitness, the effect of environmental change Δr_{ec} depends on the sign of $\dot{S}(t)$. If the susceptible population decreases ($\dot{S}(t) < 0$) then the quality of the environment worsens ($\Delta r_{ec} < 0$) leading to a decline in the infected hosts fitness. This is what happens after the emergence of a pathogen in a naive population, the density of susceptible hosts will decrease as the epidemic spreads and thus slows the growth rate of the infected population. At the endemic equilibrium (3) the density of infected hosts is stable and so the fitness of infected host must necessarily be zero. Natural selection can also induce the degradation of the environment. For instance

if a mutant pathogen arises with a higher transmission rate $\beta_m > \beta_w$ and invades the population, then the susceptible host population will decrease towards the new equilibrium value (3) corresponding to this new value of β, leading to a negative Δr_{ec}.

Mutation

We have treated in a simple case of SIR model the dynamics of the frequency of two pathogen strains and the possibility of an ESS, a strategy with which a strain cannot be invaded by any other strain. Yet the elephant in the room which we have not treated is how different strains of different genotypes (and thus potentially phenotypes) arise. Although one could invoke migration as a mechanism that introduces variability, it is through mutation that new genotypes are generated.

Viruses can exhibit very high rates of mutation which allow them to generate massive bursts of diversity over short times. It was for example estimated that HIV-1 could reach rates of 4.1×10^{-3} substitutions per nucleotide per infection (Cuevas et al., 2015). For comparison, this rate is estimated to 2.5×10^{-8} in humans (Nachman and Crowell, 2000). Of course, mutation rate varies wildly between different viruses and some exhibit rate close to that of humans. Several elements have been shown to impact mutation rate (reviewed in (Sanjuán and Domingo-Calap, 2016): RNA viruses mutate faster than DNA viruses, single stranded viruses mutate faster than double stranded ones etc.. One interesting characteristic that correlates with mutation rate is genome size: viruses with larger genomes tend to have lower mutation rates. For example, coronaviruses are the RNA viruses with the largest genomes, and they are the only RNA viruses to have evolved a proof-reading capacity for replication, thus limiting the mutation rate (Smith, Sexton, and Denison, 2014).

With such a high mutation rate, the term viral quasispecies has been used to describe the mutant distribution of viruses (also called mutant cloud or mutant spectrum) upon replication (Andino and Domingo, 2015; Domingo and Perales, 2019; Domingo, Sheldon, and Perales, 2012; Lauring and Andino, 2010). Quasispecies theory was first developed by Eigen to study the dynamics of a population of primitive replicons (Eigen, 1993). Within this framework, mutation rate is so high that it is unlikely that upon replication descendants will have the same genomic sequence as their immediate parent. At the population level this translates to a distribution of sequences rather than the traditional view of a consensus sequence and some low frequency variants. The cloud of mutants can spread in the genomic sequence space

and is only maintained, when mutation is not too strong, by negative selection which clears unfit mutants. This theoretical framework has been shown to be consistent with classic population genetics results (Wilke, 2005) and is supported by many experimental observations for example in HIV (Del Portillo et al., 2011; Jung et al., 2002) or Hepatitis C (Sobesky et al., 2007).

Mutation rate is only one part of the picture. Another key to describe the mutation process is to infer the effect of mutations on a phenotype, or directly onto fitness. Genome expression is so complex that we cannot guess the exact effect of all mutations. Protein 3D models can be used to examine the effect of an amino acid change on the binding of a protein to a ligand and predict a stronger or reduced interaction, but genome expression is so complex that a change in nucleotide at some location could for example impact the expression of the next gene. We therefore need experimental measurements to address the question of Distribution of Fitness Effects (DFE) of mutations, and most results come from what are known as mutation accumulation experiments.

Mutation accumulation (MA) experiments were first proposed by Muller to study the rate of mildly deleterious mutations in *Drosophila*. Indeed mutations are hard to study and (besides using X-rays to increase mutation rate) he proposed to allow the accumulation of many mutations in many different *Drosophila* lines (Muller, 1927). The goal of a MA experiment is to minimise as much as possible the effects of natural selection, so that a population evolves only through new mutations and genetic drift. This can be achieved through recurrent bottlenecks of a population, down to the passage of just one individual. Picking just one individual in a population and propagating it arbitrarily fixates any mutation that this individual has quasi-independently of its fitness effect. Except for highly deleterious or lethal ones, mutations will accumulate in the lines at the speed with which they would occur naturally in a genome. After each passage and in each line, a trait (usually growth rate) can be tested and using parameter estimation techniques, one can access estimates (or bounds) for the value of mutation rate, for the mean effect of mutation and for the variance in mutational effects. The vast majority of studies have found that the mean effect of mutation on fitness was deleterious (Halligan and Keightley, 2009). With this information on the mean effect of mutations on fitness $\overline{s}$ and the mutation rate U, it is possible to express the direct effect of mutation on the dynamics of mean fitness :

$$\Delta r_m = U\,\overline{s} \tag{17}$$

With a negative value of the mean effect of mutation of fitness $\overline{s}$, mutation is thus a force which directly reduces fitness over time.

Rather than average values of fitness effect, it is also possible to directly infer DFE of mutations using experimental testing of a collection of artificially induced mutations. This method requires considerable work to assess the effect on fitness of individual mutation. Sanjuán et al. used such a method to test the fitness effects of 91 single mutant clones of the vesicular stomatitis virus (VSV) generated through site-direct mutagenesis (Sanjuán, Moya, and Elena, 2004b). The fitness of these mutants is shown in Figure 2. These results show around 40% of lethal mutations and a mean effect of non lethal mutations of -13.2% on fitness. The authors note how striking it is that they found 2 out of 48 random mutations that were beneficial. It is generally accepted that mutations of beneficial effects are $\sim$ 1000-fold less common than neutral or deleterious ones (Miralles et al., 1999; Orr, 2003). In many models of adaptation a DFE with only deleterious mutations is often used, such as with a gamma distribution that has been shown to fit well observed data (Burch, Guyader, et al., 2007). Sanjuán et al. explain this over-representation of beneficial mutations with the fact that the ancestral virus they use is a chimera from two different VSV genomes and so there are "many different possible ways to optimize such genome".

This highlights a caveat of studies estimating mutational effects: they describe the effects of mutations from one ancestral genotype, in one given environment. In particular, the time during which a population has evolved in one environment can completely change the expected DFE. It is expected that as one organism adapts to one environment, the DFE of mutations will shift towards an increased proportion of deleterious mutations, and beneficial ones will get scarcer. This is supported by fitness trajectories which show saturating behaviour with time, the most famous example being the Lenski *E. coli* experiment (Lenski et al., 1991; Wiser, Ribeck, and Lenski, 2013) which has now reached more than 75,000 generations. This saturation can be explained by the hypothesis that beneficial mutations are close to exponentially distributed: there should be few mutations of large effects expected to arise early, followed by more numerous mutations of smaller effects (Orr, 2005). If mutations of large beneficial effects are possible, they are expected to fixate first through natural selection.

A possible mechanism behind this exponential distribution of beneficial effects is epistasis, which is completely circumvented when studying the effects of individual

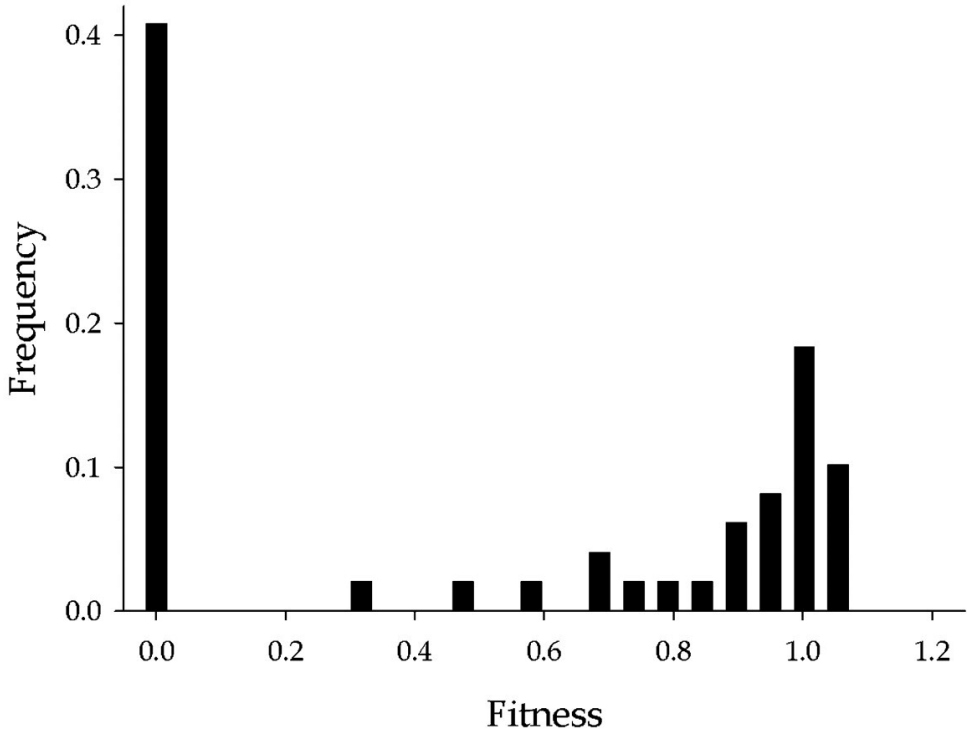


Figure 2: Frequency of fitness values associated with single-nucleotide substitutions measured for random mutations in vesicular stomatitis virus clones (VSV), from (Sanjuán, Moya, and Elena, 2004b)

mutations independently. It was shown with the VSV system described earlier that the observed effect of mutant carrying pairs of mutations was different than the expected multiplicative model would predict (Sanjuán, Moya, and Elena, 2004a). In particular they showed that the effect of pairs of beneficial mutations tended to be smaller than expected, showcasing antagonistic epistasis. Strikingly, a symmetric phenomenon was described with virus $\phi 6$: mutants with deleterious mutations were less sensitive to further deleterious mutations, highlighting positive epistasis (Burch and Chao, 2004).

Fisher's Geometric Model of adaptation: a fitness landscape

We saw that the effect of mutations on fitness is a very complex issue, and to understand it completely for substitution (one sub-type of mutation), for a single organism with genome size N_g, in a single environment, we would need to measure the fitness of all 4^{N_g} possible genotypes. To understand adaptation, let alone predict it, we must then use models of adaptation that include as many key features as possible while limiting the number of required parameters. One such model was proposed by Fisher in 1930 (Fisher, 1999), and which is now called "Fisher's Ge-

ometric Model" (FGM). In this model, one organism is defined by n independent phenotypic continuous traits. These phenotypic traits are abstract and do not represent traits like viral adsorption, burst size etc. Instead they represent underlying characteristics, and affect fitness through a transmission function. This model features stabilising selection, and thus each trait has an optimal value, meaning there is a single optimal phenotype which maximises fitness. This optimum is at the origin, and fitness decreases as the distance to the optimum in this n-space increases. Although it is possible to introduce anisotropy in selection, we will consider here an isotropic version of the model, meaning that selection is equivalent for all phenotypic traits. Resembling classic quantitative genetics selection functions (Lande and Arnold, 1983), a Gaussian transmission function has been used to link phenotype and Darwinian fitness:

$$W_{\mathbf{x}} = e^{-\frac{\|\mathbf{x}\|^2}{2}} \tag{18}$$

This transmission function is represented with a two-dimensional phenotypic space in Figure 3. Introducing mutations in this framework can be done by sampling a mutation vector from a multivariate distribution (e.g. multivariate normal distribution) which can be added to the phenotypic vector. The effect of such a mutation will depend on the original phenotype and its distance to the optimum. As represented in panels (b) and (c), we see that FGM contains features discussed in the previous section. (i) The DFE contains less beneficial mutations as a phenotype gets closer to the optimum (i.e. is better adapted to its environment) and converges to a DFE with only deleterious mutations for phenotypes which have reached the optimum. (ii) There is antagonistic epistasis between beneficial mutations, meaning that the combined fitness effect of two beneficial mutations will be lower than expected if fitness effects were multiplicative.

Overall, this model has received a strong support from experiments regarding its different features: distribution of fitness effects, fraction of beneficial mutations, epistasis etc. (reviewed in (Tenaillon, 2014)). Yet these features are static and relate to the mutation process in the FGM. Yet this model could also be used to derive the dynamics of fitness and provide testable predictions. Martin&Roques proposed a framework to study such fitness dynamics in a FGM setting (Martin and Roques, 2016). It is inspired by an approach originally from Bürger (Burger, 1991) where the dynamics of the distribution of traits under mutation and selection is followed through the cumulants of this distribution (cumulants are quantities related to moments, e.g. the mean, the variance etc.). Martin & Roques propose to

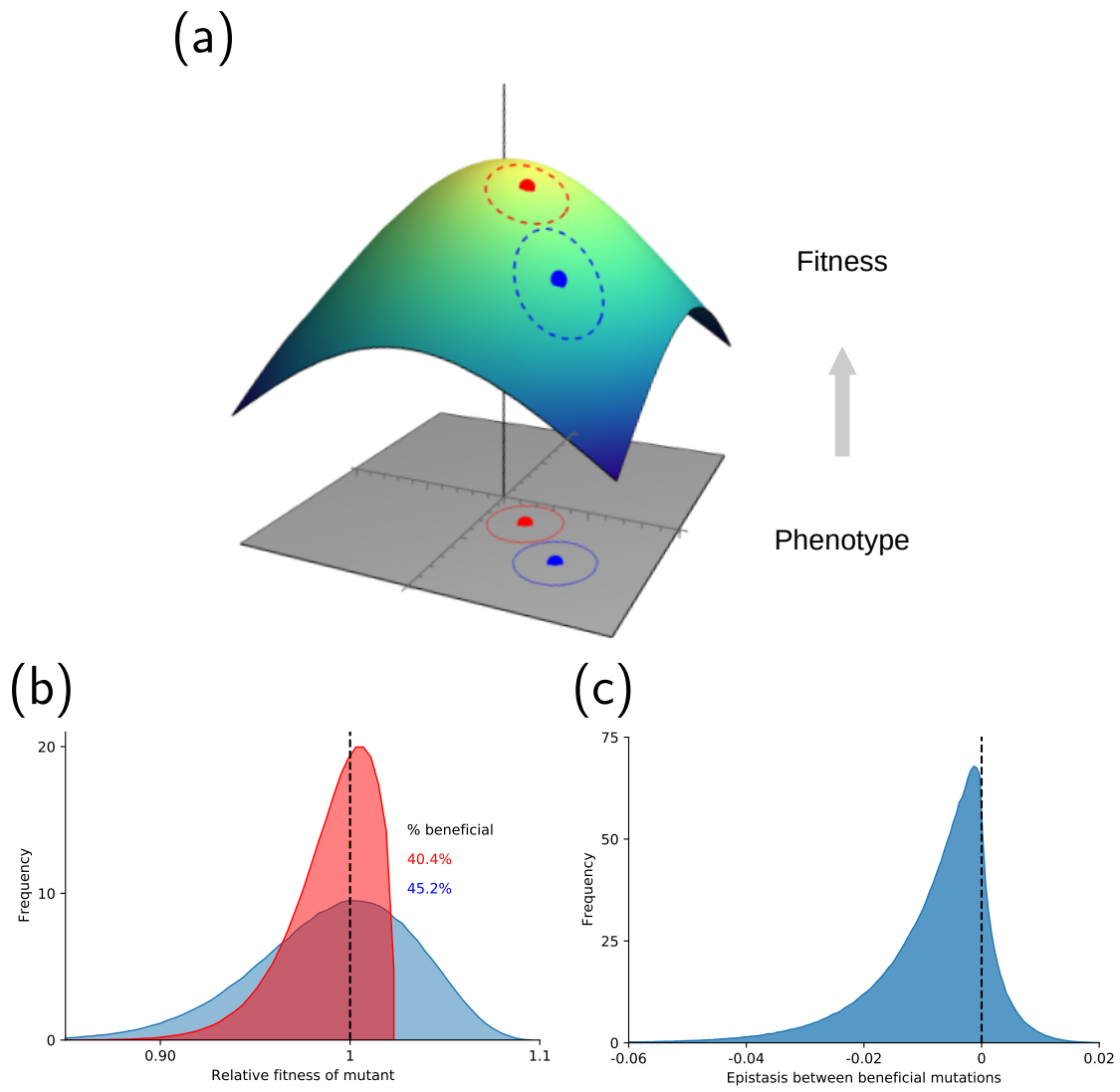


Figure 3: Schematic of Fisher's Geometric Model with phenotypic complexity n=2. (a) The Gaussian fitness transmission function is shown above the 2D phenotypic space. 2 phenotypes are shown in red and blue with a circle at a fixed phenotypic distance. (b) Distribution of Fitness Effects of Gaussian mutations from the red and blue phenotype are shown, with the associated fraction of beneficial mutations. (c) Distribution of epistasis between two beneficial mutations picked from the pool of Gaussian mutations. Epistasis is computed with : $e = Log(\frac{r_{12}}{r_0}) - Log(\frac{r_1}{r_0}) - Log(\frac{r_2}{r_0})$ where r_0 is the ancestral fitness, r_1 and r_2 are the fitness values of 2 beneficial single mutants and r_{12} is the fitness of the corresponding double mutant.

follow directly the distribution of fitness in a population by using a single equation describing the dynamics of the Cumulant Generating Functions (CGF) of the fitness distribution. This approach alleviates the need to follow individually the dynamics of each cumulant. With this approach they provide trajectories for the distribution of fitness under various assumptions on the mutation/selection regime which could prove to be testable predictions.

Objectives of Chapter 1

In this introduction we have seen the dynamics of mean fitness could be described as the sum of three forces such that:

$$\Delta r = \Delta r_{ns} + \Delta r_{ec} + \Delta r_m \tag{19}$$

where Δr_{ns} is the effect of natural selection, Δr_{ec} is the effect of environmental change (biotic or abiotic) and Δr_m is the direct effect of mutations. The direct effect of mutations Δr_m is generally negative, and many mutations come with a cost so high that the resulting fitness of the virus harbouring them is negative (Figure 2). Viruses (particularly RNA viruses) also exhibit very high mutation rates. From these observations, the idea of driving viruses to extinction through an increase in mutation rate has been proposed (J. J. Bull, Sanjuan, and Wilke, 2007; Lynch, Bürger, et al., 1993) in a process known as "lethal mutagenesis". This process has been well studied in a stochastic setting (Lynch, Bürger, et al., 1993; Lynch and Gabriel, 1990; Matuszewski et al., 2017) and particularly focusing on the role of Muller's ratchet (Muller, 1964), which predicts that the fittest genotypes will keep disappearing through drift, thus bringing down the fitness of the population and potentially to extinction. However lethal mutagenesis can also drive a population to extinction in a deterministic manner (Martin and Gandon, 2010). We have seen in this introduction that mutations have a direct negative effect on fitness, yet they generate genetic and phenotypic variance, which is the fuel of adaptation through natural selection Δr_{ns}. Thus, in case of failure to achieve lethal mutagenesis, an increased mutation rate could also lead to fitter pathogens. To get a complete picture of the process, it then seems necessary to incorporate both deleterious and beneficial mutations in models. Besides, lethal mutagenesis is mostly studied independently of the epidemiological dynamics. Yet we can expect that the dynamics of the host population will have an effect on the fitness of the pathogen population through the environmental change term Δr_{ec}.

In the first chapter, we model the joint epidemiological and evolutionary trajectories taking place within an infected host. To model mutational effects, we adapt

Fisher's geometric model to a pathogen version: instead of being translated to fitness, phenotypes are translated to a value of transmission rate, which can then be interpreted as a fitness by taking into account the density of susceptible cells (5). To better fit observations such as the DFE presented in Figure 2, we also add a category of mutations independent from phenotype: lethal mutations which lead to a non-viable pathogen which cannot transmit. This framework has been used by Martin et al. (Martin and Gandon, 2010) but their analyses were limited to the equilibrium of the system. Here we extend this work by modeling the dynamics of a distribution of transmission rates based on two methods: (i) using the dynamics of the cumulants of the distribution of transmission rate (Burger, 1991; Bürger, 2000) and (ii) using a Partial Derivative Equation (PDE) on the Cumulant Generating Function (CGF) of the distribution of transmission rates (Martin and Roques, 2016).

We also extend this framework in Appendix A to model the evolution of both transmission rate and virulence. To this end we use FGM with two distinct optima (Martin and Lenormand, 2015): one for transmission rate and one for virulence, which produces an emerging trade-off between the two life-history traits. We have seen earlier how to recover the Evolutionary Stable Strategy according to the trade-off function linking the two life-history traits. However we can expect that viruses which are not well adapted, for example from recent zoonosis, will not be on this trade-off curve but rather in the area under it (the grey area in Figure 1). With our modeling scheme, we can describe jointly the epidemiological dynamics and evolutionary dynamics of transmission rate and virulence.

2. Adaptation to host resistance

Introduction: resistance

In the previous section we assumed that the susceptible population was made up of only one type of hosts, and so a virus spreading in this population would adapt to an homogeneous environment. We discussed the effect of environment change through its effect on fitness Δr_{ec} with the example of the density of susceptible hosts decreasing. Another scenario of environment change impacting pathogen fitness is the introduction in the population of hosts resistant to infection. There are many mechanisms of resistance to viral infection in humans such as adaptive immunity, genetic factors, or vaccines.

Vaccines have been widely used as a large scale protection measure since the $20th$ century and have been incredibly efficient to reduce the number of deaths associated with viral infections. As a striking example, the vaccine against smallpox allowed for the complete eradication of the disease as declared by the World Health Organization in 1980, a disease which had previously caused hundreds of millions of deaths. Many other diseases have seen their numbers of annual cases greatly reduced.

The success of vaccination campaigns is not solely due to the individual effect of a vaccine. On top of this individual protection, there is an emerging property of resistance to infection at the population level. When a significant portion of the population is vaccinated, the circulation of the disease in the population is impaired which greatly decreases transmission and new cases, which can potentially end the epidemic.

To explore this, we can adapt the most simple epidemiological model (1) to include another compartment, that of primed hosts :

$$\begin{aligned} \dot{S}_N(t) &= b\,(1-p) - S_N(t)(d+\beta I(t)) \\ \dot{S}_P(t) &= b\,p - S_P(t)d \\ \dot{I}(t) &= \beta S_N(t) I(t) - (d+\alpha+\gamma) I(t) \\ \dot{R}(t) &= \gamma I(t) - dR(t) \end{aligned} \tag{20}$$

where S_N and S_P respectively refer to naive and primed hosts, and p is the fraction of incoming hosts who are primed. In this model we consider that recovering from the disease provides full resistance. In this case the equilibrium value for the density of infected hosts is:

$$I^\star = \frac{b\beta(1-p) - d(d+\alpha+\gamma)}{\beta(d+\alpha+\gamma)} \tag{21}$$

which decreases with the fraction of incoming primed hosts. The equilibrium given above is only valid if the basic reproduction number of the pathogen is superior to one, and its value is dependent on p following:

$$R_0 = \frac{\beta}{d+\alpha+\gamma}\frac{b(1-p)}{d} \tag{22}$$

This value is inferior to that of the same pathogen in a population of only naive hosts. This yields a threshold value on the fraction of primed hosts which drives the pathogen to extinction, which can be written as:

$$p > 1 - \frac{1}{R_0} \tag{23}$$

This corresponds to the notion of "herd immunity" which arises when a significant portion of the population is protected against a disease. In this case the spread of the pathogen is so impeded that the population as a whole is protected, including naive hosts which would otherwise be vulnerable.

Pathogen emergence in a partially resistant population

The probability of pathogen emergence in homogeneous and susceptible host population can be described with equation (4). In this equation, one can plug the result for the value of R_0 when there is a fraction p of resistant hosts in the population computed in (22) to obtain the probability of emergence in such a population, always in the case $R_0 > 1$. However, it is possible that an initially maladapted pathogen with $R_0 < 1$ generates adaptive mutation(s) before extinction. Such new mutant could then have a value of $R_0^\star > 1$ and escape extinction. We call this type of scenario an "evolutionary emergence". If we consider that mutation granting escape to host resistance, leading to a new reproductive number $R_0^\star$, takes place with a probability u for every new infection, we can approximate the probability of evolutionary emergence with (André and Day, 2005; Gandon, Hochberg, et al., 2013):

$$P_{EE} = \frac{R_0}{1 - R_0}\, u \left(1 - \frac{1}{R_0^\star}\right) \tag{24}$$

where $1 - 1/R_0^\star$ is the probability of emergence of a single mutant as shown in (4). The first quotient is the expected size for the epidemic caused by one initial pathogen with $R_0 < 1$, which is multiplied by the mutation rate u to yield the expected number of mutants produced by such an epidemic. Finally, this expected number is multiplied by the probability of emergence of the mutant from (4).

In a more complex setting, Chabas et al. (Chabas et al., 2018) used this approach to compute the probability of pathogen emergence when introduced in a population divided between susceptible and resistant hosts. In particular they study the effect of heterogeneity in the resistant fraction of the population, which can be divided in n different compartments. They show that evolutionary emergence and the spread of escape mutations in the pathogen population is more likely to occur when the host population contains an intermediate proportion of resistant hosts.

The effect of diversity of host resistance

Beside the frequency of resistant hosts in the population, there is also a strong impact of the diversity of the resistance on pathogen emergence and subsequent evolution. Indeed it has been observed that populations with poor genetic diversity were more prone to bigger epidemics (O'Brien and Evermann, 1988). This effect is supported by modeling (King and C. Lively, 2012; C. M. Lively, 2010) but the relation between epidemics sizes and host diversity is still unclear.

In their paper Chabas et al. also study the effect of the diversity of resistance in the host population on pathogen emergence. They show that the probability of pathogen emergence rapidly decreases with the diversity of resistance in the host population, because the selection coefficient associated with the escape to each individual spacer is reduced (Chabas et al., 2018). They also use an experimental system with CRISPR-resistant bacteria and bacteriophages to test this hypothesis, and manage to recover a significant effect of bacterial diversity on phage emergence.

Dynamics of a mutant escaping host resistance

Yet all vaccines are not infallible nor is natural adaptive immunity, and the SARS-CoV-2 pandemic is a prime example. With a lack of antiviral treatment, vaccines were the cornerstone of public health strategy. As of 2023, more than 13 billion doses of SARS-CoV-2 vaccine have now been administered around the world, with more than 70% of the world population now having received at least one dose. WHO reports that there has been more than 750 million confirmed cases. Such a predominant vaccine coverage and natural immunity can impose a very different environment for a spreading virus, greatly impacting the evolutionary pressures. Under these conditions, if a mutant virus arises which is even slightly able to infect vaccinated hosts, it will be greatly favoured by natural selection, which likely explains the global dominance of the Omicron Variant of Concern as of 2023. Several studies have demonstrated that this variant, first detected in late 2021, could evade neutralization when confronted to serum from vaccinated individuals. These observations of immune escape by Omicron can thus explain its quick global spread even in countries with high vaccination coverage (estimated R_0 of 3 to 5 in the UK in 2022) (Willett et al., 2022).

To study this behaviour, we can adapt the SIR to model five compartments: naive hosts S_N, primed hosts S_P, hosts infected by a wild-type pathogen I_w, host

infected by a mutant pathogen I_m and finally recovered hosts R. We consider that the wild-type pathogen is not able to infect primed hosts contrary to the mutant which can infect all types of susceptible hosts at a cost c in transmission rate:

$$\begin{aligned}
\dot{S}_N(t) &= b\,(1-p) - S_N(t)(\beta I_w(t) + \beta(1-c)I_m(t) + d) \\
\dot{S}_P(t) &= b\,p - S_P(t)(\beta(1-c)I_m(t) + d) \\
\dot{I}_w(t) &= \beta S_N(t) I_w(t) - (d+\alpha+\gamma) I_w(t) \\
\dot{I}_m(t) &= \beta(1-c) S_N(t) I_m(t) + \beta(1-c) S_P(t) I_m(t) - (d+\alpha+\gamma) I_m(t) \\
\dot{R}(t) &= \gamma(I_m(t) + I_w(t)) - dR(t)
\end{aligned} \tag{25}$$

If at least one of the pathogen is present and can emerge, that is with $R_0 > 1$, then this system will converge to an endemic equilibrium, which composition will depend on the cost parameter c of the mutant. We can study the dynamics of the frequency of the mutant by adapting equation (7) to this system:

$$\begin{aligned}
\dot{p}_m(t) &= p_m(t)(r_m(t) - \overline{r}(t)) \\
&= p_m(t)\big(1 - p_m(t)\big)\Big(\beta(1-c)\big(S_N(t) + S_P(t)\big) - \beta S_N(t)\Big)
\end{aligned} \tag{26}$$

Crucially, the sign of the difference in fitness is dependent on the densities of both primed and naive hosts at time t. If there is no cost $c = 0$ for the mutant associated with immune escape, the mutant pathogen will have strictly more resources available than the wild-type with no downside. The fitness of the hosts infected by the mutant strain will then always be higher than that of hosts infected by the wild-type strain. In such a case, the wild-type pathogen will be driven to extinction, and only the mutant pathogen will remain. In contrast, if there is a too high cost to immune escape, it could outweigh the benefit to the mutant of having higher proportion of hosts available compared to the wild-type.

There may be intermediate situations where the cost of immune escape and the densities of both types of hosts lead to the fitness of both pathogens being equal. Thus an equilibrium can be reached where both mutant and wild-type pathogens coexist. This particular case means that neither strategy is evolutionary stable, which can arise depending on two symmetric conditions: (i) the invasion fitness of the mutant in a wild-type-only endemic equilibrium $r(m, w) > 0$ and (ii) the invasion fitness of the wild-type in a mutant-only endemic equilibrium $r(w, m) > 0$. These conditions imply that both strains are able to invade without completely displacing each other, and so they can coexist.

CRISPR, an adaptive immunity system for bacteria

The question of evolutionary emergence of a pathogen that can escape pre-existing resistance in the host population is central to understand the robustness of vaccination strategies, and is also a major concern in agriculture. Genes of resistance to pathogens are often introduced and selected for in cultivated crops, and there are documented cases of emerging pathogens escaping this resistance (McDonald and Linde, 2002). Yet in both those systems – animal vaccination or resistant crops – experimental data that could complement models is hard to obtain. The problem of pathogen evolutionary emergence is intrinsically stochastic, and estimating probabilities of evolutionary emergence requires a great number of parallel replicates which is not easily feasible for the aforementioned systems.

A promising experimental pathosystem for the study of evolutionary epidemiology is that of bacteria and bacteriophages, which are bacteria-infecting viruses. A variety of bacteria and their specific bacteriophages are easily cultivated in laboratory conditions. It is thus feasible to generate great amounts of data using classic microbiology techniques to study demographic or evolutionary aspects for many replicate epidemics. Through the measurement of fitness trajectories, adaptive landscapes, mutation effects etc., phages have been used to study the joint impact of natural selection and mutation on adaptation to an homogeneous and susceptible host population (J. J. Bull, Heineman, and Wilke, 2011; J. Bull, Badgett, Rokyta, et al., 2003; J. Bull, Badgett, and Wichman, 2000; Burch and Chao, 1999; Burch, Guyader, et al., 2007). Yet another great possibility offered by phage-bacteria systems is the study of the evolution of viral escape to host resistance.

Bacteriophage abundance is estimated to be around 10-fold greater than that of their bacterial hosts (Suttle, 2005). It is then not surprising that bacteria have evolved many defense systems against phage infections. Some classic types of these systems have been described for some time now (Labrie, Samson, and Moineau, 2010), including mechanisms such as adsorption-blocking, superinfection exclusion, restriction-modification enzymes, abortive infection, Clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (cas) genes etc. (Hampton, Watson, and Fineran, 2020) Very recently, there has been a burst of discovery of new defense systems using their genomic signature and particularly their clustering in "defense islands" (Bernheim and Sorek, 2020; Doron et al., 2018; Makarova, Wolf, Snir, et al., 2011).

In this thesis, we will focus particularly on the CRISPR-Cas system, which provides a fantastic system to study the evolution of viral escape. CRISPR-Cas systems are broadly distributed across the genomes of about 42% of bacteria and 85% of archaea (Makarova, Wolf, Iranzo, et al., 2020). This defense system functions by storing short fragments of phage genetic sequences in the bacterial genome, called spacers, which are separated by unique repeat sequences. These spacers act as an immune memory repertoire, allowing for specific resistance to phages harboring the corresponding sequences in their genome, which are called protospacers. When a bacterium is exposed to a foreign DNA molecule, the Cas complex, guided by spacers, can recognize matching protospacers. The Cas complex then cleaves the foreign DNA at that site, thus stopping the infection.

This defense system is based on the recognition of complementary base pairs between bacterial spacers and phage protospacers which leads to high specificity, but also to a certain weakness in term of robustness. Only a simple mutation like a nucleotide substitution in the phage protospacer can prevent this recognition and thus lead to phage escape to the resistance provided by a specific spacer (Deveau et al., 2008). The specificity of resistance and escape being directly encoded in the host and pathogen genome means that sequencing on one hand the bacterial CRISPR locus, and on the other hand the entire phage genome (which is usually small), completely describes their respective resistance and infectivity phenotypes.

Objectives of Chapter 2

In the second chapter we compare different strategies of deployment of resistance to limit the emergence of epidemics. A population of hosts can be constituted of susceptible and potentially different types of resistant individuals. It is also possible for hosts to be multi-resistant, meaning that pathogens would need several escape mutations to infect them. We study the efficacy of three different structures of host population resistance:
- A Mixing strategy where the resistant population is made up of two single resistant hosts that we call A and B.
- A Pyramiding strategy where the resistant host population is homogeneous and double resistant AB
- A Combining strategy where the resistant host population is made of half single resistant hosts (A or B) and half double resistant hosts AB.

To study the efficacy of each of these resistance deployment strategies, we focus on a specific quantity: the probability of pathogen emergence, where pathogens escape initial extinction. We use theoretical modeling to compute this probability according to the strategy and the initial number of pathogens inoculated. We also test these analytical predictions experimentally using the CRISPR-resistant bacteria and bacteriophages system. In this system, it is possible to reproduce the strategies mentioned above and in particular the double-resistant hosts. Besides, it is possible with this system to carry out many replicate infections, which is required to estimate probabilities of pathogen emergence. Finally, we can measure the infectivity phenotypes of the bacteriophages at the end of the experiment to find whether they can infect either A, B or double resistant AB hosts.

In another experiment we use this CRISPR-resistant bacteria and phages experimental system to study the dynamics of the frequencies of different escape mutants after the initial emergence (Appendix B). The goal of this work is to monitor these frequencies through time in different treatments which we designed to assess the contribution of two parameters:
- Escape mutation rate: if a certain escape mutation happens at a higher rate, we could expect that it could arise sooner in the experiment, and so the frequency of this mutation would increase earlier.
- Selection coefficient: if a certain type of resistant is more frequent than others, an escape mutation allowing the phage to infect this particular host would be more strongly selected, and so increase in frequency faster.

To investigate the effect of escape mutation rate, we use two groups of host strains which differ in their corresponding escape mutation rate in the phage, and we manipulate the selection coefficients of the escape mutations by using different frequencies of these two groups of hosts. However, these frequencies of hosts are subject to change during the experiment, and there could also be acquisitions of additional spacers of resistance by the hosts. To circumvent these limitations, we limit host evolution by transferring each days only the phage population onto a fresh mix of bacteria with constant frequencies of hosts, thus allowing host evolution and epidemiological feedback only within each day.

3. Coevolution between viruses and their hosts

We have mentioned in the precedent section the abundance of phages in natural environments, as well as the abundance of bacterial defense systems against phages.

These two facts are not independent and highlight that in addition to the selective pressure imposed by bacteria on phage population, there is a reciprocal pressure imposed by phages on their bacterial hosts. Different phage lifestyles can lead to a variety of ecological interactions. Virulent phages – which can only transmit horizontally and require to kill their hosts to release more viral particles and spread – can impose great mortality on bacterial populations (Clokie et al., 2011). In contrast, temperate phages can integrate in their hosts genome in a process called lysogeny. Through this process, phages can reproduce vertically alongside their hosts, meaning that their fitness is closely related to the fitness of their host. Phages have thus been shown to impact positively their host's fitness by harbouring beneficial functional genes such as antibiotics resistance genes, representing *de facto* a mechanism of bacterial horizontal gene transfer (Colavecchio et al., 2017). Temperate phages have a more complex lifestyle than virulent phages, and we will focus on the latter in this thesis. The importance of ecological interactions between host and pathogen leads the way to the process of coevolution, which can be defined as the process of reciprocal adaptation and counter-adaptation between ecologically interacting species (Janzen et al., 1980).

Red queen dynamics

Coevolution is often described through the metaphor of the "Red Queen" (Van Valen, 1973). Different species experience evolutionary pressures resulting from a network of interactions, which changes over time as the different partners evolve. Generally, the adaptation of one species will degrade the effective environment experienced by other species (and thus their mean fitness through the term Δr_{ec} in equation (19)), which calls for their reciprocal adaptation in order to escape extinction. Indeed, 'it takes all the running you can do, to keep in the same place'. Inside this Red Queen framework, two main types of coevolutionary dynamics have been described (Woolhouse, Webster, et al., 2002) depending on the underlying genetic determinism of the interactions.

In Arm's Race Dynamics (ARD) adaptation is unidirectional, and the two interacting species will repetitively fix new adaptive mutations (Fig 4.a). This type of dynamics is thus driven in both host and pathogen by the selection coefficient associated with the different possible mutations, as well as the mutation rate. In this case, polymorphism is mostly transient and is only observed during the selective sweep of a beneficial mutation. In the long term, it can be imagined that all possi-

ble beneficial mutations will have been acquired, potentially leading to a stop in the coevolutionary dynamics characterized by a stable coexistence, or the extinction of one of the interacting species (or both: if the hosts go extinct, pathogens will shortly follow).

However for Fluctuating Selection Dynamics (FSD), polymorphism is maintained but the frequency of particular alleles may fluctuate through time. In particular, the selection coefficient of a given allele will often be considered dependent on the frequency of a corresponding allele in the population of the interacting species. This type of interaction can lead to cyclical dynamics for both host and pathogen allele frequencies, with a small delay or lag in the dynamics of pathogen allele frequency as the pathogen is tracking and chasing the host phenotype (Fig 4.b). This type of dynamics can be thought of as an example of Negative Frequency Dependent Selection (NFDS): if a certain host allele is over-represented in the population then the corresponding pathogen allele will be strongly selected and increase in frequency, in turn reducing, potentially to a negative value, the selection coefficient associated with the over-represented host allele.

These two models of Red Queen Dynamics represent two extremes in a potentially broad continuum of possible dynamics. Besides, for a single coevolving pair of species, there might be several genes in each partner under coevolutionary pressures: some might follow arm's race dynamics, others fluctuating selection dynamics. This depends on the genetic determinism of the interaction. Although they are very simplistic, understanding these two types of dynamics and their implications is necessary to then study more complex coevolutionary scenarios.

Investigating coevolution with time-shift experiments

During the early stage of empirical coevolution research, observation and fieldwork were the main methods employed (Ehrlich and Raven, 1964; Janzen et al., 1980). These studies made indirect inferences about the impact of reciprocal selection by analyzing spatial patterns of trait co-variation among populations and by conducting comparative and phylogenetic analyses of ecologically interacting groups. These initial investigations strongly suggested that coevolution played a significant role in driving natural selection and influencing the structure and functioning of communities. Yet, being only observational and without controlling other parameters, they did not provide conclusive proof of reciprocal evolutionary changes.

In order to address the limitations of fieldwork, the study of coevolution has been brought to the laboratory. The controlled environment in the lab enables the exclusion of extraneous sources of selection, and the use of rapidly reproducing organisms such as microbes enables direct real-time observation of coevolutionary processes. An exciting method to study coevolutionary dynamics is the use of time-shift experiments in which parasites can be tested against hosts from different evolutionary time periods, i.e. the past, present or future (Gaba and Ebert, 2009; Gandon, Buckling, et al., 2008). This allows in particular the study of the effect of environment change on adaptation and the term we called Δr_{ec} in equation (19), by comparing fitness (or a linked trait like infectivity) across the different environments experienced through time by a pathogen population. However time-shift experiments come with caveats. They do not distinguish between specific or general adaptation, in other words it can be difficult to identify the selection pressure which drove adaptation. A 'future' pathogen being very fit when exposed to 'past' hosts could be due to coevolutionary forces driving phage adaptation, but also by adaptation of the phage to the specific medium in which it is cultured. In cases where adaptation to abiotic conditions can be discarded, time-shift experiments can also fail to distinguish between different coevolutionary dynamics, such as ARD of FSD described above. Depending on the studied time scale, it is predicted that those two coevolutionary scenarios could yield similar pattern of adaptation in time-shift experiments (Gandon, Buckling, et al., 2008). More particularly with FSD, the cyclical dynamics can completely blur the signal of adaptation depending on the time scale of the experiment relative to the period of the dynamics.

Local adaptation

In parallel with time-shift experiments, local adaptation experiments can be carried out to determine whether a pathogen is better adapted to infect hosts from the same or from a different population. The idea behind such studies is in common with time-shift experiments: that pathogen will be better adapted to host that they have often encountered recently. The selection process should favor any parasite that can infect commonly occurring genotypes among its local host population. Consequently, a well-adapted population of parasites is one that is capable of infecting a relatively high proportion of host genotypes within the local population. If there is some spatial structure in the distribution of host and parasite populations, with limited migration between sites, and if there is genetic specificity involved in the infectivity of parasites and the resistance of hosts (meaning that there are no

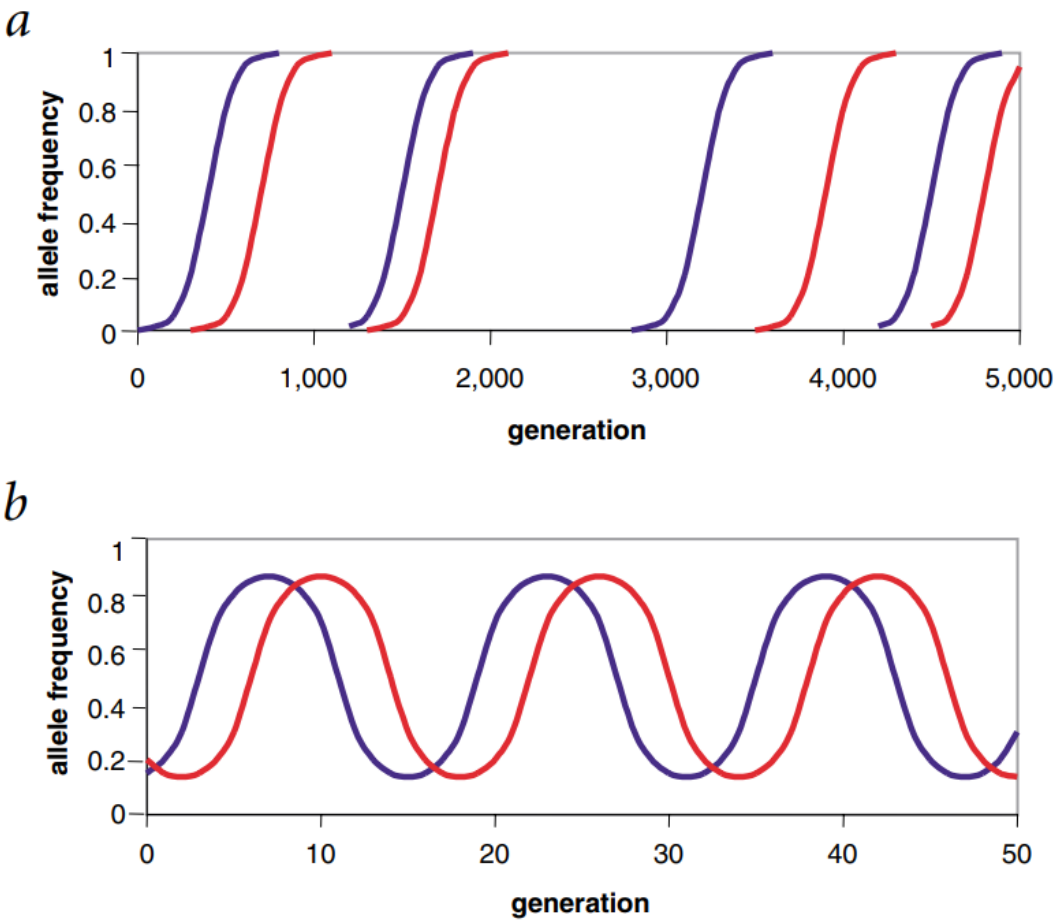


Figure 4: Allele frequency changes driven by co-evolution. a, A series of selective sweeps by host (blue line) and pathogen (red) alleles derived by mutation. Selection is directional, that is, genetic change accumulates in both populations. At any given stage of the process there may be polymorphism in either, both or neither of the two populations. b, Dynamic polymorphism in both host (blue) and pathogen (red) acting on existing genetic variation. Evolution is non-directional. At all stages of the process both populations are polymorphic. A large range of models predict, at least for certain parameter combinations, the kind of lagged limit cycles shown here. Figure from (Woolhouse, Webster, et al., 2002)

parasites universally capable of infecting all hosts nor hosts universally resistant to all parasites), the process of parasite adaptation should lead to parasites that, on average, perform better on local host genotypes than on genotypes from other host populations.

Although seemingly appealing, this hypothesis that pathogens will tend to be locally adapted is not a general rule. The same reasoning on pathogen adaptation to recent sympatric host can be applied to the adaptation of hosts to recent sympatric pathogens. It is then difficult to make prediction on whether host or pathogen will be locally adapted as it depends on the underlying genetic determinism of the interactions, the mutation rate, the generation time etc. There are documented examples of both pathogen local adaptation and local maladaptation (equivalent to host local adaptation) (Woolhouse, Webster, et al., 2002). This viral maladaptation to local hosts is well illustrated in the case of HIV viruses which were found to have a significantly lower fitness when confronted to the antibodies of the patient in which they were recovered, compared to other patients (Blanquart and Gandon, 2013). For bacteria and phages interactions, (Buckling and Rainey, 2002) showed that when coevolving 12 pairs of *pseudomonas fluorescens* with a naturally associated phage, the coevolved bacteria were more resistant to the phages with which they had coevolved than phages from other populations.

Coevolution in CRISPR systems

Many studies of coevolution, including time-shifts and local adaptation experiments, have used phenotypic measures to characterize adaptation. In particular, coevolution in host-pathogens systems has been studied by estimating proxies for pathogen fitness and host resistance. These phenotypic measurements are a natural first step to study such systems, as they do not require precise knowledge of the underlying genetic determinism of the host-pathogen interaction. Indeed measuring host resistance, and conversely pathogen infectivity, can be done without information on the mechanisms dictating the outcome of the host-pathogen interaction. However, such phenotypic approaches are insufficient to answer all the questions about coevolutionary dynamics (Gandon, Buckling, et al., 2008). Only measuring phenotypes has shortcomings as it only provides data at the population level, eg. a measure of infectivity of one population of bacteriophages against one population of bacterial hosts. When populations are diverse, assessing this diversity and obtaining the frequency of every phenotype can only be done repeating experiments on as many clones as

possible. Practically, these phenotypic approaches also require massive amounts of work to carry out the adaptation experiments we have in the two previous sections, for instance time-shift experiments require cross-infections from many pairs of time points in many replicates to provide meaningful insights (Koskella, 2014). Besides, it is necessary to decipher the genetic aspects underlying the interaction, such as the number of loci and alleles involved, to obtain a comprehensive understanding of the host-parasite interaction. This additional information is crucial in completing the overall understanding of the coevolution, and allows for the use of sequencing to monitor coevolutionary dynamics instead of phenotypic testing.

For these reasons, CRISPR-resistant bacteria and bacteriophages are a great system to study coevolution. The genetic determinism of the interaction is well known, which means that the phenotypes of resistance and infectivity of hosts and pathogens can be thoroughly described with high-throughput sequencing data. For hosts, deep sequencing of the CRISPR locus will provide the frequency of each spacer in the population. This yields the frequency of host resistance against every bacteriophage genotype. Conversely, whole-genome sequencing of the bacteriophage population provide information on the frequency of each potential escape mutations (i.e. mutations occurring in a protospacer). Coupling data on host resistance and phage infectivity yields a complete picture of the network of infection arising when a certain population of phages is confronted to a certain population of bacteria. When using this approach in a coevolution experiment, phenotypes are known for both hosts and pathogens at all potential time points and replicates. This means that fitness (or proxies for fitness) can be computed for all possible combination of host and pathogen populations, yielding *in silico* local adaptation and time-shift experiments without further experimental work.

As we have pointed out before, a growing number of bacterial defense systems against phages are being discovered, and many bacteria will carry several of these defense systems, stacking layers of resistance. For instance, many experiments have been carried out using *Pseudomonas aeruginosa* and its phage DMS3. Using this system it has been possible to show how the level of diversity of spacers in the host population can drive phage population to extinction (Common, Walker-Sünderhauf, et al., 2020; Houte et al., 2016; Morley et al., 2017) or limit their emergence (Chabas et al., 2018). Yet this experimental system has the limitation that *P. aeruginosa* can also develop resistance to phage DMS3 through the loss of its pilus (Westra, Houte, et al., 2015), which can interfere with the coevolution due to CRISPR resistance.

However in another system, that of *Streptococcus thermophilus* and its virulent phage 2972, the evolution of systems of resistance to the phage alternative to CRISPR is almost never observed (Westra and Levin, 2020). Interestingly in this system, the rate of acquisition of spacer by the host is lower compared to the *P. aeruginosa* system, which can result in extended periods of coevolution (Common, Morley, et al., 2019; Paez-Espino et al., 2015). As predicted by theory (Childs et al., 2014), this lower rate of spacer acquisition increases the likelihood of coevolution as spacer diversity is slower to build up. Phages are able to escape resistance with point mutations which in turn leads to the acquisition of additional spacers of resistance by the host population, resulting in Arms Race Dynamics. Yet in this system, the Arms Race is not symmetric: bacteria can become resistant to the whole phage population with the acquisition of one additional spacer. On the other hand, phages need to track the different individual CRISPR genotypes in the host population, with a specific combination of mutations granting escape to a single host genotype. Thus this coevolutionary scenario does not fit neither Arms Race Dynamics nor Fluctuating Selection Dynamics. A better representation for the polygenic coevolution in CRISPR pathosystems would be that of Chase Dynamics (Brockhurst et al., 2014; Gavrilets, 1997; Kopp and Gavrilets, 2006): bacteria acquire new spacers, thus resisting phages by moving in the multidimensional CRISPR genotypic space. On the other hand, phages chase the bacteria in this genotypic space with the acquisition of escape mutations. Again with this model, the phage population can only chase a limited number of host genotypes, and is thus driven to extinction when the host population becomes too diverse.

Objectives of Chapter 3

The *Streptococcus thermophilus* and phage 2972 host pathogen system offers unique possibilities to monitor coevolution. With sequencing the CRISPR locus of the bacteria and the whole genome of the phages, we can follow the frequencies of the different CRISPR host genotypes as well as the frequencies of phage escape mutations. With this information we can describe the infection network through time. In this thesis, we design a experiment to monitor coevolution in this system, starting from a diverse population of bacteria with 16 resistant as well as the fully susceptible wild-type strain.

We observed that the resistant bacteria we use widely vary in terms of growth rate, and that in a control without phages, host diversity was quickly lost through competition. As these frequencies of host shape the selection coefficient associated

with the escape mutations in the phage population, we can expect that this between-host competition will dramatically affect the evolution of the phage population. However it is not clear what this effect will be. A possible outcome would be a Kill-The-Winner scenario (Thingstad, 2000; Weinbauer, 2004) which would yield the following cycle:(1) a certain strain of bacteria outcompetes the other strains and increases in frequency, (2) phages adapt primarily to this most frequent host and the frequency of the corresponding mutation increases in the population, (3) these most frequent hosts are massively targeted by the phage population and decreases in frequency, being replaced by a new dominant bacterial strain. With our system, we are able to track whether such a cycle arises. We also use sequencing data to mimic time-shift and local adaptation experiments without requiring additional experiments.

Chapter 1:

Transient evolutionary epidemiology of viral adaptation

In preparation

Transient evolutionary epidemiology of viral adaptation

Martin Guillemet, Denis Roze, Guillaume Martin* and Sylvain Gandon*

Abstract: Viral evolution is fueled by adaptive mutations that drive the adaptation and the dynamics of the mean fitness of viral population. Yet, most mutations are not adaptive and the increase of mean fitness is hampered by deleterious and lethal mutations. This ambivalent role of mutations implies that it is unclear if a higher mutation rate boosts or slows down viral adaptation. Here we study the interplay between selection, mutation and epidemiological dynamics of viral populations under the assumption that the mutation rate is high and the effects of non-lethal mutations are small. We use this theoretical framework to show how the distribution of mutation effects can alter the transient dynamics as well as the long-term evolutionary outcome of viral populations. This work can be used to explore the feasibility of lethal mutagenesis to treat viral infections.

Introduction

The witin-host dynamics of viral infections depends both on the availability of susceptible host cells and the ability of the virus to infect and exploit these cells. This ability depends on multiple life-history traits and in particular on the transmission rate of the virus which measures the rate at which an infected host cell produces new infections. Mathematical epidemiology provides a theoretical framework to model how these life-history traits affect the dynamics of viral populations (Anderson and May, 1992; Diekmann, Heesterbeek, and Britton, 2013; Nowak and May, 2000).

Many viruses undergo high mutation rates (Sanjuán, Nebot, et al., 2010) which yields large amounts of genetic and phenotypic diversity within viral populations. This influx of mutations challenges the simplicity of classical models of viral dynamics and

has led to the concept of *quasipecies* to describe the dynamics of viruses with high mutation rates (Andino and Domingo, 2015; Domingo and Perales, 2019). Yet, the effects of high mutation rates can also be captured within the classical population genetics framework. As most mutations have deleterious effects, the constant influx of mutations generates a *mutation load* where the mean fitness of the population is lower than that of the fittest strain (Crow, 1989). In fact, some mutations can prevent viral replication and can be considered as *lethal mutations* (Sanjuán, Moya, and Elena, 2004). The massive impact of deleterious mutations on viral fitness led to the "lethal mutagenesis hypothesis" which states that there is a mutation rate above which viral population cannot grow and are driven to extinction (Bull, Sanjuan, and Wilke, 2007). Drugs increasing mutation rates are a potential therapeutic solution for many viral infections (Loeb and Mullins, 2000; Shiraki and Daikoku, 2020), including for SARS-CoV-2 (Driouich et al., 2021; Hadj Hassine, Ben M'hadheb, and Menéndez-Arias, 2022; Kaptein et al., 2020; Swanstrom and Schinazi, 2022). A better evaluation of the therapeutic potential of these drugs relies on a better understanding of the underlying dynamics leading to viral extinction when viral mutation rate is increased.

First, it is important to point out that viral extinction can occur from a fully deterministic model when the mutation load becomes overwhelmingly high (Martin and Gandon, 2010). But this effect can also be amplified in small and finite populations by Muller's ratchet (Felsenstein, 1974; Muller, 1964). In finite populations, the most fit, least-loaded genotype will be lost as drift overwhelms the effect of natural selection. This will result in a decreasing population, thus increasing the speed of the ratchet and the drop of mean fitness. There is thus a synergy between the demographic and evolutionary dynamics (Lynch, Bürger, et al., 1993; Lynch and Gabriel, 1990; Matuszewski et al., 2017). However, even a small influx of compensatory mutations can halt this process and lead to a steady state of mean fitness (Poon and Otto, 2000). The present work focuses on the analysis of deterministic models where we neglect the influence of demographic stochasticity and genetic drift. Since these effects are expected to speed up the drop of viral fitness and the risk of extinction it is important to keep in mind that our analysis is expected to yield more conservative estimations of the risk of viral extinction.

Second, it is important to realise that increasing mutation rate is a double-edged sword because some of the mutations may be beneficial. Hence, higher mutation rates induced by a mutagenic drug can potentially speed up adaptation (Bull, Joyce, et al.,

2013; Paff, Stolte, and Bull, 2014). Many models of lethal mutagenesis, do not account for beneficial and compensatory mutations, and this may amplify the efficacy of lethal mutagenesis. Another aspect which is often overlooked in previous models is the epidemiological setting. Diminishing the mean fitness of the viral population is expected to reduces the within-host growth rate. This could lead to an increased density of susceptible host cells, which may yield a better environment for the virus (i.e. more transmission opportunities), and eventually treatment failure. Both the influence of compensatory mutations and epidemiological feed-backs have been analysed in (Martin and Gandon, 2010). This study, however, focused on the long-term epidemiological and evolutionary within-host dynamics to identify the critical mutation rates allowing viral extinction. Yet, we currently lack a good understanding of the effects of higher mutation rates on the transient within-host dynamics of viral adaptation.

In the present work, we study the joint epidemiological and evolutionary within-host dynamics of a viral population subject to high mutation rates. We use Fisher's Geometric Model (FGM) to build a phenotype-to-life-history-trait map, which translates to fitness values through the epidemiological dynamics (Martin and Gandon, 2010). This geometric model of adaptation yields distributions of fitness effects of mutations that allows us to account for both deleterious and beneficial effects of mutations (Martin and Lenormand, 2006; Orr, 2000). Because, these fitness distributions depend on the parental genotype, the model allows for pervasive fitness epistasis between mutations (Tenaillon, 2014). In addition, we account for a distinct type of strictly lethal mutations. We use this model to go beyond the analysis of the joint epidemiological and evolutionary equilibrium of these populations (Martin and Gandon, 2010). In particular, we want to understand when an artificial increase of the mutation rate is expected to increase or decrease the mean fitness of the viral population.

Model

We want to model the joint epidemiological and evolutionary dynamics of a virus population spreading within a host. The virus has access to a density S of susceptible host cells and we want to model the dynamics of the density I of infected cells. We do not explicitly model the dynamics of the free virus stage because we assume the lifetime of an infected cell is much larger than a free virus particle (Martin and Gandon, 2010; Nowak and May, 2000). Host cells are produced from a reservoir at a constant rate b

and die at a per cell rate d. We assume that different strains (i.e. different phenotypes $\mathbf{x}$) may circulate within the host and $I_{\mathbf{x}}(t)$ refers to the density of host cells infected by strain $\mathbf{x}$. An infected cell of phenotype $\mathbf{x}$ infects susceptible cells with a transmission rate $\beta_{\mathbf{x}}$ and dies at a rate $(d + \alpha + Uf)$ where α is the virulence and Uf is the rate of lethal mutations. Lethal mutations lead to cells that are infected with a non-transmissible pathogen. As there is no density-dependence in our model, such cells do not influence further the dynamics of the system. Thus we do not follow the density of these cells, and we can simply treat lethal mutations as an additional death term . The overall death rate is not affected by the phenotype of the infected cell. We assume that each host cell can only be infected with a single viral strain (i.e. no multiple infections). The above life cycle yields the following system of ordinary differential equations (the upper dot represents time derivation):

$$\begin{aligned} \dot{S} &= b - S\,(\,\overline{\beta}\,I + d) \\ \dot{I} &= \overline{\beta}\,S\,I - (\alpha + d + Uf)I \end{aligned} \tag{1}$$

where $I(t) = \int I_{\mathbf{x}}(t)d\mathbf{x}$ is the total density of infected cells, $\overline{\beta} = \int p_{\mathbf{x}}\beta_{\mathbf{x}}d\mathbf{x}$ is the mean transmission rate and $p_{\mathbf{x}} = I_{\mathbf{x}}(t)/I(t)$ is the frequency of the phenotype $\mathbf{x}$ in the infected population. We can now introduce the per capita growth rate (i.e. malthusian fitness) of the phenotype $\mathbf{x}$ and the mean growth rate of the pathogen (i.e. the mean fitness) :

$$\begin{aligned} r_{\mathbf{x}} &= \beta_{\mathbf{x}}S(t) - (d + \alpha + Uf) \\ \overline{r} &= \int p_{\mathbf{x}}r_{\mathbf{x}}d\mathbf{x} = \overline{\beta}S(t) - (d + \alpha + Uf) \end{aligned} \tag{2}$$

The sign of the mean fitness can tell us whether a population of infected cells can invade a population of hosts of size S ($\overline{r} > 0$) or if it drops to extinction ($\overline{r} < 0$).

We define a phenotype $\mathbf{x}$ as a vector of size n, where each dimension refers to an independent continuous trait. To model the dependence of the transmission rate $\beta_{\mathbf{x}}$ on $\mathbf{x}$, we use a fitness landscape based on a quadratic Fisher's geometric model where the transmission rate depends on the euclidian distance of the vector $\mathbf{x}$ to the optimum (at the origin i.e. at $\mathbf{x} = (0, 0, .., 0)$) such that :

$$\beta_{\mathbf{x}} = \beta_0 - \frac{\|\mathbf{x}\|^2}{2s_\beta} \tag{3}$$

where $\|\mathbf{x}\|^2$ is the squared norm of the phenotype vector $\mathbf{x}$. The term s_β governs the curvature of the fitness landscape and is inversely linked to the strength of the directional selection towards the optimum where the transmission rate is β_0. Note that we

use this quadratic form to approximate a Gaussian shape of the fitness landscape (Martin and Gandon, 2010). Indeed, equation (3) can yield a good approximation when the virus is not too far from the optimum. Note that this approximation breaks in a certain phenotypic space far from the optimum where transmission rate can become negative, which corresponds to an additional set of lethal mutants that is ignored relative to Uf. We show how this fitness landscape links the n underlying phenotypes with the pathogens' life-history trait (i.e. transmission rate) in Figure 1. In the following, we will not work directly on the norm, but rather on the phenotypic traits themselves. We show in the Supplementary Information that we can consider that all the components x_i of the phenotype vector $\mathbf{x}$ are equally distributed without loss of generality. Considering that the distribution of each phenotypic trait x_i is Gaussian and the same for all i, we can then describe the evolutionary dynamics by following just one dimension with p_x the probability in the infected population that a phenotypic trait is of value x, $\overline{x} = \int p_x x dx$ the mean value of a phenotypic trait and $V_x = \int p_x (x - \overline{x})^2 dx$ the phenotypic variance. This framework allows us to explicitly write the mean transmission rate and the variance in transmission rate V_β:

$$\begin{aligned} \overline{\beta} &= \beta_0 - L_L - L_M \\ V_\beta &= \frac{V_x(t)}{s_\beta} (2L_L + L_M) \\ L_L &= \frac{n}{2s_\beta} \overline{x}^2 \\ L_M &= \frac{n}{2s_\beta} V_x \end{aligned} \tag{4}$$

where L_L is the lag load andL_M is the mutation load (Lande and Shannon, 1996).

The lag load L_L depends on the phenotypic distance $\overline{x}$ to the optimum and represents the difference in transmission rate between the mean phenotype and the maximum transmission rate β_0. The mutation load L_M depends on the phenotypic variance, and represents the mean difference in transmission rate between a random phenotype in the population and the mean phenotype. These load terms depend on the number of dimensions n which corresponds to the so called "cost of complexity": being away from the optimum over more dimensions increases the overall burden. The flatness term s_β decreases both the lag and mutation load, as it makes the fitness landscape flatter thus less penalizing for phenotypes far from the optimum. The variance in transmission rate depends on both the lag and the mutation loads, scaled again by the genetic variance and the shape of the fitness landscape. This variance in transmission rate, and thus the mutational and lag load, are what natural selection can act upon to increase the mean

transmission rate.

Next we need to describe the dynamics of the density $I_{\mathbf{x}}(t)$ of cells infected by each phenotype where U is the rate of mutation of the virus. This mutation process is assumed to be constant and unconditional on a transmission event. With probability f this mutation is lethal and so Uf is an additional mortality term for the infected cells. With probability $1-f$ this mutation is non-lethal and the new phenotype becomes $\mathbf{x}+\mathbf{y}$, where the mutation effect $\mathbf{y}$ is sampled in an isotropic multivariate normal distribution with mean 0 and variance λ. This mutation process yields the following integro-differential equation:

$$\dot{I}_{\mathbf{x}}(t) = \beta_{\mathbf{x}} S(t) I_{\mathbf{x}}(t) - (d+\alpha+Uf) I_{\mathbf{x}}(t) - U(1-f) I_{\mathbf{x}}(t) + U(1-f) \int I_{\mathbf{x}-\mathbf{u}}(t) \rho(\mathbf{u})\, d\mathbf{u}\,. \tag{5}$$

where ρ is the multivariate Gaussian kernel of mutational effects on phenotypes. The mutational variance λ is easily interpreted as it directly relates to the the mean effect of random mutations on transmission rate $\overline{\mu_\beta}$:

$$\overline{\mu_\beta} = \int \rho(\mathbf{u}) \big(\beta_{\mathbf{x}+\mathbf{u}} - \beta_{\mathbf{x}}\big) d\mathbf{u} = -\frac{n\lambda}{2 s_\beta} \tag{6}$$

We can directly relate this to the mean effect of mutation on fitness $\mu_r = S(t)\mu_\beta$ which is also dependent on the density of infected cells at time t. Note that this quantity is always negative and independent of the phenotype $\mathbf{x}$, meaning that mutated strains are always, on average, worse than their parental strain.

In this paper, we use different approaches to monitor the epidemiological and evolutionary dynamics described in (5). First for the sake of simplicity, we use a Weak Selection Strong Mutation (WSSM) approximation which implies that adaptation is the result of many mutations of small effects. In this regime, the distribution of phenotypes remains Gaussian. Then we relax this assumption and we study how mutations of larger effect can affect the evolutionary dynamics of the virus using a moment closure approximation. Finally we check the robustness of our approximations using numerical simulations, where we study the dynamics of discrete phenotypes on a 2D grid, with or without the assumption of small mutational effects.

Results

We use equation (5) to derive the dynamics of the distribution of phenotypes $\mathbf{x}$, through equations describing the dynamics of the cumulants of this distribution. Assuming that mutations are frequent and of small effects, we can neglect terms in higher order of the mutational variance (i.e. λ^2, λ^3 etc.) to capture viral dynamics as a function of the first two moments of the phenotypic distribution (Bürger, 2000). This is effectively the Weak Selection Strong Mutation regime (WSSM) which also corresponds to a diffusive approximation. We also assume that the distribution of underlying genetic traits $\mathbf{x}$ is initially Gaussian, and it will remain so in the WSSM regime (shown in ref (Martin and Roques, 2016) for asexuals). These derivations yield the following dynamical system:

$$\begin{aligned} \dot{\overline{x}} &= \overbrace{-\frac{V_x\, S}{s_\beta}\overline{x}}^{\text{selection}} \\ \dot{V_x} &= \underbrace{-\frac{V_x^2\, S}{s_\beta}}_{\text{selection}} + \underbrace{U(1-f)\lambda}_{\text{mutation}} \end{aligned} \tag{7}$$

Note that this evolutionary dynamics is coupled with the epidemiological dynamics (1) to form a closed system. The first equation shows how the mean phenotype goes towards the optimum at a speed governed by (i) the amount of susceptible cells $S(t)$, (ii) the phenotype variance $V_x(t)$ and (iii) the mean distance to the optimum $\overline{x}(t)$. The dynamical equation for V_x captures the balance between the effect of natural selection which decreases the variance, and the effect of mutation which introduces more genetic variation. Interestingly, the number of dimensions n only appears in the epidemiological equations through $\overline{\beta}$ in (4) and does not directly influence the dynamics of the transmission rate, neither through selection nor mutation in the WSSM limit.

The equations on the dynamics $\overline{x}$ and V_x relate to the distribution of phenotypes $\mathbf{x}$. Using the assumption that this distribution is Gaussian, we can compute dynamical equations for the mean transmission rate $\overline{\beta}$:

$$\dot{\overline{\beta}} = \overbrace{V_\beta S}^{\text{selection}} + \overbrace{U(1-f)\overline{\mu_\beta}}^{\text{mutation}} \tag{8}$$

The mean transmission rate is increased by natural selection with a speed controlled by the variance in transmission rate, scaled by the density of susceptible cells. The direct effect of mutations on the mean transmission rate is deleterious and equal to the rate of non-lethal mutations times the mean effect of these mutations on transmission rate.

Indeed, as discussed above, the expected effect on the life history-trait is deleterious (see equation (6)). In other words, we obtain a simple dynamical equation by working on the dynamics of the mean transmission rate. Then why not decide to describe the evolutionary dynamics by working directly on the life-history traits β instead of the phenotype $\mathbf{x}$? In the supplementary information, extending the approach used by Martin and Roques, 2016 we show how it is possible to track the dynamics of the Cumulant Generating Function of the distribution of transmission rates. Under the assumption that the relative strength of selection is weak and mutation is frequent, as expected, we recover equation (7).

Evolutionary equilibrium

The long-term equilibrium of the viral populations is determined both by the epidemiological and the evolutionary dynamics governed by equations (1) and (7), respectively. The ultimate endemic equilibrium of the system is given by:

$$\begin{aligned}
\overline{x}(\infty) &= 0 \\
V_x(\infty) &= \sqrt{\frac{U(1-f)\lambda s_\beta}{S(\infty)}} = \frac{\sqrt{U(1-f)\lambda}\,(2\gamma - n\sqrt{U(1-f)\lambda})}{4\,(\alpha+d+Uf)} \\
S(\infty) &= \frac{\alpha+d+Uf}{\beta_0} + \frac{n\,\sqrt{U(1-f)\lambda}\,(2\gamma + n\sqrt{U(1-f)\lambda})}{8\,s_\beta\,\beta_0^2} \\
I(\infty) &= \frac{b}{\alpha+d+Uf} - \frac{d}{\beta_0} - \frac{d}{\alpha+d+Uf}\frac{n\,\sqrt{U(1-f)\lambda}\,(2\gamma + n\sqrt{U(1-f)\lambda})}{8\,s_\beta\,\beta_0^2}
\end{aligned} \tag{9}$$

where $\gamma = \sqrt{4\,s_\beta\,(\alpha+d+Uf)\beta_0 + \frac{n^2U(1-f)\lambda}{4}}$. Interestingly, we note that if the transmission rate of every infected cell was $\beta = \beta_0$ then the equilibrium densities would be $S(\infty) = \frac{\alpha+d+Uf}{\beta_0}$ and $I(\infty) = \frac{b}{\alpha+d+Uf} - \frac{d}{\beta_0}$. As expected, the constant influx of non-lethal mutations introduces a mutation load that leads to a reduced density of infected cells at equilibrium (and an increased density of susceptible cells).

The above equations can be used to identify the critical mutation rate above which the viral population goes to extinction (see also supplementary information). Interestingly, Figure 2 shows that the fraction of lethal mutations has a non-monotonous effect on the critical mutation rate. We get for the equilibrium mean fitness of the infected population:

$$\overline{r}(\infty) = S(\infty)\Big(\beta_0 - \frac{n}{2}\sqrt{\frac{U(1-f)\lambda}{S(\infty)\,s_\beta}}\Big) - (d+\alpha+Uf) \tag{10}$$

The critical mutation rate can be obtained solving this equation for $\overline{r}(\infty) = 0$ after setting $S(\infty) = b/d$, the density of susceptible cells in the absence of viruses. This expression is useful to discuss the effects of lethal mutations on viral dynamics. The final term in (10) accounts for the direct effect of lethal mutations: increasing the proportion of lethal mutations increases the *death rate* of infections, and consequently decreases the critical mutation rate leading to viral extinctions. Yet, lethal mutations have an additional effect on the *birth rate* of infections. Indeed, the first term in equation (10) refers to the rate of new infections and this rate drops with the mutation load. This mutation load drops when most mutations are lethal because only viable mutations are accounted for in this load. Hence, increasing the proportion of lethal mutations decreases the mutation load. This effect is relatively small when the number n of phenotypic dimensions is small, but it can overwhelm the direct effect of lethal mutations when n becomes large. In other words, we find that the effect of the cost of complexity, which is expected to decrease the critical mutation rate for larger values of n, is actually dependent on the proportion of lethal mutations f. In fact, Figure 2 shows that beyond a given level of complexity, increasing the proportion of lethal mutations requires larger mutation rates to drive viral populations to extinction. We can also note that when mutations are all lethal, the number of phenotypic dimension has no impact on the critical mutation rate for pathogen extinction. An additional aspect which is hidden in equation (10) is that increasing the proportion of lethal mutations can increase the density of infected cells, thus providing a benefit for viral fitness.

Transient dynamics

We can jointly use equations (1), (4), (7) and (8) to follow the joint transient dynamics of the hosts and viral populations. In Figure 3 we explore the dynamics of the mean transmission rate $\overline{\beta}(t)$ from several initial conditions. We vary the initial distance of the mean phenotype to the the optimal phenotype, and for each of these distances we contrast a scenario where we start from a clonal population ($V_x(0) = 0$, black line) and a scenario with some standing genetic variance ($V_x(0) = 0.1$, red line). Regardless of these initial conditions, the dynamics converge to the same equilibrium, which is given by (9). However, the initial condition dictates the initial dynamics and the speed at which this equilibrium is reached. First, the standing genetic variance induces a mutation load, which explains the drop in the initial value of $\overline{\beta}$ in Figure 3. Second, the absence of a genetic variation in the clonal population implies that the speed of adaptation is initially very slow. In fact, the mean transmission rate initially drops

because of the effect of deleterious mutations (see equation 8). Genetic variation first needs to build up before adaptation can act on the mean transmission rate. In contrast, the speed of adaptation is faster with standing genetic variance. This faster adaptation allows to rapidly overcome the initial mutation load and the mean transmission rate $\overline{\beta}$ becomes rapidly higher than in initially clonal populations.

However, the epidemiological dynamics of the virus population is driven by $\overline{r}$ and not by $\overline{\beta}$. To better understand the dynamics of viral adaptation, it is useful to decompose the dynamics of viral mean fitness into separate effects following the framework of Gandon&Day (Gandon and Day, 2009):

$$\dot{\overline{r}} = \Delta r_{ns} + \Delta r_m + \Delta r_{ec} \tag{11}$$

with Δr_{ns}, Δr_m and Δr_{ec} refer to the changes in mean fitness due to natural selection, mutation and environmental change, respectively.

$$\Delta r_{ns} = V_r = S^2(t) V_\beta = \frac{V_x(t) S(t)^2}{s_\beta} (2L_L + L_M) \tag{12}$$

$$\Delta r_m = -S(t)\, U(1-f) \overline{\mu_\beta} \tag{13}$$

$$\Delta r_{ec} = \overline{\beta}(t) \dot{S}(t) \tag{14}$$

First, as expected from Fisher's fundamental theorem the change of mean fitness from natural selection is always positive and equal to the variance in fitness. This variance increases with the lag load (the farther a phenotype is from the optimum, the larger the strength of selection towards this optimum) and the mutation load (even if this load has a negative impact on mean fitness, it has a positive influence on the speed of adaptation). Selection is also fueled by the population of susceptible cells and the phenotypic variance, scaled by the shape of the fitness landscape, s_β. Second, the effect of mutations on mean fitness is simply the mean effect of mutations on fitness $\overline{\mu_\beta}$ multiplied by the influx of non-lethal mutation $U(1-f)$. This effect is always negative in our model. Note that this quantity is exactly equal to the drop in mean fitness in mutation accumulation experiments where the radical bottlenecking at each passage ensures that natural selection does not operate (because the variance in fitness $V_r = 0$). Finally, the third term accounts for the environmental change consecutive to a drop in the density of susceptible cells. This final term can be either positive or negative, depending on the change in the density of susceptible host cells. During the initial

phase of an infection, the density of susceptible cells is expected to drop and to have a negative impact on the growth rate of the epidemic (density-dependent regulation). In contrast, during the initial phase of therapy, drugs are expected to reduce the density of infected cells and, consequently, the density of susceptible cells may increase. As illustrated in Figure 4, this epidemiological feedback may have a huge impact on the dynamics of mean fitness.

Relaxing the weak selection approximation

The above analysis relies on a weak selection assumption which allowed us to focus on first order terms in λ and neglect higher order terms as $\lambda \ll U$ (see (Martin and Roques, 2016)). In other words, we assumed that the adaptation of the pathogen proceeds through the accumulation of many small-effect mutations, and that mutations of strong effects are rare and can be neglected. This assumption implies that the phenotypic distribution is always Gaussian. In the following, we explore what happens when we relax this assumption and account for stronger effects of mutations. These larger effect mutations induce a deviation from the Gaussian distribution of phenotypes that can be captured as a first approximation by the dynamics of the third and fourth cumulants. The dynamics of the phenotypic distribution satisfies (see supplementary information):

$$\begin{aligned}
\dot{\overline{x}}(t) &= -\frac{S(t)}{2s_\beta}\left(K_3(t) + 2V_x(t)\,\overline{x}(t)\right) \\
\dot{V}_x(t) &= -\frac{S(t)}{2s_\beta}\left(K_4(t) + 2\overline{x}(t)\,K_3(t) + 2V_x^2(t)\right) + U\,(1-f)\,\lambda \\
\dot{K}_3(t) &= -\frac{S(t)}{s_\beta}\left(\overline{x}\,K_4(t) + 3K_3(t)\,V_x(t)\right) \\
\dot{K}_4(t) &= -\frac{S(t)}{s_\beta}\left(4K_4(t)\,V_x(t) + 3K_3^2(t)\right) + 3U\,(1-f)\,\lambda^2\,.
\end{aligned} \tag{15}$$

Where $K3(t)$ and $K4(t)$ are respectively the third and fourth cumulants of the distribution of phenotypes. Note that we recover the dynamics of system (7) for the mean phenotype and the phenotypic variance, but with the additional effect of $K3$ and $K4$, the third and fourth cumulants of the phenotypic distribution. As expected, when larger effects of mutations can be neglected (i.e. here $\lambda^2 = 0$), both $K3$ and $K4$ converge to 0 and we recover (7). Yet, when λ^2 cannot be neglected, the influx of new mutations increase the cumulant $K4$. A positive $K4$ means that the phenotypic distribution is both more peaked around the mean phenotype and more heavily tailed, with less intermediate phenotypes than in a Gaussian distribution. This fourth cumulant is also expected

to generate transiently some skewness K_3 (which sign is inverse to th sign of $\overline{x}$). This is expected to transiently increase the variance in fitness and speed up viral adaptation. In the long run, however, the skewness K_4 is expected to vanish as the viral population gets closer to the optimal phenotype. Yet, the kurtosis remains even at the equilibrium and it decreases the equilibrium phenotypic variance. This lower phenotypic variance results in a lower mutation load. As discussed above with the effect of lethal mutations, increasing λ has a similar effect as in the WSSM because it increases the efficacy of natural selection and allows to get rid of deleterious mutations.

The above analysis breaks down when λ becomes too high relative to the standing variance (see Supplementary Information). But we can use another approximation to describe the viral dynamics under a regime of mutation where the variance of mutation overwhelms the effect of the parental strain. The classical "House of Cards" approximation has been used to derive the equilibrium mutation load (Bürger, 2000). After incorporating the influence of epidemiological feedbacks we obtain the following expectation for the phenotypic variance at equilibrium:

$$V_x^{HC}(\infty) = \frac{2\,U(1-f)\,s_\beta}{n\,S(\infty)} \tag{16}$$

A striking feature of this regime is that the equilibrium phenotypic variance is independent of the mutational variance λ. We compare in figure 6 the values of equilibrium phenotypic variance predicted by our models and under the "House of Cards" (HC) approximation with numeric simulations. We explore this with several values of s_β which varies the strength of selection and thus increases the effect of mutations on transmission rate. We see that the variance predicted with the relaxed WSSM is very accurate when the strength of selection is lower, but as previously mentioned, this approximation breaks down when the effect of mutations becomes high. In this case (low s_β) the HC approximation is more accurate and the predicted variance is smaller than that computed with a WSSM approximation. Different expectations on the equilibrium phenotypic variance (and thus mutation load) leads to different thresholds for the critical mutation rate. Indeed, if the expected load is lower, the mutation rate required to achieve lethal mutagenesis is higher. We show how the critical mutation rate changes with the strength of selection under the three approximations in the Supplementary Information. Interestingly, the critical mutation rate predicted under the HC approximation depends only on the demographic parameters but not on parameters that describe the fitness landscape. Overall, the critical mutation rate in the HC approximation is the same as the one obtained under the WSSM approximation but if

all mutations were lethal.

Discussion

In this work we built a model for the within-host adaptation of pathogens through the evolution of the transmission rate. We assume an explicit model of mutations on n underlying phenotypic traits determining transmission rate $\beta_{\mathbf{x}}$ *via* an optimization function. We couple this model of adaptation with an epidemiological model, and demonstrate that the demographic response in density of susceptible cells imposes a feedback on the adaptation of the pathogen. We use a modeling approach based on the cumulants of the phenotypic or transmission rate distribution, either through the Cumulant Generating Funcion, or by direct derivation of the dynamics of the cumulants. This direct derivation approach is treated thoroughly by Bürger (Bürger, 2000) in a case where the fitness is only dependent on the traits. We extended this work to a case of an asexual pathogen, which fitness is dependent on the dynamic density of hosts. We also chose an explicit mutational scheme with non-lethal mutations affecting the phenotype, and lethal mutations which are treated as a genotype-independent additional mortality. We explore the effects of the two types of mutation, and find a striking effect when the effect of non-lethal mutations in strong. For example we find that at higher phenotypic dimension - which increases the effect of mutations - non-lethal mutation can be more deleterious than strictly lethal mutations to the survival of the pathogen. Indeed, lethal mutations have a strong but instantaneous effect, whereas non-lethal deleterious mutations can accumulate over time and bring down the mean transmission rate of the virus, and thus its fitness. Thus we highlight the importance of the number of phenotypic dimensions or phenotypic complexity, which can be defined as the number of quasi-independant traits under optimizing selection. There have been several attempts to estimate this value of complexity for different organisms (reviewed in (Tenaillon, 2014)) and estimates vary broadly with different methods. Yet some methods yield complexity values of up to 40 for viruses, showing that considering relatively high values of n could be relevant in studying evolutionary dynamics.

We also compare our results to the classical "House of Cards" approximation which describes evolutionary equilibria when mutations are of strong effects ($\lambda \gg U$). This approach leads to a lower equilibrium phenotypic variance, which we show translates into a higher critical mutation rate. Interestingly, in this regime, mutations that we model as non-lethal have such strong effects that they are $de facto$ considered lethal. In fact, this prediction is the same as that obtained in the WSSM regime, when considering

that all mutations are lethal ($f = 1$).

In this work we studied the effect of epidemiology and the demographic feedbacks it imposes on the adaptation of a pathogen. This context leads to a pathogen fitness that is dependent on the density of susceptible cells and in return, the density of susceptible cells changes with the fitness of the pathogen. We show that the speed of pathogen adaptation increases with the availability of susceptible cells. As the mean fitness of the pathogen population increases, the density of susceptible cell decreases which slows down the speed of adaptation

Effects of stochasticity

Our analysis does not account for the potential effects of demographic stochasticity, which is expected to have major effects on the risk of extinction when the viral population size becomes low. It is important to distinguish between two effects of stochasticity.

First, stochasticity could affect the demographic dynamics without altering viral evolution. Such an approach has been used to compute probabilities of extinction when a deterministic model of evolution is coupled with stochastic growth rates in the absence of feedbacks (Anciaux et al., 2019). Interestingly, this analysis allowed to identify a maximal and a minimal value of the mutation rate allowing the population to escape extinction. The maximal value is analogous to our critical mutation rate, but the minimal mutation rate results from the demographic stochasticity which may yield viral extinction if beneficial mutations are not frequent enough to rescue an originally maladapted population. This analysis, however, did not account for epidemiological feedbacks on the viral population growth rates. These demographic feedbacks are likely to limit the risk of extinction if viral maladaptation results in an influx of susceptible cells. It would be interesting to extend their analysis with our current model to see how these feedbacks are expected to alter the predictions on the minimal mutation rate allowing evolution rescue.

Second, as the viral population gets smaller, genetic drift is expected to alter evolutionary dynamics. More specifically, Muller's ratchet has often been invoked to explain how the accumulation of deleterious mutations eventually leads to the decrease of population mean fitness. Taking into account this additional layer of stochasticity would further decrease pathogen mean fitness through an added drift load. Yet the magnitude of this effect could be minimal, as it has been shown with the FGM that Muller's ratchet is effectively negligible, except in very small populations (Poon and Otto, 2000).

Extension to the study of the evolution of two life-history traits

We built a model in which the phenotype of a virus only influences its transmission rate. However we could consider a dependency to other life-history traits such as virulence. In many studies, an explicit trade-off is used where transmission rate is a function of virulence (or vice versa). The underlying hypothesis is often related to the viral load: if there are more viral particles being produced, there is an increased cost on the host cell or organism, which leads to an additional mortality. In our model, we could add an additional dependence of virulence on the phenotype much like we did with transmission rate, except that the optimum for virulence is a minimum. If the evolutionary optima of several life-history traits are not superimposed on the landscape, we can expect the emergence of a trade-off between maximising transmission rate and minimizing virulence. An interesting aspect of this trade-off is that the effect of transmission rate on fitness is scaled by the density of infected cells, which is not the case for the effect of virulence (Day and Proulx, 2004; Gandon and Day, 2009). Thus selection is dependent on the epidemiological environment and the optimal strategy may not be the same according to the susceptible cells demographic parameters. Our model, which takes into account the demography of susceptible hosts is then particularly suited to study this behavior. However this would require additional analytic work to model the effect of a mutation on two life-history traits simultaneously, which depends on the angle between the direction to the two optima in the fitness landscape(Martin and Lenormand, 2015).

References

Anciaux, Y., A. Lambert, O. Ronce, L. Roques, and G. Martin (2019). "Population persistence under high mutation rate: from evolutionary rescue to lethal mutagenesis". In: *Evolution* 73.8, pp. 1517–1532 (cit. on p. 14).

Anderson, R. M. and R. M. May (1992). *Infectious diseases of humans: dynamics and control*. Oxford university press (cit. on p. 1).

Andino, R. and E. Domingo (2015). "Viral quasispecies". In: *Virology* 479, pp. 46–51 (cit. on p. 2).

Bull, J. J., P. Joyce, E. Gladstone, and I. J. Molineux (2013). "Empirical complexities in the genetic foundations of lethal mutagenesis". In: *Genetics* 195.2, pp. 541–552 (cit. on p. 2).

Bull, J. J., R. Sanjuan, and C. O. Wilke (2007). "Theory of lethal mutagenesis for viruses". In: *Journal of virology* 81.6, pp. 2930–2939 (cit. on p. 2).

Bürger, R. (2000). *The mathematical theory of selection, recombination, and mutation*. John Wiley & Sons (cit. on pp. 7, 12, 13).

Crow, J. F. (1989). "Some possibilities for measuring selection intensities in man". In: *Human biology* 61.5/6, pp. 763–775 (cit. on p. 2).

Day, T. and S. R. Proulx (2004). "A general theory for the evolutionary dynamics of virulence". In: *The American Naturalist* 163.4, E40–E63 (cit. on p. 15).

Diekmann, O., H. Heesterbeek, and T. Britton (2013). *Mathematical tools for understanding infectious disease dynamics*. Vol. 7. Princeton University Press (cit. on p. 1).

Domingo, E. and C. Perales (2019). "Viral quasispecies". In: *PLoS genetics* 15.10, e1008271 (cit. on p. 2).

Driouich, J.-S., M. Cochin, G. Lingas, G. Moureau, F. Touret, P.-R. Petit, G. Piorkowski, K. Barthélémy, C. Laprie, B. Coutard, et al. (2021). "Favipiravir antiviral efficacy against SARS-CoV-2 in a hamster model". In: *Nature communications* 12.1, p. 1735 (cit. on p. 2).

Felsenstein, J. (1974). "The evolutionary advantage of recombination". In: *Genetics* 78.2, pp. 737–756 (cit. on p. 2).

Gandon, S. and T. Day (2009). "Evolutionary epidemiology and the dynamics of adaptation". In: *Evolution* 63.4, pp. 826–838 (cit. on pp. 10, 15).

Hadj Hassine, I., M. Ben M'hadheb, and L. Menéndez-Arias (2022). "Lethal mutagenesis of RNA viruses and approved drugs with antiviral mutagenic activity". In: *Viruses* 14.4, p. 841 (cit. on p. 2).

Kaptein, S. J., S. Jacobs, L. Langendries, L. Seldeslachts, S. Ter Horst, L. Liesenborghs, B. Hens, V. Vergote, E. Heylen, K. Barthelemy, et al. (2020). "Favipiravir at high doses has potent antiviral activity in SARS-CoV-2- infected hamsters, whereas hydroxychloroquine lacks activity". In: *Proceedings of the National Academy of Sciences* 117.43, pp. 26955–26965 (cit. on p. 2).

Lande, R. and S. Shannon (1996). "The role of genetic variation in adaptation and population persistence in a changing environment". In: *Evolution*, pp. 434–437 (cit. on p. 5).

Loeb, L. A. and J. I. Mullins (2000). "Perspective-Lethal Mutagenesis of HIV by Mutagenic Ribonucleoside Analogs". In: *AIDS research and human retroviruses* 16.1, pp. 1–3 (cit. on p. 2).

Lynch, M., R. Bürger, D. Butcher, and W. Gabriel (1993). "The mutational meltdown in asexual populations". In: *Journal of Heredity* 84.5, pp. 339–344 (cit. on p. 2).

Lynch, M. and W. Gabriel (1990). "Mutation load and the survival of small populations". In: *Evolution* 44.7, pp. 1725–1737 (cit. on p. 2).

Martin, G. and S. Gandon (2010). "Lethal mutagenesis and evolutionary epidemiology". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1548, pp. 1953–1963 (cit. on pp. 2, 3, 5).

Martin, G. and T. Lenormand (2006). "A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species". In: *Evolution* 60.5, pp. 893–907 (cit. on p. 3).

— (2015). "The fitness effect of mutations across environments: Fisher's geometrical model with multiple optima". In: *Evolution* 69.6, pp. 1433–1447 (cit. on p. 15).

Martin, G. and L. Roques (2016). "The nonstationary dynamics of fitness distributions: asexual model with epistasis and standing variation". In: *Genetics* 204.4, pp. 1541–1558 (cit. on pp. 7, 8, 11).

Matuszewski, S., L. Ormond, C. Bank, and J. D. Jensen (2017). "Two sides of the same coin: A population genetics perspective on lethal mutagenesis and mutational meltdown". In: *Virus evolution* 3.1, vex004 (cit. on p. 2).

Muller, H. J. (1964). "The relation of recombination to mutational advance". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1.1, pp. 2–9 (cit. on p. 2).

Nowak, M. and R. M. May (2000). *Virus dynamics: mathematical principles of immunology and virology: mathematical principles of immunology and virology*. Oxford University Press, UK (cit. on pp. 1, 3).

Orr, H. A. (2000). "Adaptation and the cost of complexity". In: *Evolution* 54.1, pp. 13–20 (cit. on p. 3).

Paff, M. L., S. P. Stolte, and J. J. Bull (2014). "Lethal mutagenesis failure may augment viral adaptation". In: *Molecular biology and evolution* 31.1, pp. 96–105 (cit. on p. 3).

Poon, A. and S. P. Otto (2000). "Compensating for our load of mutations: freezing the meltdown of small populations". In: *Evolution* 54.5, pp. 1467–1479 (cit. on pp. 2, 14).

Sanjuán, R., A. Moya, and S. F. Elena (2004). "The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus". In: *Proceedings of the National Academy of Sciences* 101.22, pp. 8396–8401 (cit. on p. 2).

Sanjuán, R., M. R. Nebot, N. Chirico, L. M. Mansky, and R. Belshaw (2010). "Viral mutation rates". In: *Journal of virology* 84.19, pp. 9733–9748 (cit. on p. 1).

Shiraki, K. and T. Daikoku (2020). "Favipiravir, an anti-influenza drug against life-threatening RNA virus infections". In: *Pharmacology & therapeutics* 209, p. 107512 (cit. on p. 2).

Swanstrom, R. and R. F. Schinazi (2022). "Lethal mutagenesis as an antiviral strategy". In: *Science* 375.6580, pp. 497–498 (cit. on p. 2).

Tenaillon, O. (2014). "The utility of Fisher's geometric model in evolutionary genetics". In: *Annual review of ecology, evolution, and systematics* 45, pp. 179–201 (cit. on pp. 3, 13).

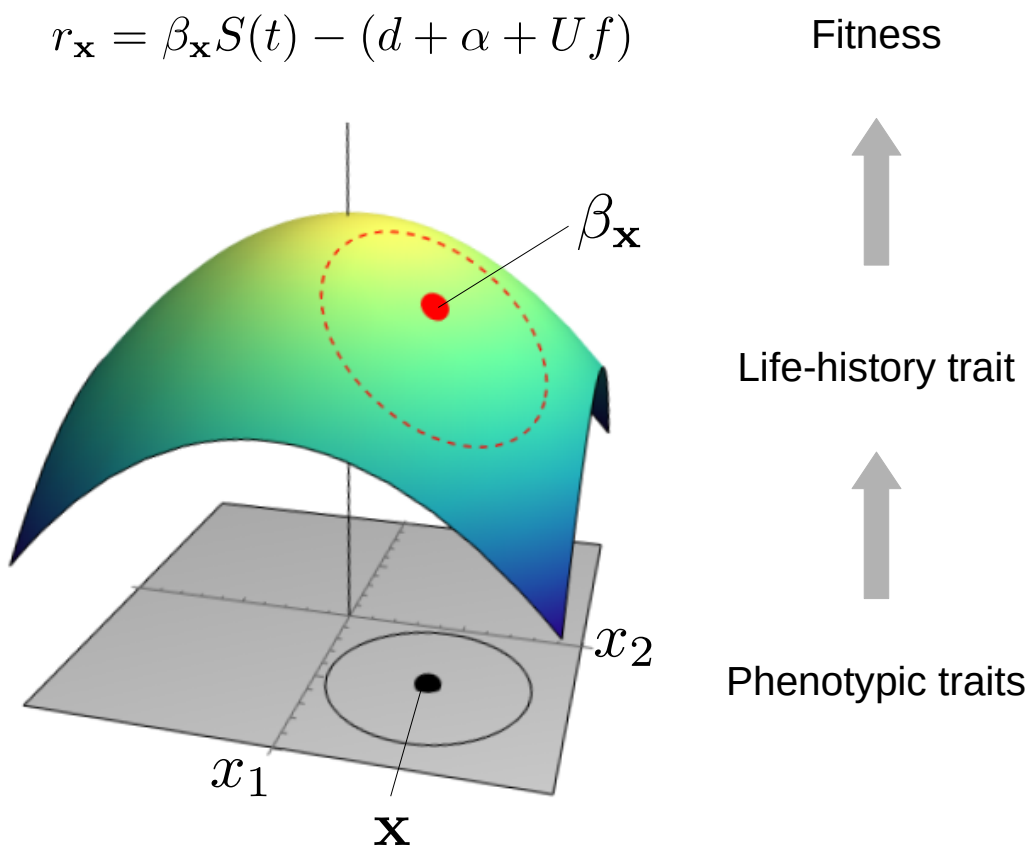


Figure 1: The fitness landscape links the underlying phenotypic traits with life-history traits and malthusian fitness. We represent our phenotype to life-history trait landscape in two dimensions ($n = 2$). A phenotype $\mathbf{x}$ is translated to a transmission rate $\beta_{\mathbf{x}}$ using equation (3). This life-history trait is translated to a fitness value with equation (2), which depends on time through the number of susceptible cells $S(t)$. A black circle is shown around phenotype $\mathbf{x}$, which is translated to a dashed red circle of transmission rates, showing how the FGM distorts the distributions.

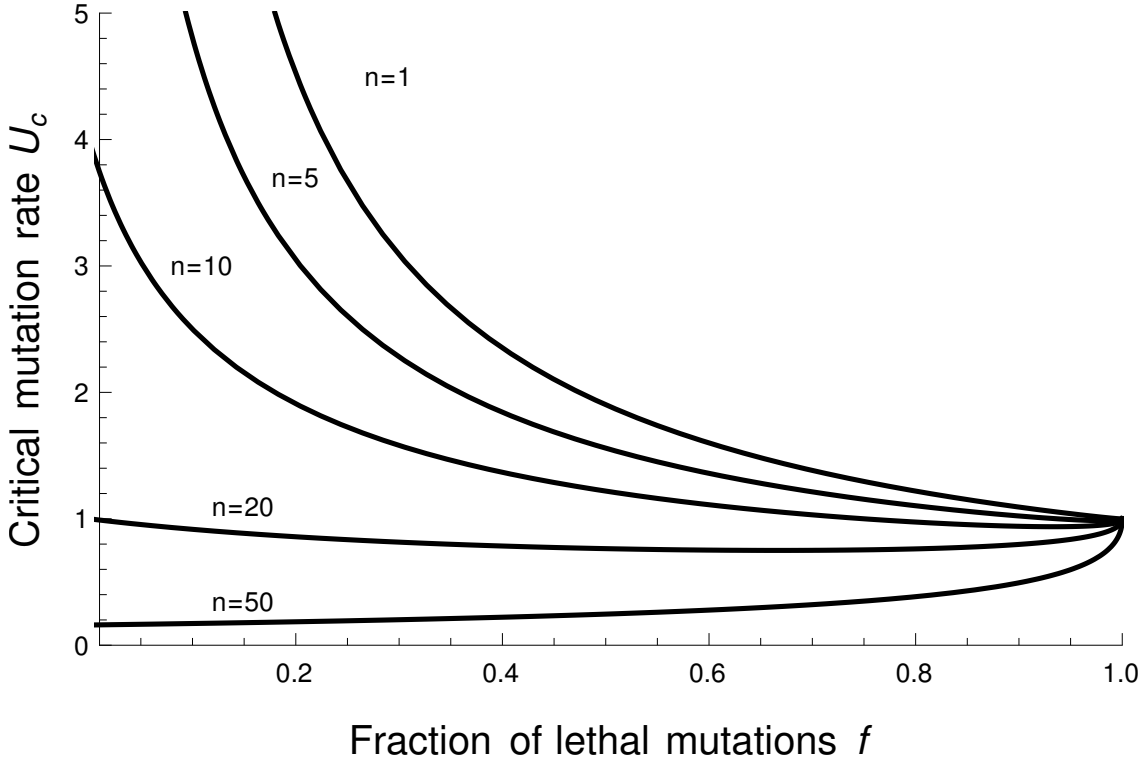


Figure 2: Effect of the proportion of lethal mutations and the phenotypic dimension of the fitness landscape on the critical mutation rate. The plot shows the effect of both fraction of lethal mutations f and number of dimensions n on the critical mutation rate required for lethal mutagenesis. The parameters used were $b = 2, d = 1, \beta_0 = 2, \alpha = 2, s_\beta = 1, \lambda = 0.005$.

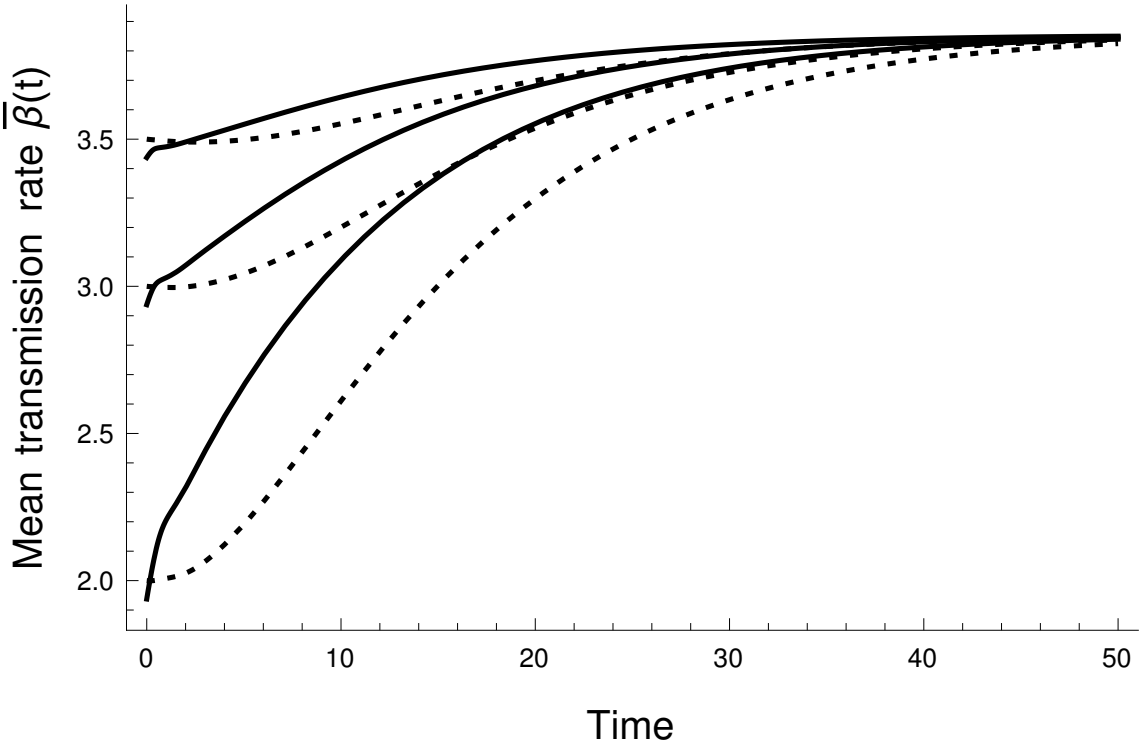


Figure 3: Adaptation dynamics depend on the initial phenotypic variance. We show the dynamics of mean transmission rate $\overline{\beta}(t)$ through time starting from a clonal populations at different initial values of $\overline{x}$. Represented in dashed black the infected population is initially clonal with $V_x(0) = 0$ while in solid black the population is initially diverse with $V_x(0) = 0.025$. The parameters were $b = 4, d = 1, \beta_0 = 4, \alpha = 2, s_\beta = 1, U = 1, f = 0.4, \lambda = 0.005, n = 5$. Initial conditions were $I(0) = 0.1, S(0) = b/d$ and we used three initial values for $\beta_{\overline{x}}$ were $2, 3$ and 3.5.

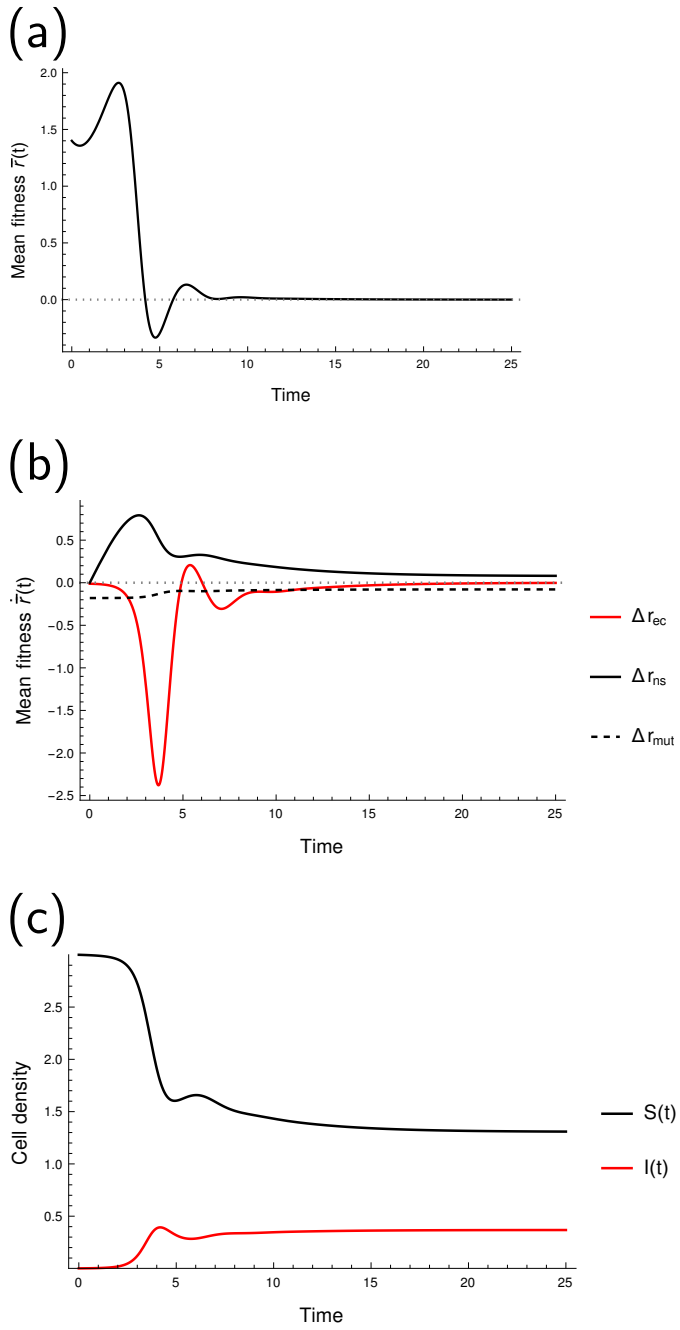


Figure 4: The three forces driving the dynamics of mean fitness. Mean fitness $\overline{r}$ starting from a clone ($V_x(0) \ll 1$) is shown through time (a) and the different components of its dynamics are represented in (b). The solid line in (b) shows the effect of natural selection (Δr_{ns}), the dashed black line shows the effect of mutation (Δr_{mut}) and the red solid line represent the feedback from the demography of the susceptible cells i.e. the environmental change (Δr_{ec}). The epiodemiological dynamics of the system are shown in (c). The parameters were $b = 3, d = 1, \beta_0 = 4, \alpha = 1, s_\beta = 1, U = 4, f = 0.4, \lambda = 0.005, n = 10$. Initial conditions were $I(0) = 0.001, S(0) = b/d, V_x(0) = 1e-8, \overline{\beta}(0) = 2$

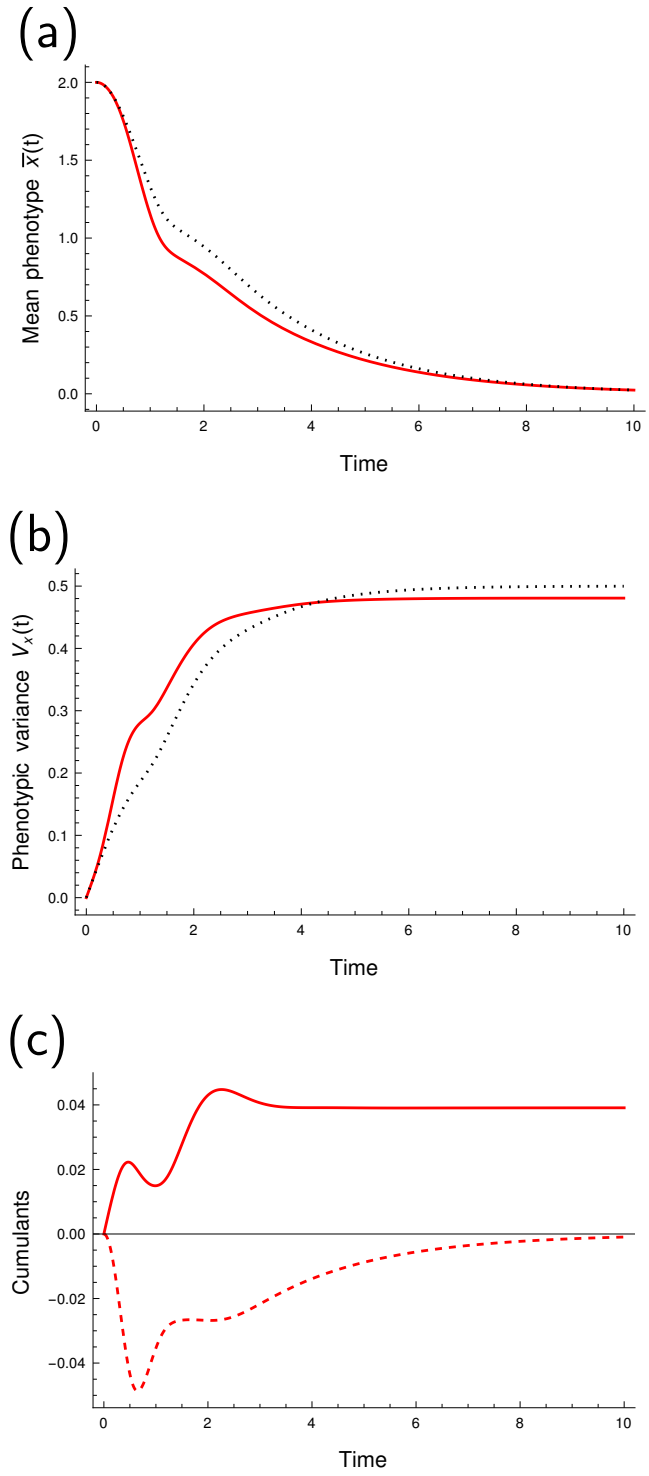


Figure 5: Higher moment mutations reduce genetic variance. The evolutionary dynamics are shown according to the WSSM model (black) or taking into account mutations of larger effect (red). We show (a) the mean phenotype $\overline{x}$, (b) the phenotypic variance V_x, (c) the genetic variance and finally (d) the third (dashed line) and fourth cumulant (solid line) of the genetic distribution in the full model. The parameter values used are: $b = 4, d = 1, \beta_0 = 4, \alpha = 1, s_\beta = 1, U = 4, f = 0.4, \lambda = 0.1, n = 20$. Initial conditions were $I(0) = 0.001, S(0) = b/d, V_x(0) = 1e-8, \overline{\beta}(0) = 2$

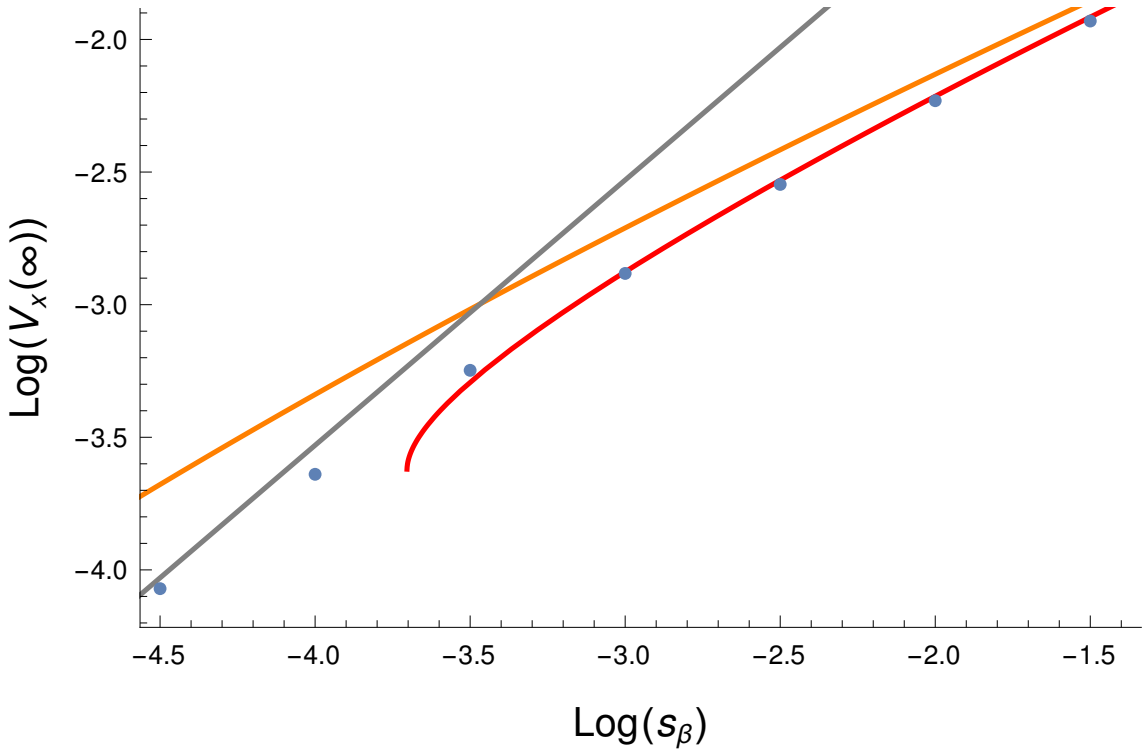


Figure 6: Comparison of the equilibrium variance under different regimes. The equilibrium variance V_x is shown as a function of the selection parameter s_β in a log-log scale. The results are shown for three derivations: the WSSM approximation in orange, the results from the cumulant approach up to K4 in red, and the "House of Cards" (HC) approximation in grey. Data from numerical simulations are shown as blue points. The parameters used were: $b = 3, d = 1, \beta_0 = 4, \alpha = 1, U = 4, f = 0.4, \lambda = 0.05, n = 2$.

Supplementary Information

In this supplementary information we detail the derivation of the theoretical results presented in the main text. In section 1, we start by presenting the main assumptions on the fitness landscape, the effects of mutations and the epidemiological dynamics. In sections 2 and 3, we present different ways to analyse the joint epidemiological and evolutionary analysis of this model and obtain analytical approximations. In section 4, we detail the numerical model used to check our analytical approximations.

1. The model

We model the within-host dynamics of of a viral population that spreads in a large population of susceptible cells. We assume that different strains of the virus may have different phenotypes. Each phenotype $\mathbf{x}$ is defined as a vector of size n, representing n independent continuous phenotypic traits. Each host cell is assumed to be infected by a single strain (no multiple infections) and the total density of infected cells is noted $I = \int I_{\mathbf{x}} d^n\mathbf{x}$ iand $p_{\mathbf{x}} = I_{\mathbf{x}}/I$ is the frequency of cells infected by strains with phenotype $\mathbf{x}$. We assume that the viral transmission rate $\beta_{\mathbf{x}}$ is governed by the values of the n underlying phenotypic traits through a Gaussian transmission function. This Gaussian is well approximated with a quadratic function when phenotypes are not too far from the optimum, ensuring no negative transmission rate values. This quadratic form yields:

$$\beta_{\mathbf{x}} = \beta_0 - \frac{\|\mathbf{x}\|^2}{2s_\beta} \tag{1}$$

where $\|\mathbf{x}\|$ is the norm of vector $\mathbf{x}$ and s_β measures the strength of selection towards a single optimum at $\mathbf{x} = (0, 0, ..., 0)$ of maximum transmission rate β_0 (see **Figure 1** in the main text). Different virus strains may have different transmission rates $\beta_{\mathbf{x}}$ and the average transmission rate is given by:

$$\overline{\beta} = \int p_{\mathbf{x}} \beta_{\mathbf{x}} d^n\mathbf{x} = \beta_0 - \frac{\sum_{i=1}^{n} (V_{x,i} + \overline{x_i}^2)}{2s_\beta}, \tag{2}$$

where $\overline{x_i}$ and $V_{x,i} = \mathrm{E}\left[(x_i - \overline{x_i})^2\right]$ are respectively the mean and the variance for the phenotypic trait i.

We model the whithin-host spread of the viral population through the dynamics of S and I the densities of uninfected and infected cells, respectively. We assume there is a constant influx b of susceptible cells which die at constant rate d. A cell infected with a virus strain with phenotype $\mathbf{x}$ infects new susceptible cells at a rate $\beta_{\mathbf{x}}$ and die at a rate $d+\alpha$ where α measures the increased mortality rate induced by the infection. Viral mutations may occur at a constant rate U and among those mutations a fraction f may be lethal for the virus. Consequently, the rate of lethal mutation Uf may be treated as an additional mortality term for infected cells. Note that, in the following, the dependence to t of most dynamical variables is dropped to simplify the notation. This yields the following dynamical system:

$$\dot{S} = b - (\overline{\beta} I + d)S \tag{3}$$

$$\dot{I} = \overline{\beta} SI - (d + \alpha + Uf)I \tag{4}$$

In the absence of the pathogen the density of the uninfected cells is equal to $S_0 = b/d$. In such a naive host population a pathogen with average transmission rate $\overline{\beta}$ can spread if and only if its basic reproductive ratio:

$$R_0 = \frac{\overline{\beta} S_0}{d + \alpha + Uf} > 1 \tag{5}$$

In other words a condition for the viability of the pathogen population at equilibrium is $\overline{\beta} > \frac{d(d+\alpha+Uf)}{b}$. We can write the dynamics of $I_{\mathbf{x}}$ the density host cells infected by strain x using the following integro-differential equation:

$$\dot{I}_{\mathbf{x}} = \beta_{\mathbf{x}} S I_{\mathbf{x}} - (d + \alpha + Uf) I_{\mathbf{x}} - U(1-f) I_{\mathbf{x}} + U(1-f) \int I_{\mathbf{x\text{-}u}} \rho(\mathbf{u}) d^n \mathbf{u}\,. \tag{6}$$

The first two terms describe the *birth rate* and the *death rate* of infections, respectively. The following term describe the mutation away from strain $\mathbf{x}$, and the last term describes the mutations from all other strains to strain $\mathbf{x}$. At rate $U(1-f)$ (the effect lethal mutations have already been accounted in the *death rate* of infections), strain **x-u** mutates to phenotype $\mathbf{x}$ with a probability $\rho(\mathbf{u})$ following an isotropic multivariate normal distribution $\mathcal{N}(0, I_n\sqrt{\lambda})$, where refers to the variance of the phenotypic effect on each phenotypic dimension.

2. Dynamics of the cumulants of the distribution of phenotypes

In this section we use the integro-differential equation (6) to derive the cumulants of the distribution of phenotypes $\mathbf{x}$. Cumulants are quantities directly related to the moments of the distribution as we show below. We use cumulants over moments because of a property of Gaussian distributions: their cumulants of order > 3 are equal to zero. We will use this property for "moment closure", ie. to neglect higher order cumulant in our dynamical equations to obtain a closed system of differential equations. Interestingly, one can relax this Gaussian approximation to a certain degree by taking into account these high order cumulants. In all our derivations we will assume that cumulants of order ≥ 5 can be neglected.

In the following, $C_{ij} = \mathrm{E}\left[\left(x_i - \overline{x_i}\right)\left(x_j - \overline{x_j}\right)\right]$ denotes the phenotypic covariance between traits i and j, while $C_{ijk} = \mathrm{E}\left[\left(x_i - \overline{x_i}\right)\left(x_j - \overline{x_j}\right)\left(x_k - \overline{x_k}\right)\right]$, $C_{ijkl}...$ denote higher-order moments. Cumulants K_U can be defined in terms of moments C_U from the definition of the cumulant generating function. In particular, we have:

$$K_{ij} = C_{ij} \tag{7}$$

$$K_{ijk} = C_{ijk} \tag{8}$$

$$K_{ijkl} = C_{ijkl} - C_{ij}C_{kl} - C_{ik}C_{jl} - C_{il}C_{jk} \tag{9}$$

$$\begin{aligned} K_{ijklm} = {} & C_{ijklm} - C_{ijk}C_{lm} - C_{ijl}C_{km} - C_{ijm}C_{kl} - C_{ikl}C_{jm} - C_{ikm}C_{jl} \\ & - C_{ilm}C_{jk} - C_{jkl}C_{im} - C_{jkm}C_{il} - C_{jlm}C_{ik} - C_{klm}C_{ij} \end{aligned} \tag{10}$$

$$K_{ijklmn} = C_{ijklmn} - \sum C_{ijkl}C_{mn} - \sum C_{ijk}C_{lmn} + 2\sum C_{ij}C_{kl}C_{mn} \tag{11}$$

where in the last expression the sums include all terms obtained by permuting indices. If the distribution of phenotypes is Gaussian, all cumulants of order ≥ 3 are zero. In the following, we derive expressions for the dynamics of mean phenotypes ($\overline{x_i}$), phenotypic variances and covariances (C_{ij}) and cumulants of order 3 and 4 (K_{ijk}, K_{ijkl}).

Changes in phenotype frequencies $p_{\mathbf{x}} = I_{\mathbf{x}}(t)/I_{\mathbf{x}}$ among the non-lethal viruses are given by:

$$\begin{aligned} \dot{p}_{\mathbf{x}} = {} & S\left(t\right)\left[\beta_{\mathbf{x}} - \overline{\beta}\right] p_{\mathbf{x}} - U\left(1-f\right) p_{\mathbf{x}} \\ & + U\left(1-f\right) \int p_{\mathbf{x}-\mathbf{u}} \rho\left(\mathbf{u}\right) d^n\mathbf{u}\,. \end{aligned} \tag{12}$$

2.1. Dynamics of mean phenotypes

Changes in mean phenotypes are given by:

$$\dot{\overline{x_i}} = \int x_i\, \dot{p}_{\mathbf{x}} d^n\mathbf{x}\,. \tag{13}$$

From equations 12 and 2, this gives:

$$\begin{aligned}\dot{\overline{x_i}} = -\frac{S\left(t\right)}{2s_\beta}\left[\int x_i \sum_{j=1}^{n} x_j^2\, p_{\mathbf{x}} d^n\mathbf{x} - \overline{x_i}\sum_{j=1}^{n}\left(V_{x,j}+\overline{x_j}^2\right)\right] \\ - U\left(1-f\right)\overline{x_i} + U\left(1-f\right)\int\int x_i\, p_{\mathbf{x}-\mathbf{u}}\rho\left(\mathbf{u}\right) d^n\mathbf{u}\, d^n\mathbf{x}\,.\end{aligned} \tag{14}$$

The last line of equation 14 cancels, while the integral on the first line may be written as:

$$\begin{aligned}\sum_{j=1}^{n}\mathrm{E}\left[x_i\, x_j^2\right] &= \sum_{j=1}^{n}\mathrm{E}\left[\left(x_i-\overline{x_i}+\overline{x_i}\right)\left(x_j-\overline{x_j}+\overline{x_j}\right)^2\right] \\ &= \sum_{j=1}^{n}\left[K_{ijj}+2C_{ij}\,\overline{x_j}+\overline{x_i}\left(V_{x,j}+\overline{x_j}^2\right)\right]\end{aligned} \tag{15}$$

Finally giving:

$$\dot{\overline{x_i}} = -\frac{S\left(t\right)}{2s_\beta}\sum_{j=1}^{n}\left(K_{ijj}+2C_{ij}\,\overline{x_j}\right). \tag{16}$$

2.2. Dynamics of second moments

Similarly, changes in phenotypic variances and covariances are given by:

$$\begin{aligned}\dot{C}_{ij} &= \int\left(x_i-\overline{x_i}\right)\left(x_j-\overline{x_j}\right)\dot{p}_{\mathbf{x}} d^n\mathbf{x} \\ &= -\frac{S\left(t\right)}{2s_\beta}\left[\int\left(x_i-\overline{x_i}\right)\left(x_j-\overline{x_j}\right)\sum_{k=1}^{n}x_k^2\, p_{\mathbf{x}} d^n\mathbf{x} - C_{ij}\sum_{k=1}^{n}\left(V_{x,k}+\overline{x_k}^2\right)\right] \\ &\quad - U\left(1-f\right)C_{ij} + U\left(1-f\right)\int\int\left(x_i-\overline{x_i}\right)\left(x_j-\overline{x_j}\right)\, p_{\mathbf{x}-\mathbf{u}}\rho\left(\mathbf{u}\right) d^n\mathbf{u}\, d^n\mathbf{x}\,.\end{aligned} \tag{17}$$

The integral in brackets on the second line of equation 17 may be written as:

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}\left[\left(x_i-\overline{x_i}\right)\left(x_j-\overline{x_j}\right)\left(x_k-\overline{x_k}+\overline{x_k}\right)^2\right] &= \sum_{k=1}^{n}\left(C_{ijkk}+2\overline{x_k}K_{ijk}+C_{ij}\overline{x_k}^2\right) \\ &= \sum_{k=1}^{n}\left(K_{ijkk}+C_{ij}V_{x,k}+2C_{ik}C_{jk}+2\overline{x_k}K_{ijk}+C_{ij}\overline{x_k}^2\right)\end{aligned} \tag{18}$$

(using equation 9) while the double integral on the third line of equation 17 may be written as:

$$\int p_{\mathbf{x}}\left[\int\rho\left(\mathbf{u}\right)\left(x_i-\overline{x_i}+u_i\right)\left(x_j-\overline{x_j}+u_j\right)d^n\mathbf{u}\right]d^n\mathbf{x} = C_{ij}+\delta_{ij}\lambda \tag{19}$$

where $\delta_{ij}=1$ if $i=j$, and 0 otherwise. Putting everything together yields:

$$\dot{C}_{ij} = -\frac{S\left(t\right)}{2s_\beta}\sum_{k=1}^{n}\left(K_{ijkk}+2\overline{x_k}K_{ijk}+2C_{ik}C_{jk}\right)+\delta_{ij}\,U\left(1-f\right)\lambda \tag{20}$$

2.3. Dynamics of third cumulants

The change in $K_{ijk} = C_{ijk}$ is given by:

$$\begin{aligned}\dot{K}_{ijk} &= \int (x_i - \overline{x_i})(x_j - \overline{x_j})(x_k - \overline{x_k})\,\dot{p}_{\mathbf{x}} d^n\mathbf{x} \\ &\quad + \int \frac{d\,(x_i - \overline{x_i})(x_j - \overline{x_j})(x_k - \overline{x_k})}{dt} p_{\mathbf{x}} d^n\mathbf{x} \\ &= \left(\int (x_i - \overline{x_i})(x_j - \overline{x_j})(x_k - \overline{x_k})\,\dot{p}_{\mathbf{x}} d^n\mathbf{x}\right) - \dot{\overline{x_i}}\,C_{jk} - \dot{\overline{x_j}}\,C_{ik} - \dot{\overline{x_k}}\,C_{ij}\,.\end{aligned} \tag{21}$$

As before, the integral on equation 21 writes:

$$\begin{aligned}&-\frac{S(t)}{2s_\beta}\left[\int (x_i - \overline{x_i})(x_j - \overline{x_j})(x_k - \overline{x_k}) \sum_{l=1}^{n} x_l^2\, p_{\mathbf{x}} d^n\mathbf{x} - K_{ijk} \sum_{l=1}^{n} \left(V_{x,l} + \overline{x_l}^2\right)\right] \\ &+ U(1-f) \int\int (x_i - \overline{x_i})(x_j - \overline{x_j})(x_k - \overline{x_k})\; p_{\mathbf{x}-\mathbf{u}}\rho(\mathbf{u})\, d^n\mathbf{u}\, d^n\mathbf{x} \\ &- U(1-f)\, K_{ijk}\,.\end{aligned} \tag{22}$$

The last two lines of equation 22 vanish under the assumption that third moments of the distribution of mutational effects ρ are zero, while the integral on the first line can be written as:

$$\sum_{l=1}^{n} \mathrm{E}\left[(x_i - \overline{x_i})(x_j - \overline{x_j})(x_k - \overline{x_k})(x_l - \overline{x_l} + \overline{x_l})^2\right] = \sum_{k=1}^{n} \left(C_{ijkll} + 2\overline{x_l} C_{ijkl} + K_{ijk}\overline{x_l}^2\right) \tag{23}$$

The moments C_{ijkll} and C_{ijkl} can be expressed in terms of second-order moments and third / fourth-order cumulants using equations 9 and 10 (assuming that the fifth-order cumulant K_{ijkll} equals zero). Putting everything together, one finally obtains:

$$\dot{K}_{ijk} = -\frac{S(t)}{s_\beta} \sum_{l=1}^{n} \left(\overline{x_l}\, K_{ijkl} + K_{ijl}\, C_{kl} + K_{ikl}\, C_{jl} + K_{jkl}\, C_{il}\right)\,. \tag{24}$$

2.4. Dynamics of fourth cumulants

Finally, from equation 9 we have:

$$\begin{aligned}\dot{K}_{ijkl} &= \dot{C}_{ijkl} - \dot{C}_{ij}\, C_{kl} - C_{ij}\, \dot{C}_{kl} - \dot{C}_{ik}\, C_{jl} - C_{ik}\, \dot{C}_{jl} \\ &\quad - \dot{C}_{jk}\, C_{il} - C_{jk}\, \dot{C}_{il}\,.\end{aligned} \tag{25}$$

The change in C_{ijkl} is given by:

$$\begin{aligned}\dot{C}_{ijkl} &= \int (x_i - \overline{x_i})(x_j - \overline{x_j})(x_k - \overline{x_k})(x_l - \overline{x_l})\,\dot{p}_{\mathbf{x}} d^n\mathbf{x} \\ &\quad + \int \frac{d\,(x_i - \overline{x_i})(x_j - \overline{x_j})(x_k - \overline{x_k})(x_l - \overline{x_l})}{dt} p_{\mathbf{x}} d^n\mathbf{x},\end{aligned} \tag{26}$$

the second line of equation 26 being equal to:

$$-\dot{\overline{x_i}}\,K_{jkl} - \dot{\overline{x_j}}\,K_{ikl} - \dot{\overline{x_k}}\,K_{ijl} - \dot{\overline{x_l}}\,K_{ijk}, \tag{27}$$

while the first line is:

$$\begin{aligned} &-\frac{S\left(t\right)}{2s_\beta}\Bigg[\int\left(x_i-\overline{x_i}\right)\left(x_j-\overline{x_j}\right)\left(x_k-\overline{x_k}\right)\left(x_l-\overline{x_l}\right)\sum_{m=1}^{n}x_m^2\,p_{\mathbf{x}}d^n\mathbf{x} \\ &\qquad\qquad -C_{ijkl}\sum_{m=1}^{n}\left(V_{x,m}+\overline{x_m}^2\right)\Bigg] \\ &+U\left(1-f\right)\int\int\left(x_i-\overline{x_i}\right)\left(x_j-\overline{x_j}\right)\left(x_k-\overline{x_k}\right)\left(x_l-\overline{x_l}\right)\,p_{\mathbf{x}-\mathbf{u}}\rho\left(\mathbf{u}\right)d^n\mathbf{u}\,d^n\mathbf{x} \\ &-U\left(1-f\right)C_{ijkl}\,. \end{aligned} \tag{28}$$

The first integral is computed as before, and yields:

$$C_{ijklmm} + 2\overline{x_m}\,C_{ijklm} - C_{ijkl}\,V_{x,m}\,. \tag{29}$$

The moments C_{ijklmm}, C_{ijklm} and C_{ijkl} can be expressed in terms of second-order moments and third / fourth-order cumulants using equations 9, 10 and 11 (assuming $K_{ijklm}=0$, $K_{ijklmm}=0$). From this, one obtains that the change in K_{ijkl} is given by:

$$\begin{aligned} \dot{K}_{ijkl} = -\frac{S\left(t\right)}{s_\beta}\sum_{m=1}^{n}\big(&K_{ijkm}\,C_{lm}+K_{ijlm}\,C_{km}+K_{iklm}\,C_{jm}+K_{jklm}\,C_{im} \\ &+K_{ijm}\,K_{klm}+K_{ikm}\,K_{jlm}+K_{ilm}\,K_{jkm}\big) \\ &+\left(\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{jl}+\delta_{il}\delta_{kj}\right)U\left(1-f\right)\lambda^2\,. \end{aligned} \tag{30}$$

2.5. Simplification for isotropic mutation and selection

If the initial distribution of phenotypes $\mathbf{x}$ is a multivariate normal $\mathcal{N}(0, I_n\sqrt{V_x})$ with I_n the identity matrix of size n, we have initially independence of all phenotypic trait x_i. This implies that only cumulants of the same phenotypic trait (eg. C_{iiii}) are non-zero. If these cumulants are initially zero, one can check with their dynamical equations that they will remain zero throughout the dynamics because selection and mutation do not generate anisotropy. Overall, this means that the dynamics of the distribution of phenotypes $\mathbf{x}$ only depends on the dynamics of x_i, K_{ii}, K_{iii}, K_{iiii}. We thus simplify our notations using, for all phenotypic dimensions i :

$$\overline{x} = \overline{x_i} \tag{31}$$

$$V_x = K_2 = V_{x,i} = K_{ii} \tag{32}$$

$$K_3 = K_{iii} \tag{33}$$

$$K_4 = K_{iiii} \tag{34}$$

Finally we can write the dynamical system for the cumulants as:

$$\dot{\overline{x}} = -\frac{S}{2s_\beta}\left(K_3 + 2V_x\,\overline{x}\right) \tag{35}$$

$$\dot{V_x} = -\frac{S}{2s_\beta}\left(K_4 + 2\overline{x}\,K_3 + 2V_x^2\right) + U\left(1-f\right)\lambda \tag{36}$$

$$\dot{K_3} = -\frac{S}{s_\beta}\left(\overline{x}\,K_4 + 3K_3\,V_x\right) \tag{37}$$

$$\dot{K_4} = -\frac{S}{s_\beta}\left(4K_4\,V_x + 3K_3^2\right) + 3U\left(1-f\right)\lambda^2\,. \tag{38}$$

After some time the system reaches an equilibrium where $\overline{x} = K_3 = 0$ and $S\left(t\right) = S_{eq}$. The equilibrium values of V_x and K_4 depend on $U\left(1-f\right)\lambda$ (mutation) and on $\frac{S_{eq}}{s_\beta}$ (selection).

With the isotropy in phenotypic traits, we can also rewrite the expression of the mean transmission rate as:

$$\overline{\beta} = \beta_0 - \frac{nV_x}{2s_\beta} - \frac{n\overline{x}}{2s_\beta} \tag{39}$$

The equilibrium values of V_x and K_4 must satisfy:

$$V_x = \sqrt{\frac{U\left(1-f\right)\lambda}{A} - K_4/2} \tag{40}$$

$$K_4 = \frac{3U\left(1-f\right)\lambda^2}{4V_x A} \tag{41}$$

$$A = \frac{S_{eq}}{s_\beta} \tag{42}$$

We can express the equilibrium density of susceptible cells as a function of the equilibrium phenotypic variance, yielding:

$$A = \frac{2(d+\alpha+Uf)}{2\beta_0 s_\beta - nV_x} \tag{43}$$

First, we can work in the WSSM regime thus neglecting terms in λ^2. In this case we have $K_4 = 0$ and we get the following expression for equilibrium variance:

$$V_x^{WSSM}(\infty) = \frac{-X + \sqrt{X\left(X + 16\frac{\beta_0 s_\beta}{n}\right)}}{4} \tag{44}$$

with:

$$X = \frac{nU(1-f_L)\lambda}{d+\alpha+Uf} \tag{45}$$

We can also relax the WSSM approximation and consider higher order mutational effects directly with equation (40). However, we cannot get a simple expression for the equilibrium variance in this case, but we can study the predictions numerically (Figure in Main text). Interestingly, when λ or selection are very high the above dynamical system does not reach a stable equilibrium with a positive V_x. A necessary condition for a positive V_x to exist is $V_x > \frac{3}{8}\lambda$. In other words, the standing phenotypic variance has to be substantially higher than the mutational variance. This yields threshold values on λ below which our resolution does not work and the distribution of the phenotypic variation cannot be described by the first 4 moments.

On the other side of the spectrum, there are regimes where mutations are rare but with larger effects. For this case, the 'House of Cards' approximation has been developed by Turelli (Turelli, 1984), based on the assumption that the effect of a mutation is independent on the original phenotype in which it appeared. This approximation, applied to our model with demographic feedback, leads to a formula for the equilibrium variance:

$$V_x^{HC}(\infty) = \frac{2\,U(1-f)\,s_\beta}{n\,S(\infty)} \tag{46}$$

which simplifies to:

$$V_x^{HC}(\infty) = \frac{2\,\beta_0\,(1-f)\,s_\beta\,U}{n\,(\alpha+d+U)} \tag{47}$$

The main difference with our previous results is that the equilibrium variance is independent of the mutational variance λ. We use our simulation model to compare the different expressions for equilibrium variance in the different regimes. We see that the cumulant approach up to K4 provides the best fit when s_β is high ie. when selection is weak. In case of extremely low selection, this approximation becomes equivalent to the WSSM regime because the term in λ^2 vanishes. However, the fourth cumulant approximation crumbles with higher selection where it yields negative values for the variance. In this case, the House of Cards approximation yields the best result. Indeed in this regime, the distribution of phenotypes is more peaked around the optimum and the distribution is very far from Gaussian, which is the basis of the WSSM approximation.

2.6. Critical Mutation Rates

The critical mutation rate is the value above which the pathogen population goes extinct because the mutation load is too high (i.e., the transmission rate is too low). In this case the equilibrium density of susceptible cells is:

$$S_{eq} = \frac{d + \alpha + Uf}{\overline{\beta}} = \frac{b}{d} \tag{48}$$

With:

$$S_{eq} = \frac{d + \alpha + Uf}{\overline{\beta}}\overline{\beta} = \beta_0 - \frac{nV_x}{2s_\beta} \tag{49}$$

The critical value of the phenotypic variance is:

$$V_x^c = \frac{2s_\beta}{n}\left(\beta_0 - \frac{(d + \alpha + Uf)d}{b\beta_0}\right) \tag{50}$$

Then one can use 44 to derive the critical value of the mutation rate U_c. We find that when $f > 0$:

$$U_c^{WSSM} = \frac{A + B - \sqrt{A(A + 2B)}}{8f^2 s_\beta d} \tag{51}$$

with:

$$A = (1 - f)\lambda\,\beta_0\,b\,n^2 \tag{52}$$

$$B = 8f\,s_\beta(\beta_0^2\,b - d(d + \alpha)) \tag{53}$$

In the special case when $f = 0$:

$$U_c^{WSSM} = \frac{4s_\beta(\beta_0\,b - d(d + \alpha))^2}{\beta_0\,b\,d\,\lambda\,n^2} \tag{54}$$

In the same way we can compute the critical mutation rate In the House of Cards regime which gives:

$$U_c^{HC} = \frac{b\beta_0 - d(d + \alpha)}{d} \tag{55}$$

Strinkingly, this threshold of critical mutation rate is not dependent on other parameters affecting the effect of mutations λ, s_β, n and f. The critical mtuation rate only depends on demographic parameters. This means that in this regime, mutations that we model as 'non-lethal' have such strong effects that they are *de facto* lethal when computing a critical mutation rate. In fact, this critical mutation rate is exactly the one predicted in the WSSM regime with equation (51) when all mutations are lethal i.e. $f = 1$.

We cannot get an explicit expression for the critical mutation rate in the model where we consider more cumulants. We can however get these values numerically and we compare the three models in figure S3.

3. Evolutionary dynamics using a partial differential equation on the CGF of the distribution of transmission rates

In this appendix we derive the evolutionary dynamics by following directly the distribution of transmission rates $\beta_{\mathbf{x}}$ instead of the phenotypes $\mathbf{x}$. To do so we follow the dynamics of the Cumulant Generating Function of the distribution of transmission rate using the method from Martin&Roques (Martin and Roques, 2016). This type of function is useful as it directly gives access to the cumulants of the distribution such as the mean and the variance of the distribution of transmission rates. It could also be used to easily explore other mutational regimes as one would just need to plug the Moment Generating Function of another distribution of mutational effects to get the dynamics. In the following, we build a partial derivative equation (PDE) on this Cumulant Generating Function. Finally we use a Gaussian approximation for the phenotypes, and a WSSM approximation to linearize the PDE and obtain a system of ordinary differential equations describing the evolutionary dynamics.

To elucidate the dynamics of the distribution of transmission rates we use particular generating functions of the distribution. We define:

$$M_t(z) = \int I_{\mathbf{x}} e^{\beta_{\mathbf{x}} z} d^n\mathbf{x} \tag{56}$$

$$C_t(z) = \log(M_t(z)) \tag{57}$$

where $M_t(z)$ and $C_t(z)$ are respectively the density Moment Generating Function (dMGF) and density Cumulant Generating Function (dCGF). The term "density" refers to the use of $I_{\mathbf{x}}$ instead of $p_{\mathbf{x}}$ in these definitions. We use these functions instead of 'regular' generating functions because we are interested in following the density of infected cells, and not just the distribution of phenotypes. A consequence of the use of of dMGF and dCGF is that:

$$M_t(0) = I \tag{58}$$

$$C_t(0) = \log(I) \tag{59}$$

In the following we refer to the partial derivatives according to t (resp. z) with ∂_t (resp. ∂_z). Taking the first derivative with respect to z and setting z gives access to the mean transmission rate:

$$\partial_z M_t(z) = \int I_{\mathbf{x}} \beta_{\mathbf{x}} e^{\beta_{\mathbf{x}} z} d^n\mathbf{x} \tag{60}$$

$$\partial_z M_t(0) = \int I_{\mathbf{x}} \beta_{\mathbf{x}} d^n\mathbf{x} = I\overline{\beta} \tag{61}$$

$$\partial_z C_t(z) = \frac{\int I_{\mathbf{x}} \beta_{\mathbf{x}} e^{\beta_{\mathbf{x}} z} d^n\mathbf{x}}{\int I_{\mathbf{x}} e^{\beta_{\mathbf{x}} z} d^n\mathbf{x}} = \frac{\partial_z M_t(z)}{M_t(z)} \tag{62}$$

$$\partial_z C_t(0) = \frac{\int I_{\mathbf{x}} \beta_{\mathbf{x}} d^n\mathbf{x}}{I} = \int p_{\mathbf{x}} \beta_{\mathbf{x}} d^n\mathbf{x} = \overline{\beta} \tag{63}$$

Going further, one can check that the nth derivative of $C_t(z)$ with respect to z, evaluated in $z = 0$ yields the nth cumulant of the distribution of β. Now we can describe the dynamics of the distribution of transmission rates, neglecting the non-lethal mutations, with a partial derivative equation on the dMGF or dCGF of this distribution:

$$\partial_t M_t(z) = \int \partial_t(I_{\mathbf{x}})\, e^{\beta_{\mathbf{x}} z} d^n\mathbf{x} = S \int \beta_{\mathbf{x}} I_{\mathbf{x}} e^{\beta_{\mathbf{x}} z} - (\alpha + d + Uf) \int I_{\mathbf{x}} e^{\beta_{\mathbf{x}} z} \tag{64}$$

$$= S\partial_z M_t(z) - (\alpha + d + Uf) M_t(z) \tag{65}$$

$$\partial_t C_t(z) = \frac{\partial_t M_t(t)}{M_t(z)} = S\frac{\partial_z M_t(z)}{M_t(z)} - (\alpha + d + Uf) = S\,\partial_z C_t(z) - (\alpha + d + Uf) \tag{66}$$

We note that setting $z = 0$ in the above PDE yields exactly the dynamics of the density of infected cells as presented in the epidemiological model. Yet this PDE can be used to go deeper in the description of the dynamics of the phenotypic distributions of the viral population. In the next section we implement the mutational scheme of non-lethal mutations by studying the effect of mutations on the dCGF $C_t(z)$

3.1. Adding non-lethal mutations

A non-lethal mutation of effect s in an infected cell of transmission rate $\beta_{\mathbf{x}}$ has the following effect on the dMGF of the transmission rate:

$$\underset{mut}{\Delta} M_t(z|(s, \beta_{\mathbf{x}})) = I\,U(1-f)\,\Delta t(e^{(\beta_{\mathbf{x}}+s)\,z} - e^{\beta_{\mathbf{x}}\,z}) = I\,U(1-f)\,e^{\beta_{\mathbf{x}}\,z}(e^{s\,z} - 1) \tag{67}$$

Taking expectations over the distribution of mutational effects s in background $\beta_{\mathbf{x}}$:

$$\underset{mut}{\Delta} M_t(z|\beta_{\mathbf{x}}) = \int \underset{mut}{\Delta} M_t(z|(s, \beta_{\mathbf{x}}))\, f(s|\beta_{\mathbf{x}})\, ds = I\,U(1-f)\,\Delta t\, e^{\beta_{\mathbf{x}}\,z}(M^s(z, \beta_{\mathbf{x}}) - 1) \tag{68}$$

with $f(s|\beta_{\mathbf{x}})$ the distribution of mutational effects in background $\beta_{\mathbf{x}}$, and $M^s(z, \beta_{\mathbf{x}})$ the MGF of the distribution of mutational effects in background $\beta_{\mathbf{x}}$. Then taking expectations over all phenotypes $\mathbf{x}$:

$$\underset{mut}{\Delta} M_t(z) = \int p_{\mathbf{x}} \underset{mut}{\Delta} M_t(z|\beta_{\mathbf{x}})\, d^n\mathbf{x} = I\, U(1-f)\Delta t \int p_{\mathbf{x}} e^{\beta_{\mathbf{x}}\, z}(M^s(z, \beta_{\mathbf{x}}) - 1) d^n\mathbf{x} \quad (69)$$

$$= I\, U(1-f)\Delta t(\overline{e^{\beta_{\mathbf{x}}\, z} M^s(z, \beta_{\mathbf{x}})} - \overline{e^{\beta_{\mathbf{x}} z}}) \quad (70)$$

Where the overbar refers to the average over all values of $\beta_{\mathbf{x}}$ at time t. In continuous time, as $\Delta t \to 0$, we use the fact that $\underset{mut}{\Delta} C_t(z) = \underset{mut}{\Delta} M_t(z)/M_t(z)$ to obtain :

$$\underset{mut}{\Delta} C_t(z) = U(1-f)\Delta t(\frac{\overline{e^{\beta_{\mathbf{x}}\, z} M^s(z, \beta)}}{\overline{e^{\beta_{\mathbf{x}}\, z}}} - 1) \quad (71)$$

To go further with the expression, we need to express the MGF of the distribution of mutational effects $M^s(z, \beta)$. Note that it is not dependent on the number of infected, and so this is a classic MGF and not a density MGF. In the next two sections we derive an expression in the cases where the phenotype dimensions are $n = 1$ or $n > 1$. We already note that the two expressions will be consistent.

3.2. MGF of the DFE in one dimension ($n = 1$)

A mutation of effects s on β, can be expressed with the phenotype x in which it happens and effect u of the mutation of the phenotype as follows:

$$(s|x) = \beta_{x+u} - \beta_x = -\frac{2\, x\, u + u^2}{2 s_\beta} \quad (72)$$

When $u \sim \mathcal{N}(0, \sqrt{\lambda})$, we can express the distribution of (s|x) using a noncentral Chi-squared distribution dependent on the distance to the optimum (Martin2014):

$$(s|x) \sim s_0(x) - \frac{\lambda}{2 s_\beta} \chi^2_1\Big(\frac{2\, s_\beta\, (\beta_0 - \beta_x)}{\lambda}\Big) \quad (73)$$

The MGF of this distribution (s|g) is:

$$M_s(z, x) = \Big(1 + \frac{z\lambda}{s_\beta}\Big)^{-\frac{1}{2}} Exp\Big(\frac{z^2\lambda(\beta_0 - \beta_x)}{s_\beta + z\lambda}\Big) = M_*(z)\, e^{-\omega(z)\, \beta_x} \quad (74)$$

3.3. MGF of the DFE in higher dimension ($n \geq 2$)

Similarly to the case $n = 1$, A mutation of effects s on $\beta_{\mathbf{x}}$, can be expressed with an mutational effect $\mathbf{u}$ on the phenotype $\mathbf{x}$ in which it happens as follows:

$$(s|\beta_{\mathbf{x}}) = \beta(\mathbf{x} + \mathbf{u}) - \beta(\mathbf{x}) = -\frac{2\, \mathbf{x}.\mathbf{u} + \mathbf{u}.\mathbf{u}}{2 s_\beta} \quad (75)$$

Using polar coordinates for the mutation $\mathbf{u}$, let $r = \|\mathbf{u}\|$ and $\theta = cos(\widehat{(\mathbf{x}, \mathbf{u})})$.

$$\begin{cases} \mathbf{u}.\mathbf{u} = r^2 \\ \mathbf{u}.\mathbf{x} = r\ \|\mathbf{x}\|\ \theta \end{cases} \tag{76}$$

We want to compute the MGF of the distribution of fitness effects s which is

$$M_s(z, \|\mathbf{x}\|) = \mathbb{E}\left(e^{sz}\right) = \mathbb{E}_{r,\theta}\left[e^{z\left(-\frac{r^2}{2s_\beta} - \frac{r\,\theta\,\|\mathbf{x}\|}{s_\beta}\right)}\right] = \mathbb{E}_r\left[\mathbb{E}_\theta\left[e^{z\left(-\frac{r^2}{2s_\beta} - \frac{r\,\theta\,\|\mathbf{x}\|}{s_\beta}\right)}\right]\right] \tag{77}$$

As $\mathbf{u} \sim \mathcal{N}(0, I_n\sqrt{\lambda})$ we then have that $r^2 \sim \lambda\chi_n^2 = \Gamma\left(\frac{n}{2}, 2\lambda\right)$ and using the distribution of angles to one optimum in a n-sphere from (Martin and Lenormand, 2015) we get

$$M_s(z, \mathbf{x}) = \left(\frac{\lambda z}{s_\beta} + 1\right)^{-\frac{n}{2}} Exp\Big(\frac{\lambda z^2 \|\mathbf{x}\|^2}{2s_\beta^2 + 2\lambda s_\beta z}\Big) \tag{78}$$

Finally by using the dependence of $\beta_{\mathbf{x}}$ on $\|\mathbf{x}\|$ we get

$$M_s(z, \mathbf{x}) = \Big(1 + \frac{z\lambda}{s_\beta}\Big)^{-\frac{n}{2}} Exp\Big(\frac{z^2\lambda(\beta_0 - \beta_{\mathbf{x}})}{s_\beta + z\lambda}\Big) = M_*(z)\, e^{-\omega(z)\,\beta_{\mathbf{x}}} \tag{79}$$

with $M_*(z) = \left(\frac{\lambda z}{s_\beta} + 1\right)^{-\frac{n}{2}} e^{\frac{\lambda z^2 \beta_0}{s_\beta + \lambda z}}$ is the MGF of the DFE in a phenotype at the optimum ($\mathbf{x} = 0$) and $\omega(z) = \frac{\lambda\, z^2}{s_\beta + \lambda\, z}$ is the linear effect of β on the CGF of the DFE (Martin and Roques, 2016).

3.4. An alternative way to derive the WSSM approximation

The MGF of the distribution of fitness effects $M_s(z)$ computed above can be plugged in (71) which yields:

$$\frac{\underset{mut}{\Delta}\, C_t(z)}{\Delta t} = U(1-f)\left(M^*(z)\,\frac{\overline{e^{\beta_{\mathbf{x}}(z-\omega(z))}}}{\overline{e^{\beta_{\mathbf{x}}\, z}}} - 1\right) = U(1-f)\left(M^*(z)\, e^{C_t(z-\omega(z)) - C_t(z)} - 1\right) \tag{80}$$

This expression is only dependent on the CGF of the distribution of transmission rates β and thus we obtain a closed system. Adding this mutation term to the PDE describing selection (66) yields:

$$\partial_t C_t(z) = S\,\partial_z C_t(z) - (\alpha + d + U\,f) + U(1-f)\left(M^*(z)\, e^{C_t(z-\omega(z)) - C_t(z)} - 1\right) \tag{81}$$

This PDE is not solvable analytically, but by differentiating it according to z and setting $z = 0$, we can access the dynamics of the cumulants of the distribution of transmission rates $\beta_{\mathbf{x}}$.

$$\partial_t C_t(0) = \partial_t \log\big(I(t)\big) = -(\alpha + d) + \overline{\beta} S(t) \tag{82}$$

$$\partial_t \partial_z C_t(0) = \partial_t \overline{\beta} = -\frac{U(1-f)\,\lambda}{2 s_\beta} + S(t) V_\beta \tag{83}$$

$$\partial_t \partial_z^2 C_t(0) = \partial_t V_\beta = \frac{U(1-f)\,\lambda\big(3\lambda + 8 s_\beta(\beta_0 - \overline{\beta})\big)}{(2 s_\beta)^2} + S(t)\,\partial_z^3 C_t(0) \tag{84}$$

However, we get that the dynamics of the cumulant of order n is dependent on the value of the cumulant of order $n+1$. In contrast to section 2, we focus here on the distribution of transmission rates $\beta_{\mathbf{x}}$ and not on the distribution of phenotypes $\mathbf{x}$. The distribution of transmission rate is not Gaussian, and so we cannot neglect cumulants of high order to close the dynamics. Thus we cannot get a closed system of differential equations without further assumptions.

To simplify the resolution of the PDE, we first make the assumption that the distribution of phenotypes $\mathbf{x}$ is Gaussian. With this assumption, we can compute the general form of $C_t(z)$ the dCGF of the distribution of transmission rates. Knowing the form of the dCGF and plugging it into the PDE (81), we directly compute the partial derivatives according to z and t to obtain an ordinary differential equation. We find the form of the dCGF of transmission rates much in the same way as we did the MGF of mutational effects. Indeed, mutation is gaussian and centered around the phenotype in which it appears, and in parallel the distribution of phenotypes is gaussian and centered around the mean phenotype. We study $\delta\beta = \beta(\bar{x} + u) - \beta(\bar{x})$ where $u \sim \mathcal{N}\left(0, I_n \sqrt{V_x}\right)$ with $V_x(t)$ the variance in each phenotypic trait. Thus with the same derivations as earlier, we get:

$$M_t^{\delta\beta}(z) = \Big(1 + \frac{z V_x(t)}{s_\beta}\Big)^{-\frac{n}{2}} Exp\Big(\frac{z^2 V_x(t)(\beta_0 - \beta_{\mathbf{x}})}{s_\beta + z V_x(t)}\Big) = M_*(z)\, e^{-\omega(z)\,\beta_{\mathbf{x}}} \tag{85}$$

From which follows:

$$\begin{aligned} M_t(z) &= \int I_{\mathbf{x}} e^{\beta_{\mathbf{x}} z} d^n\mathbf{x} = \int I_{\mathbf{x}} e^{(\beta_{\overline{x}} + u) z} du \\ &= I\big(e^{\beta_{\overline{x}} z} + M_t^{\delta\beta}(z)\big) \end{aligned} \tag{86}$$

where $\overline{x} = \int p_{\mathbf{x}} g d^n\mathbf{x}$ is the mean phenotype and thus $\beta_{\overline{x}}$ is the transmission rate of the mean phenotype (note that $\beta_{\overline{x}} \neq \overline{\beta}$). Finally, the dCGF on the distribution of transmission rates $\beta_{\mathbf{x}}$, under the assumption that phenotypes $\mathbf{x}$ are Gaussian, is:

$$C_t(z) = Log(M_t(z)) = \frac{z\,(s_\beta\,\beta_{\overline{x}}(t) + z\,\beta_0\,V_x(t))}{z\,V_x(t) + s_\beta} - \frac{n}{2}\log\left(1 + \frac{z\,V_x(t)}{s_\beta}\right) + \log(I) \tag{87}$$

The range of validity of this approximation is discussed by Martin&Roques (Martin and Roques, 2016) and boils down in our model to a lower bound on the mutation rate relative to the strength of mutation and selection:

$$U \gg \frac{n^2 \lambda}{4 s_\beta} \tag{88}$$

The Gaussian approximation can only be valid when mutations have small effects. Thus we linearize the PDE (81) using a Taylor Series of λ at the first order which gives:

$$\partial_t C_t(z) = S\,\partial_z C_t(z) - (\alpha + d + U f) + \frac{\mu^2\, z^2 \left(\beta_0 - \partial_z C_t(z)\right) - z\,\mu^2 \frac{n}{2}}{s_\beta} \tag{89}$$

This dCGF is dependent on time through parameters linked to the distribution of the underlying phenotypic traits: the transmission rate of the mean phenotype $\beta_{\overline{x}}$ and the phenotypic variance V_x. Finally, plugging the form (87) in (89) and solving for all z and z^2 finally gives:

$$\dot{\beta}_{\overline{x}}(t) = \frac{2\,V_x\,S}{s_\beta}\big(\beta_0 - \beta(\overline{x})\big) \tag{90}$$

$$\dot{V_x}(t) = -\frac{V_x^2\,S}{s_\beta} + U(1-f)\lambda \tag{91}$$

$$\tag{92}$$

Instead of the dynamics of the transmission rate of the mean phenotype $\beta_{\overline{x}}$, we prefer to follow directly the dynamics of the mean phenotype $\overline{x}$ like in appendix A. From the expression of the dynamics of the transmission rate of the mean phenotype $\beta_{\overline{x}}(t)$, the dynamics of the mean phenotype is easily recovered and follows:

$$\dot{\overline{x}}(t) = -\frac{V_x(t)\,S(t)}{s_\beta}\overline{x}(t) \tag{93}$$

which is consistent with the dynamics recovered in section 2.

4. Numerical simulations

To check our analytic results, we compare them with phenotype-centered simulations in two dimensions $n = 2$. We build a grid of size $l * l$ with l an odd number, where each square of the grid corresponds to a phenotype. The phenotypic trait value goes from $-x_{max}$ to x_{max} and the phenotype step between each square is $\delta_x = \frac{2x_{max}}{l}$ as shown in

Figure S1. To each square (i, j) of the grid is associated at a given time t a number of infected individuals $I(i, j, t)$ of phenotype (x_i, x_j).

At each time interval dt, susceptible cells $S(t)$ grow with a constant rate b and die with rate d. Infected cells $I(i, j, t)$ grow by infecting susceptible cells $S(t)$ depending on the transmission rate of their phenotype $\beta(x_i, x_j)$, and die with $d + \alpha + Uf$, such that

$$\begin{aligned} I(i,j,t+dt) &= I(i,j,t) + dtI(i,j,t)\big(S(t)\beta(x_i,x_j) - (d+\alpha+Uf)\big) \\ S(t+dt) &= S(t) + dt(b - S(t)(d + \sum_{i,j} I(i,j,t)\beta(x_i,x_j))) \\ \beta(i,j) &= \beta_0 - \frac{x_i^2 + x_j^2}{2s_\beta} \end{aligned} \tag{94}$$

To stick with our analytic model we use a quadratic link function between phenotypes and transmission rate that could lead to negative values of β. To add the effect of mutations, we use two different model. The first model M_g features Gaussian mutation. It is the exact transposition of the integro-differential equation describing the dynamics of the distribution of phenotypes on a grid, such that by mutation:

$$I(i,j,t+dt) = I(i,j,t) + dt\, U\,(1-f)\Big(\big(\sum_{k,m} I(i-k, j-m, t)\rho(k)\rho(m)\big) - I(i,j,t)\Big) \tag{95}$$

where ρ is the Gaussian kernel $\mathcal{N}(0, \sqrt{\lambda})$. We make sure to normalize the kernel so that the sum of probabilities of the mutations from the center square of the grid is equal to 1.

The second model M_d models mutation as a diffusion in the phenotypic space and represents the WSSM regime. Starting from the integro-differential equation (6), we can take a Taylor expansion around $I_{\mathbf{x}-\mathbf{u}}(t)$ in small $\mathbf{u}$. The incoming mutation term in the integral then equals to:

$$\int \big(I_{\mathbf{x}}(t)\rho(\mathbf{u}) - \mathbf{u}\rho(\mathbf{u})\nabla I_{\mathbf{x}}(t) + \mathbf{u}.\mathbf{u}\rho(\mathbf{u})\nabla^2 I_{\mathbf{x}}(t) + \mathbf{u}.\mathbf{u}.\mathbf{u}\rho(\mathbf{u})\nabla^3 I_{\mathbf{x}}(t) + o(\mathbf{u}.\mathbf{u}.\mathbf{u})\big) d^n\mathbf{u} \tag{96}$$

Where $\nabla I_{\mathbf{x}}(t)$ is the gradient operator of $I_{\mathbf{x}}(t)$ according to phenotype $\mathbf{x}$. As the mutation kernel is symetric, the second and fourth term vanishes. Higher order terms in $\mathbf{u}$ are neglected as they would lead to terms in higher orders of λ. The integral yields:

$$I_{\mathbf{x}}(t) + \frac{\lambda}{2}\nabla^2 I_{\mathbf{x}}(t) + o(\lambda) \tag{97}$$

The first term cancels out with the outgoing mutation term in (6) and we finally get a diffusive form of the integro-differential equation:

$$\frac{\partial I_{\mathbf{x}}(t)}{\partial t} = \beta_{\mathbf{x}} S I_{\mathbf{x}} - (d + \alpha + Uf) I_{\mathbf{x}} + U(1-f)\frac{\lambda}{2}\nabla^2 I_{\mathbf{x}}(t) \tag{98}$$

Modelling this equation with discrete phenotypes on the phenotype grid coupled with the epidemiological dynamics yields:

$$I(i,j,t+dt) = I(i,j,t) + dt\, U(1-f)\frac{\lambda}{2\delta_x^2}\,\Delta_I \tag{99}$$

$$\Delta_I = I(i+1,j,t) + I(i-1,j,t) + I(i,j+1,t) + I(i,j-1,t) - 4I(i,j,t) \tag{100}$$

We compare these simulations to the analytic models we use in the main text in Figure**??**. We see that the diffusion model M_d squarely fits the WSSM analytic predictions for the two mutation rates considered. However, the cumulant model taking into account larger mutational effects (λ and λ^2) with K3 and K4 shows a faster adaptation of the mean transmission rate and a reduced equilibrium mutational load. The Gaussian mutation model M_g (considering all orders of λ) yields an even faster adaptation and even smaller equilibrium mutational load.

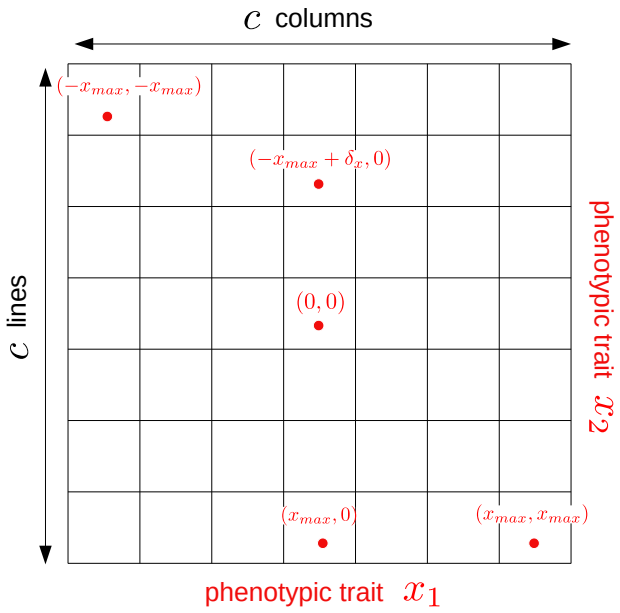


Figure S1: Grid of phenotypes for numeric simulations. Example phenotype values are represented in red on the grid.

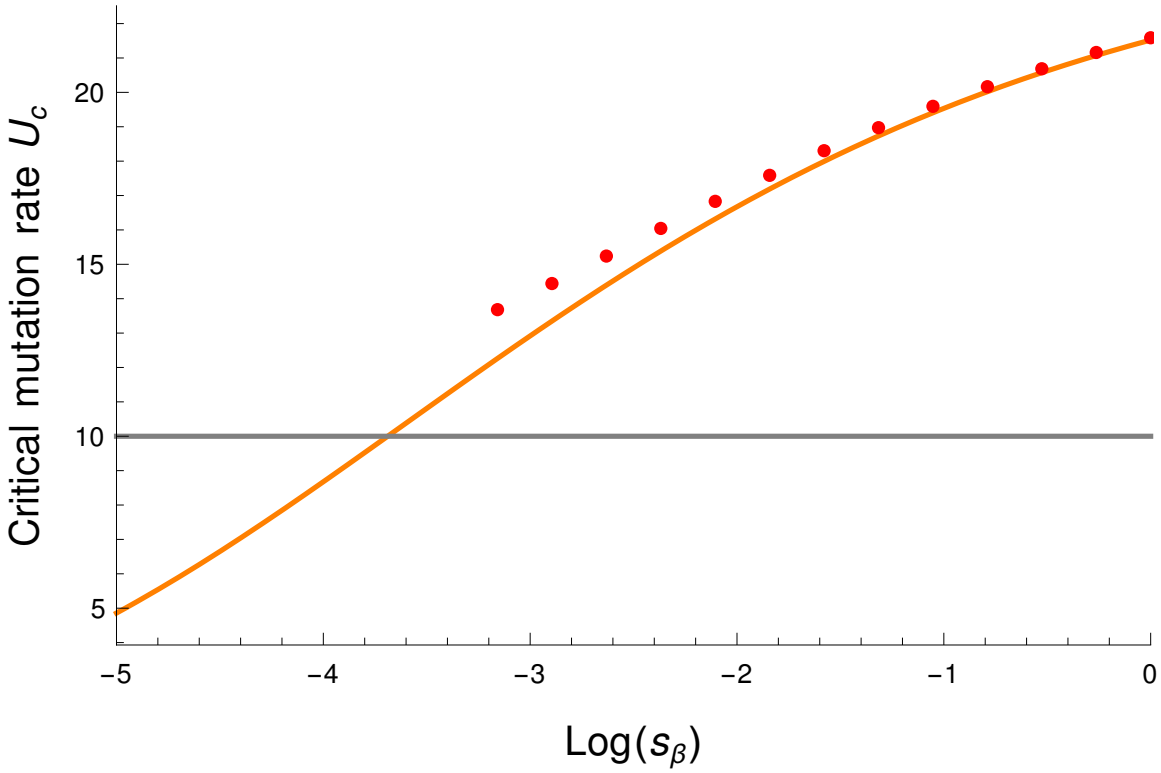


Figure S2: Comparison of the critical mutation rate for lethal mutagenesis under different regimes. The critical mutation rate U_c is shown as a function of the selection parameter s_β in a log-log scale. The results are shown for three derivations: the WSSM approximation in orange, the results from the cumulant approach up to K4 in red, and the House of Cards (HC) approximation in grey. Data from numerical simulations are shown as blue points. The parameters used were: $b = 3, d = 1, \beta_0 = 4, \alpha = 1, U = 4, f = 0.4, \lambda = 0.05, n = 2$.

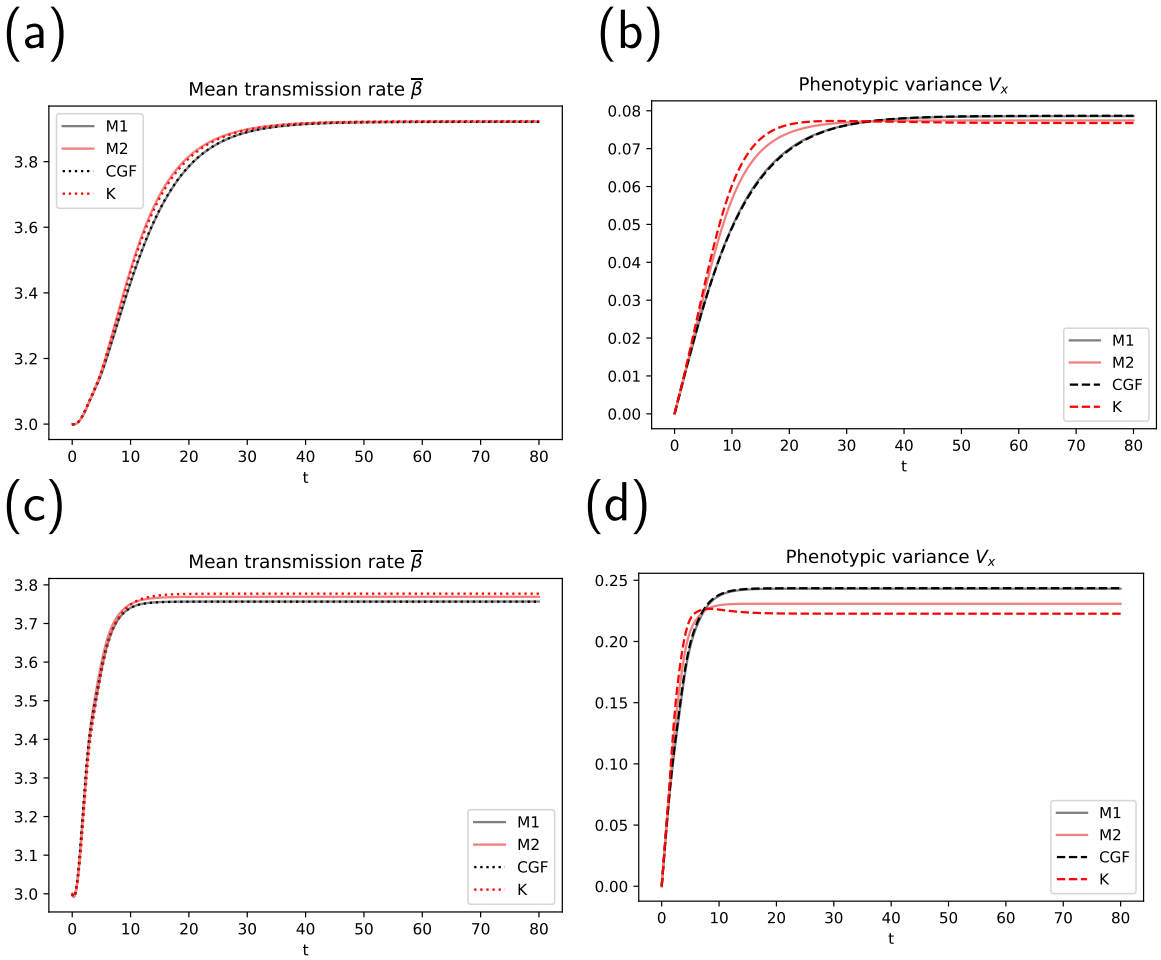


Figure S3: Comparison of the analytic en numeric predictions for (a,c) the mean transmission rate and (b,d) the phenotypic variance. The analytic models are depicted with dashed lines, the numeric models with solid lines. Model M_d (diffusion) is shown in grey, model M_g (Gaussian mutation) in light red, the CGF model in black and the cumulant model in red. Panels (a,b): $\lambda = 0.005$; Panels (c,d): $\lambda = 0.05$ The other parameters common two all panels were: $b = 4, d = 1, \beta_0 = 4, \alpha = 2, s_\beta = 1, U = 1, f = 0.4, n = 5$

Chapter 2:

Building pyramids against the evolutionary emergence of pathogens

Submitted to Proceedings of the Royal Society B
July 2023 version

Building pyramids against the evolutionary emergence of pathogens

Sylvain Gandon*[1], Martin Guillemet[1], François Gatchitch[1], Antoine Nicot[1],

Ariane C. Renaud[2,3], Denise M. Tremblay[3,4], and Sylvain Moineau[2,3,4]

1: CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

2: Département de biochimie, de microbiologie et de bio-informatique, Faculté des sciences et de génie, Université Laval, Québec, Canada, G1V 0A6.

3: Groupe de recherche en écologie buccale, Faculté de médecine dentaire, Université Laval, Québec City, G1V 0A6, Canada

4: Félix d'Hérelle Reference Center for Bacterial Viruses, Université Laval, Québec City, G1V 0A6, Canada

* Correspondence: sylvain.gandon@cefe.cnrs.fr

Abstract

Mutations allowing pathogens to escape host immunity promote the spread of infectious diseases in heterogeneous host populations and can lead to major epidemics. Understanding the conditions that slow down this evolution is key for the development of durable control strategies against pathogens. Several earlier theoretical studies have explored the efficacy of different deployment strategies of host resistance across space and time in agriculture. Because these studies focus on diverse optimization criteria, they sometimes yielded contrasting recommendations regarding their relative efficacy. Besides, these analyses are limited by the paucity of experimental data on the durability of proposed control schemes. Here we use theory and experiments to compare the efficacy of three resistance strategies: (i) a *mixing* strategy where the host population contains two single-resistant genotypes, (ii) a *pyramiding* strategy where the host carries a double-resistant genotype, (iii) a *combining* strategy where host population is a mix of a single-resistant genotype and a double-resistant genotype. First, we use evolutionary epidemiology theory to clarify the interplay between demographic stochasticity and evolutionary dynamics to show that the *pyramiding* strategy always yields lower probability of evolutionary emergence. We also show that when only single-resistance genotypes are available, we should maximize diversity of resistance to minimize the risk of evolutionary emergence. Second, we tested experimentally these predictions by using virulent bacteriophages introduced into bacterial populations where we manipulated the

diversity and the depth of immunity using a CRISPR-Cas system. We showed that pyramiding multiple defenses into the same host genotype and avoiding combination with single-defense genotypes is a robust way to reduce pathogen evolutionary emergence. These results have practical implications for the optimal deployment of host resistance in agriculture and biotechnology but also for vaccination against pathogens.

Keywords: evolutionary emergence, epidemiology, demographic stochasticity, host resistance, CRISPR, infectious disease control.

Introduction

The spread of pathogen epidemics is driven by the composition of host populations and, in particular, by the fraction f_R of resistant hosts. Larger values of f_R generate "herd immunity" in well-mixed populations because a randomly chosen susceptible host is expected to be surrounded by many resistant neighbors. Since resistant hosts cannot be successfully infected, their presence shields susceptible individuals from the risk of being infected and reduce the spread of a given pathogen. In fact, the epidemic is expected to stop growing when $f_R > 1 - 1/R_0$, where R_0 refers to the basic reproduction ratio of the pathogen. This theoretical framework provides key guidelines for the deployment of control measures like vaccination [1,2] or the deployment of resistant varieties of crops in agriculture [3,4].

The efficacy of these control strategies, however, is challenged by the potential acquisition of escape mutations allowing the pathogen to infect a resistant host. Whether those variants will appear, establish, and spread depend on multiple evolutionary forces, including the composition of the host population. For instance, a larger fraction f_R of resistant hosts is expected to reduce the growth rate of the wild-type pathogen and, consequently, to limit the influx of escape mutations. But a larger fraction f_R of resistant hosts is also expected to increase the fitness benefit associated with an escape mutation. This will increase the probability of establishment of a given mutation (i.e. lower risk of stochastic extinction when rare) and it will also increase the speed at which this variant will spread. The balance between these two opposite effects may thus result in a higher risk of pathogen adaptation for intermediate frequency of resistance. Hence, a better understanding of the influence of the host population composition on pathogen adaptation may help identify more durable control strategies.

Many theoretical studies have explored complex ecological scenarios to evaluate the impact of various strategies for the deployment of host resistance genotypes across space

and time in agriculture [3,5–14]. In particular, several studies contrasted the efficacy of *mixing* multiple single host resistant genotypes with the efficacy of *pyramiding* multiple resistant genes within a single genotype. Earlier models did not incorporate demographic feedbacks or any influence of demographic stochasticity and focused on the long-term deterministic evolutionary outcomes. Under these conditions the *mixing* strategy can outperform the *pyramiding* strategy because the former strategy can prevent the spread of pathogens carrying multiple escape mutations [5,6]. More recent studies challenged this guidance and relied on more realistic simulation models to highlight the importance of epidemiology, demographic stochasticity, and spatial structure on both the epidemiology and the evolution of the pathogen [9–11,14,15]. Taken together, the available theoretical literature may appear confusing because distinct studies make different recommendations on the optimal strategy for the deployment of resistance against a pathogen [16]. This confusion stems from the different assumptions of the models (e.g. with or without demography, with or without stochasticity) but also on the different optimality criteria used to identify the most effective pathogen control strategies (e.g. no evolution of multi-escape mutations, minimal disease incidence) [8,17,18]. Besides, experimental studies needed to evaluate the durability of control strategies are notoriously difficult to carry out in most pathosystems, and in particular in agriculture [8,13,19]. This lack of experimental studies does not help to elucidate the pros and cons of distinct deployment of resistance strategies.

Here we develop a joint theoretical and experimental approach to analyse the durability of different strategies for the deployment of host resistance. We focus on a very specific quantity to evaluate the efficacy of a control strategy: the probability of pathogen emergence with (or without) viral adaptation. This quantity provides a relevant measure of control efficacy because it accounts for both short-term (epidemiology) and long-term (evolution) dynamical processes [20]. Experimental measurements of evolutionary emergence, however, are challenging because the stochastic nature of pathogen extinction requires a large number of replicate populations to measure the probability of emergence. These experiments require also the ability to manipulate the composition of the host population and to track the evolution of the pathogen population. These hurdles can be overcome by studying the evolutionary emergence of virulent bacteriophages in bacterial populations that use the adaptive CRISPR immunity to prevent phage infections: (i) many replicates can be carried out simultaneously using bacteria and phages in 96-well plates [20]; (ii) CRISPR immunity provides a very convenient way to manipulate both the *diversity* of host immunity (different bacteria derived from the same population can carry different “spacers” in their CRISPR array [20,21]) and the *depth* of host immunity (multiple “spacers” can be stacked within the CRISPR array of the same multiresistant bacterium [22]); (iii) the mechanism of phage adaptation to CRISPR-based immunity is well documented: virulent

phages escape CRISPR resistance through mutation in their target sequence (the "protospacer") [23–26].

In the next sections we present the theoretical framework used to compute the probability of evolutionary emergence of pathogens after being introduced in a heterogeneous host population. We use this model to understand the effect of multiple factors on the fate of the pathogen: (i) the number of viruses introduced, (ii) the proportion of resistant hosts, (iii) the diversity and the depth of immunity of resistant hosts. This allows us to contrast the influence of different strategies of resistance deployment on the probability of pathogen evolutionary emergence. In a second step, we manipulated the heterogeneity of bacterial CRISPR immunity to test the validity of our theoretical predictions on the evolutionary emergence of phage populations.

Materials and Methods

Theory

Pathogen emergence is defined as the ability to escape early extinction and thus to initiate an epidemic after the introduction a small quantity of pathogens in the host population. This probability of emergence depends both on the host (e.g. proportion and diversity of resistant hosts) and the pathogen (e.g. inoculum size, genetic composition) [20,27–29]. For a pathogen to emerge, we assume that the host population contains a fraction $\left(1-f_R\right)$ of individuals fully susceptible to the pathogen while the remaining fraction f_R of the population is resistant. Among the resistant hosts, we consider three alternative scenarios (**Figure 1**):

(i) a *mixing* scenario, in which the resistant fraction of the population is a mix of two single-resistance genotypes (A+B) aiming at distinct pathogen *target sites* (a target site is defined here as a region of the pathogen genome recognized by immune effectors and where a mutation or a deletion may allow escape recognition by host immunity). We allow the frequency of the two resistant hosts to vary and f_A refers to the frequency of the resistant host A among the resistant hosts ($f_B=1-f_A$ is the frequency of the resistant host B among the resistant hosts);

(ii) a *pyramiding* scenario, in which the resistant fraction of the population is monomorphic with a double-resistance genotype (AB);

(iii) a *combining* scenario, in which the resistant fraction of the population results from a mix of single-resistance genotypes (say A) and a double-resistance genotype (AB). We allow the frequency of the two resistant hosts to vary and f_A refers to the frequency of resistant host A among the resistant hosts ($f_{AB}=1-f_A$ is the frequency of the resistant host AB

among the resistant hosts). Note that the *pyramiding* scenario is a limit case of the *combining* scenario when $f_A = 0$.

The efficacy of resistance is assumed to be complete (no infection if the host is resistant) but the pathogen can evade recognition by acquiring escape mutations in the corresponding immunity target sites. Therefore, a pathogen with escape mutation i can infect a fraction $(1-f_R)+f_R P_i$ of the total host population, where $P_i = \sum_{h \in \{A,B,AB\}} f_h P_i^h$ is the fraction of the resistant hosts that can be infected by the pathogen with genotype i since P_i^h measures the ability of the pathogen genotype i to infect of the host genotype $h \in \{\varnothing, A, B, AB\}$. In the *pyramiding* scenario, pathogens with less than 2 escape mutations can only infect a fraction $(1-f_R)$ of the total host population and only pathogens with 2 escape mutations can infect all the hosts. In the *mixing* and in the *combining* scenario, the fitness of a single escape mutant depends on the identity of the single-resistance genotype in the host population and the composition of the resistant population (see **supplementary information**).

We further assume a classic birth-death process to model the epidemiological dynamics where a host infected with a pathogen that does not carry escape mutations spreads this pathogen in a fully susceptible host population at rate b and dies at rate d. Host resistance prevents infection and may thus affect the effective birth rate, but without affecting d. Whereas escape mutations may allow the pathogen to infect a larger fraction of the host population, they also carry a fitness cost c which causes pathogens with i escape mutations to reproduce at rate $b_i = b(1-c)^i$. The probability of acquiring an escape genotype $i \in \{A, B, AB\}$ by mutation is noted μ_i and may vary among target sites. Crucially, the rate of acquisition of 2 escape mutations is expected to be much smaller than the rate of single escape mutations: $\mu_{AB} \approx \mu_A \mu_B \ll \mu_A, \mu_B$. For the sake of simplicity, we assume that escape mutations are fixed and cannot revert to the ancestral types. These reversions to the wild-type target are expected to have a negligible effect on the probability of evolutionary emergence when the target site mutation rate remains small [20,30].

We detailed in the **supplementary information** how we compute the probability of emergence (with or without viral evolution) after the introduction of an inoculum of V phage particles in a heterogeneous bacterial host population. We also derive approximations for the probability of evolutionary emergence inspired from models of evolutionary rescue. Those approximations help to contrast the effects of the composition of the host population on the risk of evolutionary emergence.

Experiments

We used the Gram-positive bacterial strain *Streptococcus thermophilus* DGCC 7710 which is susceptible to the virulent phage 2972. We also used three CRISPR-resistant clones (also referred as bacteriophage-insensitive mutants: BIMs) that were derived from *S. thermophilus* DGCC 7710 and differ only in their CRISPR arrays (**Tables S1** and **S2**). Two of these clones carried a single additional spacer (strains A and B) targeting the genome of phage 2972, while the remaining clone carried a combination of these two spacers (strain AB) precisely obtained using the approach developed by Hynes et al. [22]. The addition of a single spacer in the CRISPR1 array of *S. thermophilus* DGCC 7710 provides a robust resistance against infection by the wild-type virulent phage 2972 [23–25] (**Table S2**). The rate at which phage 2972 acquires mutations allowing to escape CRISPR immunity was found to be approximately equal to $2.8*10^{-7}$ mutations/locus/replication [31]. The acquisition of a single escape mutation may or may not yield significant fitness costs for the phage [24,31].

We monitored the dynamics of the phage population after introducing an inoculum of V viruses in each well of a 96-well plate containing 200 μL of replicate bacterial populations with a proportion $f_R = 90\%$ of resistant cells and $1-f_R = 10\%$ of susceptible cells. This virus inoculum was sampled from a lysate obtained after amplifying a single plaque of the wild-type phage 2972 on *S. thermophilus* DGCC 7710 (the initial frequency of single and double escape mutants was estimated in **Table S3**). We manipulated the composition of the resistant bacterial population to produce three experimental treatments to test the predictions of the theoretical model (**Figure 1**): (i) *mixing* (strains A and B in equal frequency), (ii) *pyramiding* (only strain AB), (iii) *combining* strains A and AB in equal frequency (combining A) or *combining* strains B and AB in equal frequency (combining B). After an overnight incubation (22 h) we quantified the abundance and the evolution of the phage after spotting a fraction of each replicate ($2\,\mu L$) on a lawn of the different bacterial strains to measure: (i) the presence/absence of phages using a lawn of susceptible cells (ii) the presence/absence of escape mutations in the phage population using lawns of single-resistance bacteria (A or B) and a lawn of double-resistance bacteria (AB) [20,31].

We used logistic regression models with the presence/absence on susceptible bacteria (or on resistant bacteria) as the response variable as a function of the inoculum size and the composition of the host population (see **supplementary information**).

Results

Emergence and evolutionary emergence

We derive the probability of pathogen emergence after the introduction of an inoculum of V pathogens. This inoculum is sampled in a population where some phage genotypes may

already carry escape mutations: p_i refers to the frequency of genotype $i \in \{\varnothing, A, B, AB\}$. In the following we focus mainly on scenarios where the frequencies of preexisting escape mutations remain low (i.e. $p_{\varnothing} \approx 1$). **Figure 2** shows the effect of the inoculum size and the frequency of resistance on pathogen emergence under different deployment strategies.

In the absence of pathogen mutations ($p_{\varnothing}=1$ and $\mu=0$) the probability of pathogen emergence is equal to $P_E=1-(f_R+1/R_0)^V$ when $f_R+1/R_0>1$, where $R_0=b/d$ is the basic reproduction ratio of the pathogen [20]. As indicated with a dashed line in **Figure 2**, this probability of pathogen emergence drops rapidly with the increase in the proportion of resistant hosts and pathogen emergence becomes impossible when $f_R>1-1/R_0$. Note that this threshold is independent of the deployment strategy because they all share the same value of f_R.

However, the pathogen population may avoid extinction through the acquisition of escape mutations. The term *evolutionary emergence* refers to these situations where emergence is consecutive to pathogen evolution [28,29]. In **Figure 2** we compare the probabilities of evolutionary emergence in a symmetric scenario where $f_A=1/2$ for increasing values of f_R (**Figure 2A**) and V (**Figure 2B**). Crucially, the probability that the pathogen adapts to host resistance depends on the deployment of host resistance strategies and the *pyramiding* treatment always yields lower probability P_{EE} of evolutionary emergence. Indeed, in both the *mixing* and the *combining* treatments, the presence of a single-resistance genotype provides a "stepping stone" allowing the acquisition of a first escape mutation allowing the virus to recover the ability to grow in the host population. Besides, the acquisition of this first escape mutation may allow the pathogen to acquire later on the ability to escape both types of resistance. The lower probability to acquire both escape mutations at the same time explains the step-like shape of the probability of emergence in **Figure 2** (see also **Figure S1**). As expected, preexisting mutations always increase the probability of pathogen emergence and allow the pathogen population to escape extinction even in the extreme case where $f_R=1$ and no fully susceptible hosts are present in the host population (**Figure 2A**).

We can generalize these results for asymmetric scenarios where $f_A \neq 1/2$. Interestingly, variations of f_A have different effects in the mixing and combining treatments (**Figure 3**). In the *mixing* treatment, the probability of emergence is minimized when f_A is close to ½ and thus when the amount of diversity is maximized in line with the effect of diversity discussed in Chabas et al [20]. In the *combining* treatment, the risk of emergence is

minimized when $f_A=0$ because this is the case where all the resistant hosts carry two resistances (i.e. *pyramiding* treatment).

The influence of host composition on the probability of evolutionary emergence can be captured within the framework of evolutionary rescue models. This framework is relevant as soon as $f_R>1-1/R_0$ because the wild-type virus is doomed to go extinct when the proportion f_R of resistant hosts leads to a negative growth rate of the wild-type virus population . We derive approximations for the probability of evolutionary emergence under the assumption that the viral mutation rate is small (**supplementary information**). In the symmetric scenario (i.e. $f_A=1/2$) this yields:

$$\begin{aligned}
&\text{Mixing: } P_{EE}^{M}\approx 2V\mu\left(1-\frac{d}{b(1-c)(1-f_R/2)}\right)+O(\mu^2)\\
&\text{Combining: } P_{EE}^{C}\approx V\mu\left(1-\frac{d}{b(1-c)(1-f_R/2)}\right)+O(\mu^2)\\
&\text{Pyramiding: } P_{EE}^{P}\approx O(\mu^2)
\end{aligned} \qquad (1)$$

This approximation captures both the effect of a larger inoculum size and the effect of treatment on P_{EE} illustrated in **Figure 2**. Note that larger inoculum sizes are also expected to increase P_{EE} via the introduction of pre-existing mutants, not modelled in (1). The above approximation is particularly useful to discuss the effect of the composition of the host population. In particular, in the *mixing* strategy the P_{EE} is expected to be twice larger than in the *combining* strategy in the symmetric scenario. And both these strategies are expected to have higher P_{EE} than the *pyramiding* strategy because μ_{AB} is assumed to be much smaller than μ_A and μ_B.

Experiments

Increasing the size V of the virus inoculum increased the ability to observe the presence of phages on fully susceptible bacterial populations (Type II Anova: LR Chi-square =3744.2, df=1, P<2.2×10^{-16}) and reached its maximal value when $V>10^3$. We found an effect of host treatment on the probability to detect phages on fully susceptible bacteria which is difficult to interpret because it interacts with the inoculum size (**supplementary information**). Importantly, note that this treatment effect is not due to pathogen evolution since pathogen evolution is not detectable when $V<10^3$.

It is tempting to equate our measure of the presence of phages on susceptible bacteria with the probability of emergence P_E. Yet, as soon as the wild-type phages start to replicate, the proportion of susceptible bacteria is expected to drop and f_R is expected to be ≈ 1 after the overnight culture. So, the presence/absence of phage on susceptible bacteria may actually result from the detection of some of the phages that have been inoculated but did not adsorb to a host cell yet. In the following, we prefer to focus on the analysis of the presence/absence of phage able to replicate on different types of resistant hosts (i.e. host A, B or AB) because it provides an unambiguous measure of the probability of pathogen adaptation to host immunity.

Our analysis of the probability of the phage to adapt to at least one type of resistance confirms our predictions on the effect of inoculum and host composition (**Figures 4 and 5**). In particular, we recover the predicted relationship $P_{EE}^{M} > P_{EE}^{C} > P_{EE}^{P}$ when we focused on the *Combining B* treatment (i.e. a combination of strains B and AB): Tukey HSD test, $P_{EE}^{M} - P_{EE}^{CB}$ =0.78, z=3.08, P=0.011 ; $P_{EE}^{CB} - P_{EE}^{P}$=2.67, z=9.40, P<0.001. However, we find no significant differences between the probabilities of viral evolution in the *Mixing* and in the *Combining A* treatments (i.e. a combination of strains A and AB): Tukey HSD test, $P_{EE}^{M} - P_{EE}^{CA}$=0.12, z=0.49, P=0.96. This suggests that the probability for a virus of acquiring an escape mutation against resistance A is higher than against resistance B. Note that the expected twofold increase in the probability of viral evolution in the *Combining* treatment relative to the *Mixing* treatment (see equation (1)) lies in the 95% confidence intervals we compute: P_{EE}^{M} / P_{EE}^{CA} = 1.13 [0.60;2.13] ; P_{EE}^{M} / P_{EE}^{CA} = 2.17 [1.14;4.15] (**Figure 5B**, red dashed line).

Interestingly, similar treatment effects were found when we analysed the ability of phage 2972 to acquire both escape mutations (**Figure S4**). In particular, we found that the *Mixing* treatment was most favourable for the emergence of double escape mutations (Tukey HSD test, $P_{EE_2}^{M} - P_{EE_2}^{CA}$=1.75, z=6.71, P<0.001 ; $P_{EE_2}^{M} - P_{EE_2}^{CB}$=0.84, z=3.39, P=0.0039; $P_{EE_2}^{M} - P_{EE_2}^{P}$ =1.78, z=6.81, P<0.001), even if none of the bacteria carry both resistance in this treatment. This effect likely results from the sequential acquisition of multiple mutations, which is facilitated in the mixing treatment. In other words, the *mixing* strategy is far less durable than the *pyramiding* strategy. Besides, as predicted by our theoretical model, the probability of evolutionary emergence under the *combining* strategy falls in between the two other strategies and confirms that the presence of single-resistant genotypes speeds up the acquisition of escape mutations and promotes evolutionary emergence even when some hosts are multiresistant.

Discussion

In this study, we have explored the influence of several factors such as pathogen life history traits (birth and death rates), mutation rates, pathogen initial inoculum size, fraction and depth of host resistance, on the ultimate fate of a pathogen introduced in a heterogeneous host population. In particular, we showed that larger inoculum size favors the emergence and the adaptation of the pathogen to the host population because of two main effects. First, larger inoculum size increases the probability of the introduction of a preexisting escape mutation which further increases the evolutionary potential of the pathogen population. Second, even in the absence of preexisting mutations in the inoculum, a larger inoculum size of the wild-type pathogen provides more opportunities for the emergence of escape genotypes by mutation.

Our theoretical analysis yielded clear predictions on the effect of the host composition on the probability of evolutionary emergence of a pathogen: *pyramiding* is the most effective way to reduce the risk of pathogen adaptation, even in the presence of preexisting escape mutant in the pathogen inoculum (**Figure 2**). The worst strategy is the fully asymmetric *mixing* strategy (e.g. $f_A=1$) because it takes a single escape mutant to exploit the whole host population. The fully symmetric *mixing* strategy is better than the asymmetric *mixing* strategy because, as shown by previous studies, higher host diversity reduces the probability of evolutionary emergence [20]. The efficacy of the *combining* treatment is intermediate and is very sensitive to the relative proportion of single and multiple resistances. In particular, we showed that the overlap between the resistance genes carried by single- and double-resistant host genotypes in the *combining* treatment may greatly enhance the risk of evolutionary emergence because escaping single-resistance may provide a “stepping stone” towards the acquisition of multiple escape mutations.

Our experimental results confirmed both the positive effect of larger inoculum size and the hierarchy in the efficacy of different host treatments on the probability of pathogen adaptation. Note, however, that our model oversimplifies several features of the pathogen dynamics taking place in our experiments. First, we modeled viral growth as a “birth-death” process while the reproduction of our virulent phage follows a “burst-death” cycle. The burst-death process is expected to alter the variance associated with the reproduction event and may thus alter the predictions on the evolutionary outcome [32], but see [20] for a comparison between these two ways to model pathogen dynamics. Second, we assumed the fraction of the different host genotypes to be constant throughout the experiment. This is a very rough approximation because we know that the fraction of the susceptible hosts will drop relatively rapidly when the wild-type genotype spreads. Consequently, the fraction f_R of resistant hosts is expected to increase rapidly through time. Similarly, the relative fraction of

the different types of hosts is expected to vary with time after the emergence of single escape mutations that will exploit specifically a fraction of these resistant hosts. Yet, the good match between our theoretical predictions and our experimental results suggests that the conclusions of our model are robust to the specific details of the epidemiology of the pathogen.

Our conclusions are also consistent with a review of the available empirical studies on the durability of different crop protection programs which concluded that *pyramiding* is the most durable strategy [8]. For instance, the durability of wheat cultivars was associated with the pyramiding of multiple resistant genes [7]. A few experimental studies have tracked the evolution of the pathogens over several generations and demonstrated the beneficial impact of the *pyramiding* strategy. A study on the evolution of plant-parasitic nematode showed that the use of pyramided genotypes protected the plant-crop over several years [33]. Another study on transgenic broccoli plants indicated that the expression of multiple *Bacillus thuringiensis* (Bt) toxins hampered the epidemiology and evolution of a major insect pest, the diamondback moth (*Plutella xylostella*) [34–36]. In addition, this latter study revealed the detrimental effect of combining the same resistance genes in different plant varieties for the durability of resistance. Indeed, as in our *combining* strategy, the advantage of using plant genotypes containing two dissimilar Bt toxin genes for resistance management may be compromised if they share similar toxins with single-gene plants that are deployed simultaneously.

Conclusions and broader implications

Microbes carrying CRISPR-Cas immunity against virulent bacteriophages provide ideal biological models to obtain experimental measures of the probability of pathogen emergence and evaluate their ability to escape host resistance under different control strategies [20]. Besides, the specificity of CRISPR-Cas immunity to bacteriophages is very similar to the classical gene-for-gene model of specificity driving the coevolution between many plants and their pathogens. Our biological experiments confirm our theoretical predictions on the influence of (i) the resistance strategy and (ii) the initial dose of the pathogen. In particular, we find that the *pyramiding* strategy is a more effective way to reduce the evolutionary emergence of the pathogen. These microbiological assays confirm that exposing pathogens to a mix of different host genotypes carrying a low number of resistance genes facilitate the adaptation of the pathogen because it provides multiple routes (with slower slopes) towards complex pathogen genotypes carrying multiple escape mutations. This result does not conflict with the positive effect of host diversity for the reduction of pathogen evolutionary emergence [20,37]. But for a given amount of host resistance diversity, the present study shows that stacking this diversity in a limited number of genotypes is a more effective

strategy than using a mixture of single-resistance host genotypes to prevent pathogen emergence. The success of the pyramiding strategy may explain why many microbes carry several genes coding for distinct defence systems in their genome [38,39]. While these genomic defence islands may provide immunity against a wide variety of diverse phages, they may also limit the emergence and evolution of bacteriophage variants, thereby increasing the persistence of microbes in various ecosystems.

While our results are relevant for several areas, including for crop management in agriculture [8,33,40] as well as in food fermentation [41], they may also hold for the management of drugs and vaccines. In AIDS, for instance, the success of the combination therapy is arguably due to the use of the *pyramiding* strategy where the patient is treated simultaneously with multiple drugs [42–44]. A similar conclusion was reached with a theoretical model that explored alternative treatment strategies against bacteria as a combination therapy (*pyramiding*) outperforms other ways to use available antibiotics [45–47]. In malaria, the use of artemisinin-containing combination therapies (*pyramiding*) is also believed to provide a way to slow the spread of antimalarial drug resistance [48–51]. These results suggest also that the use of phage cocktails in phage therapy is likely to be more effective because the pathogenic bacteria will have difficulties to evolve resistance against multiple phages [52,53].

The durability of vaccines may also be explained by the *pyramiding* effect. Unlike therapeutic drugs, some vaccines often elicit multiple immune responses against several pathogen targets and this could explain why resistance to vaccines evolves usually more slowly than drug resistance [54]. The recognition of the value of immune diversity could lead to new vaccination strategies. For example, the deployment of different vaccines among different individuals to create a mosaic of vaccination has the potential to outperform conventional vaccination [55]. Moreover, several studies demonstrated that combining multiple immune response to different epitopes can increase significantly the efficacy of vaccination [56–60]. The rise of mRNA vaccines [61,62] may facilitate the development of such new generation of multivalent vaccines that could use the pyramiding effect to increase their durability.

Pyramiding multiple defenses in the same host may thus provide a durable strategy for both prophylactic and therapeutic control of infectious diseases. The recognition of the value of *pyramiding* is ancient [63] but we hope our work clarifies the complex interplay between demography, evolution, and stochasticity. The influence of many other factors remains to be investigated. For instance, our model does not account for the change in the composition of the host population after the start of a viral epidemic. Time-inhomogeneous branching process models could be developed to better understand the influence of these epidemiological feedbacks on evolutionary emergence. In addition, the importance of

epistasis in fitness among the escape mutations carried by the pathogen is expected to affect the probability of emergence (see **Figure S1**). Patterns of epistasis are also likely to have an impact on the influence of pathogen recombination [64,65]. Our model does not allow for coinfections and, consequently, does not allow for recombination. The influence of genetic recombination on the robustness of the *pyramiding* effect remains to be investigated. Finally, our joint theoretical and experimental study could be extended to explore a wider range of deployment strategies in space and time [44,66–69]. This approach could thus be used to identify and to test the durability of new strategies to limit the emergence and the evolution of pathogens.

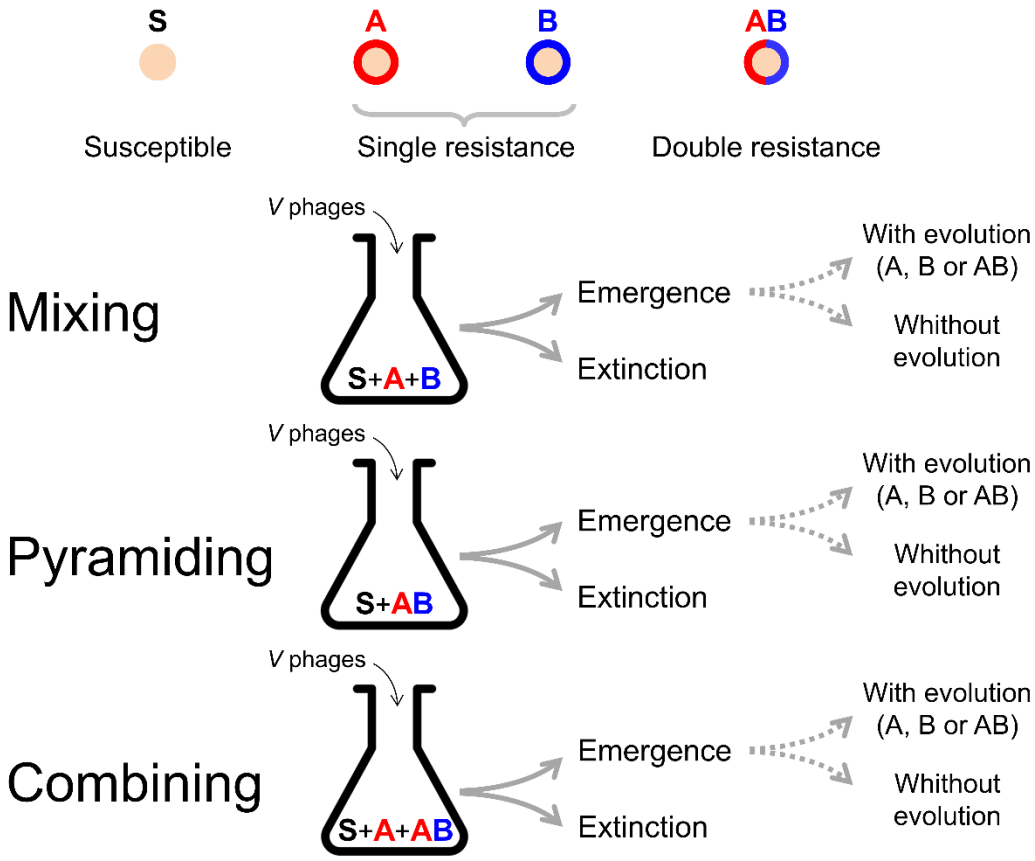


Figure 1: Schematic representation of the *mixing*, *pyramiding,* and *combining* scenarios. In each scenario the host population is a mix of a proportion $(1-f_R)$ of susceptible bacteria (S) and a proportion f_R of resistant bacteria (A, B, and AB). In our experiment we used $f_R=0.9$. The composition of the population of resistant bacteria differs between the *mixing* (1:1 mix of two single-resistant hosts A+B), the *combining* (1:1 mix of a single-resistant host, A (combining A) or B (combining B), and a double-resistant host AB), and the *pyramiding* (a double-resistant host AB) scenarios. After the inoculation of V phages the viral population may either go extinct or produce an epidemic. The virus epidemic may either result from the replication of the ancestral virus (no evolution) or in the additional replication of phage genotypes carrying escape mutations against A, B, or AB. We carried out these experiments in 96-well plates that allowed us to replicate our inoculation experiment in 96 host populations (each replicate population was 200 μL). After an overnight incubation the cultures (22 h) we measured (i) the occurrence of phage epidemics (i.e., emergence) by plating a fraction (2 μL) of each replicate population on a lawn of sensitive cells (*S. thermophilus* DGCC 7710) and (ii) the presence of escape phage mutants (i.e., evolutionary emergence) by plating a fraction (2 μL) of each replicate population on a lawn of singly resistant (A or B) or doubly resistant (AB) bacteria (see **Supplementary Information**).

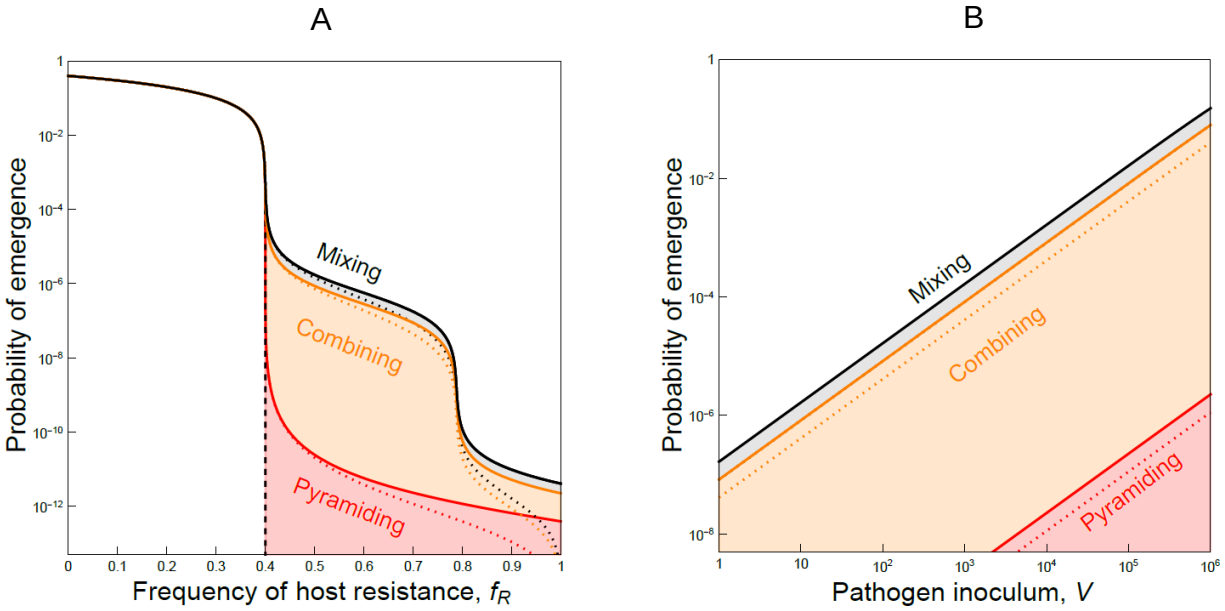


Figure 2: Pathogen emergence under the *mixing* (black), *combining* (orange), and *pyramiding* (red) scenarios. In (A) we show the effect of f_R, the fraction of resistant hosts in the host population, on the probability of pathogen emergence after the introduction of a single virus ($V=1$) with or without preexisting mutations (full or dotted lines, respectively). The probability of emergence in the absence of pathogen evolution is indicated with the dashed black line. The color shading refers to *evolutionary emergence* in the different scenarios (i.e., pathogen emergence resulting from pathogen evolution). In (B) we show the effect of V, the size of the virus inoculum, on the probability of pathogen emergence with or without preexisting mutations when $f_R=0.7$. Other parameter values: $b=1.66$, $d=1$, $c=0.01$, $\mu=10^{-6}$, $p_\varnothing=1-p_A-p_B-p_{AB}$, $p_A=p_B=10^{-6}$, $p_{AB}=10^{-12}$ (without preexisting mutations: $p_\varnothing=1$, $p_A=p_B=p_{AB}=0$).

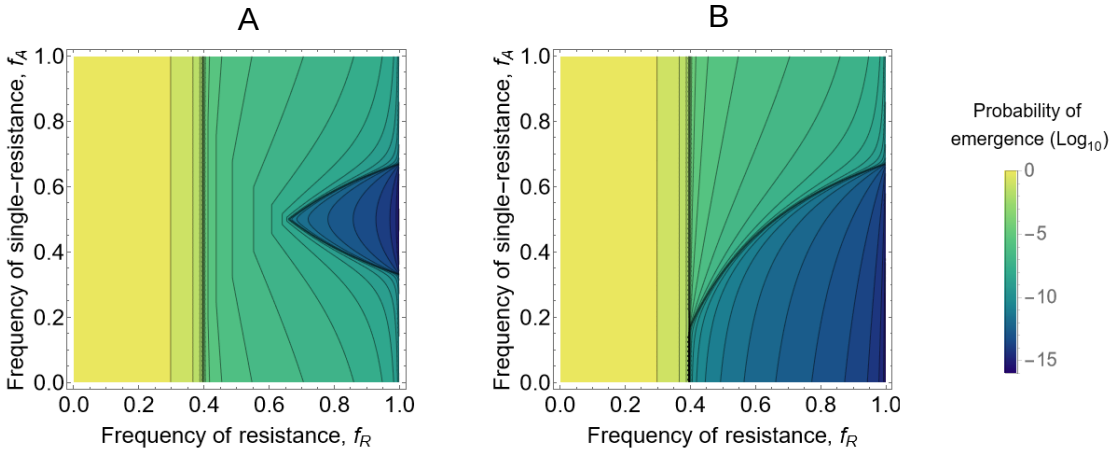


Figure 3: Pathogen emergence under (A) the *mixing* and (B) the *combining* scenarios. We show the effect of f_R, the fraction of resistant hosts in the host population, and f_A, the fraction of a single-resistant host in the resistant host population on the probability of pathogen emergence after the introduction of a single virus ($V=1$) and without preexisting mutations ($p_\varnothing=1$, $p_A=p_B=p_{AB}=0$). Other parameter values: $b=1.66$, $d=1$, $c=0.1$, $\mu=10^{-6}$.

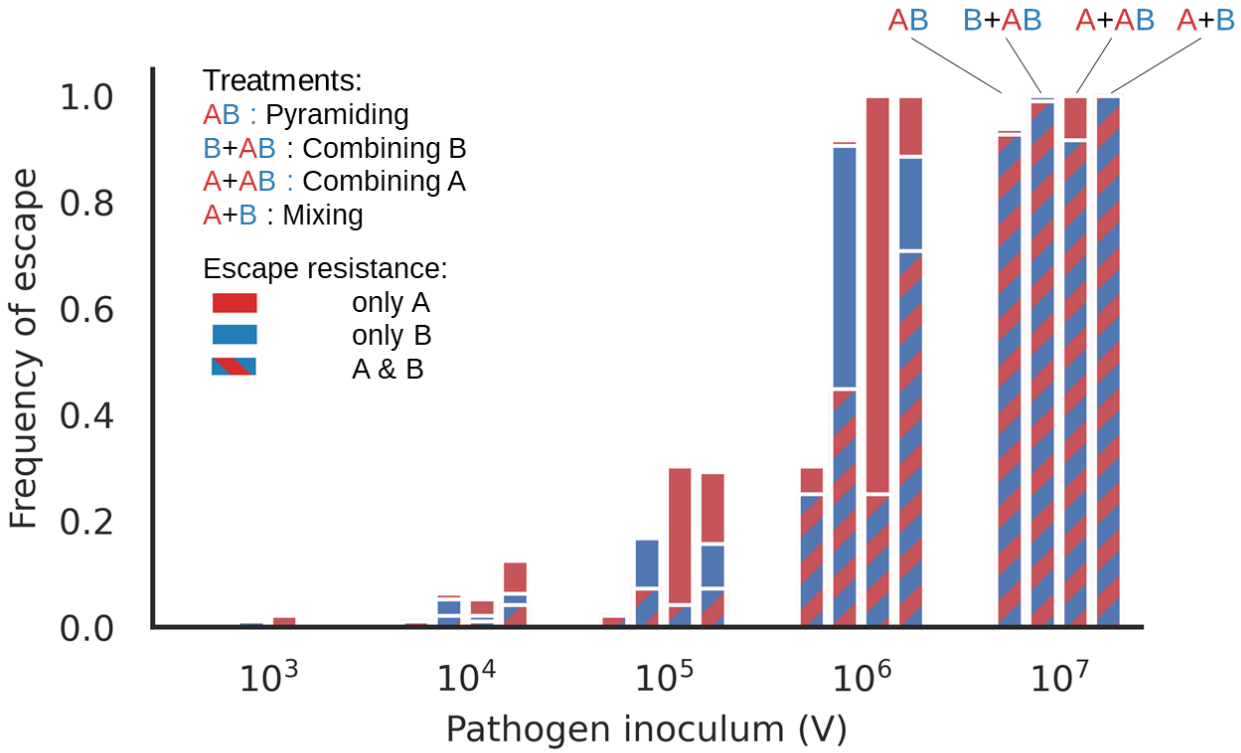


Figure 4: Probability of pathogen evolutionary emergence for different inoculum dose (V) and for different resistance treatments (mixing, combining, and pyramiding). We plot the proportion of populations (among the 96 experimental replicates) that resulted in a virus amplification on different hosts. The colored bars show the frequency of emergence of pathogens which could infect resistant hosts A (red bars) or B (blue bars). The hatched red and blue bars represent the frequency of emergence of double mutant pathogens able to infect both types of resistant hosts. The three types of emergence are always stacked in the same order from the bottom: AB, B, and A.

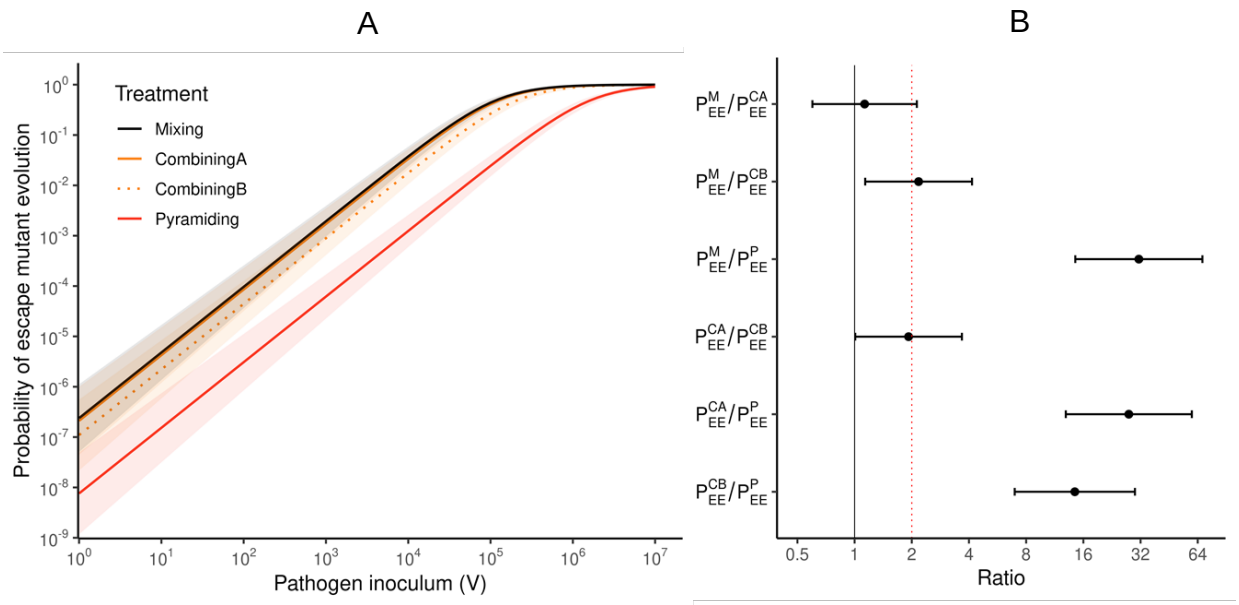


Figure 5: Probability of evolutionary emergence is higher in the *mixing* treatment and lowest in the *pyramiding* treatment. (A) We plot here the estimation of the probability of evolutionary emergence (i.e., the probability to evolve at least one escape resistance) against the inoculum size V and the resistance treatment. The model can be written as $logit(P_{EE}^{T})$ $a_T \log(V)+b_T$, where the slope parameter is the same for all treatments (see **supplementary information**). The lines indicate the prediction of the statistical model for the different treatments and the shaded areas show 95% confidence interval. (B) We compare the estimated values of b_T for all pairs of treatment and we plot $e^{b_{T1}-b_{T2}} \approx P_{EE}^{T1}/P_{EE}^{T2}$. The error bars show 95% confidence interval and the red dashed line refers to a two-fold difference in the probability of emergence. This two-fold effect is expected when we compare the mixing and the combining treatments (see equation (1)).

References:

1. Anderson, R.M., and May, R.M. (1985). Vaccination and herd immunity to infectious diseases. Nature *318*, 323–329. 10.1038/318323a0.

2. Ashby, B., and Best, A. (2021). Herd immunity. Curr Biol *31*, R174–R177. 10.1016/j.cub.2021.01.006.

3. Gilligan, C.A. (2008). Sustainable agriculture and plant diseases: an epidemiological perspective. Philos Trans R Soc Lond B Biol Sci *363*, 741–759. 10.1098/rstb.2007.2181.

4. Gilligan, C.A., and van den Bosch, F. (2008). Epidemiological models for invasion and persistence of pathogens. Annu Rev Phytopathol *46*, 385–418. 10.1146/annurev.phyto.45.062806.094357.

5. Sasaki, A. (2000). Host-parasite coevolution in a multilocus gene-for-gene system. Proceedings of the Royal Society of London. Series B: Biological Sciences. 10.1098/rspb.2000.1267.

6. Segarra, J. (2005). Stable Polymorphisms in a Two-Locus Gene-for-Gene System. Phytopathology *95*, 728–736. 10.1094/PHYTO-95-0728.

7. Brown, J.K.M. (2015). Durable Resistance of Crops to Disease: A Darwinian Perspective. Annual Review of Phytopathology *53*, 513–539. 10.1146/annurev-phyto-102313-045914.

8. Delmotte, F., Bourguet, D., Franck, P., Guillemaud, T., Reboud, X., Vacher, C., and Walker, A.-S. (2016). Combining Selective Pressures to Enhance the Durability of Disease Resistance Genes. Front. Plant Sci. *7*. 10.3389/fpls.2016.01916.

9. Djidjou-Demasse, R., Moury, B., and Fabre, F. (2017). Mosaics often outperform pyramids: insights from a model comparing strategies for the deployment of plant resistance genes against viruses in agricultural landscapes. New Phytol. *216*, 239–253. 10.1111/nph.14701.

10. Rimbaud, L., Papaïx, J., Barrett, L.G., Burdon, J.J., and Thrall, P.H. (2018). Mosaics, mixtures, rotations or pyramiding: What is the optimal strategy to deploy major gene resistance? Evol Appl *11*, 1791–1810. 10.1111/eva.12681.

11. Elisabeth Lof, M., de Vallavieille-Pope, C., and van der Werf, W. (2017). Achieving Durable Resistance Against Plant Diseases: Scenario Analyses with a National-Scale Spatially Explicit Model for a Wind-Dispersed Plant Pathogen. Phytopathology *107*, 580–589. 10.1094/PHYTO-05-16-0207-R.

12. Mikaberidze, A., McDonald, B.A., and Bonhoeffer, S. (2015). Developing smarter host mixtures to control plant disease. Plant Pathology *64*, 996–1004. 10.1111/ppa.12321.

13. Mundt, C.C. (2002). Use of multiline cultivars and cultivar mixtures for disease management. Annu Rev Phytopathol *40*, 381–410. 10.1146/annurev.phyto.40.011402.113723.

14. Sapoukhina, N., Durel, C.-E., and Le Cam, B. (2009). Spatial deployment of gene-for-gene resistance governs evolution and spread of pathogen populations. Theor Ecol *2*, 229. 10.1007/s12080-009-0045-5.

15. Bourget, R., Chaumont, L., and Sapoukhina, N. (2013). Timing of pathogen adaptation to a multicomponent treatment. PLoS ONE *8*, e71926. 10.1371/journal.pone.0071926.

16. Gibson, A.K. (2022). Genetic diversity and disease: The past, present, and future of an old idea. Evolution. 10.1111/evo.14395.

17. van den Bosch, F., and Gilligan, C.A. (2003). Measures of Durability of Resistance. Phytopathology *93*, 616–625. 10.1094/PHYTO.2003.93.5.616.

18. van den Bosch, F., and Gilligan, C.A. (2008). Models of Fungicide Resistance Dynamics. Annual Review of Phytopathology *46*, 123–147. 10.1146/annurev.phyto.011108.135838.

19. Burdon, J.J., Barrett, L.G., Rebetzke, G., and Thrall, P.H. (2014). Guiding deployment of resistance in cereals using evolutionary principles. Evol Appl *7*, 609–624. 10.1111/eva.12175.

20. Chabas, H., Lion, S., Nicot, A., Meaden, S., Houte, S. van, Moineau, S., Wahl, L.M., Westra, E.R., and Gandon, S. (2018). Evolutionary emergence of infectious diseases in heterogeneous host populations. PLOS Biology *16*, e2006738. 10.1371/journal.pbio.2006738.

21. van Houte, S., Ekroth, A.K.E., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S., Boots, M., Paterson, S., Buckling, A., *et al.* (2016). The diversity-generating benefits of a prokaryotic adaptive immune system. Nature *532*, 385–388. 10.1038/nature17436.

22. Hynes, A.P., Labrie, S.J., and Moineau, S. (2016). Programming Native CRISPR Arrays for the Generation of Targeted Immunity. mBio *7*, e00202-16. 10.1128/mBio.00202-16.

23. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712. 10.1126/science.1138140.

24. Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400. 10.1128/JB.01412-07.

25. Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71. 10.1038/nature09523.

26. Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. U.S.A. *108*, 10098–10103. 10.1073/pnas.1104144108.

27. Antia, R., Regoes, R.R., Koella, J.C., and Bergstrom, C.T. (2003). The role of evolution in the emergence of infectious diseases. Nature *426*, 658–661. 10.1038/nature02104.

28. Gandon, S., Hochberg, M.E., Holt, R.D., and Day, T. (2013). What limits the evolutionary emergence of pathogens? Philos. Trans. R. Soc. Lond., B, Biol. Sci. *368*, 20120086. 10.1098/rstb.2012.0086.

29. André, J.-B., and Day, T. (2005). The effect of disease life history on the evolutionary emergence of novel pathogens. Proc. Biol. Sci. *272*, 1949–1956. 10.1098/rspb.2005.3170.

30. Alexander, H.K., and Day, T. (2010). Risk factors for the evolutionary emergence of pathogens. J R Soc Interface *7*, 1455–1474. 10.1098/rsif.2010.0123.

31. Chabas Hélène, Nicot Antoine, Meaden Sean, Westra Edze R., Tremblay Denise M., Pradier Léa, Lion Sébastien, Moineau Sylvain, and Gandon Sylvain (2019). Variability in the durability of CRISPR-Cas immunity. Philosophical Transactions of the Royal Society B: Biological Sciences *374*, 20180097. 10.1098/rstb.2018.0097.

32. Martin, G., Aguilée, R., Ramsayer, J., Kaltz, O., and Ronce, O. (2013). The probability of evolutionary rescue: towards a quantitative comparison between theory and evolution experiments. Philos Trans R Soc Lond B Biol Sci *368*, 20120088. 10.1098/rstb.2012.0088.

33. Djian-Caporalino, C., Palloix, A., Fazari, A., Marteu, N., Barbary, A., Abad, P., Sage-Palloix, A.-M., Mateille, T., Risso, S., Lanza, R., *et al.* (2014). Pyramiding, alternating or mixing: comparative performances of deployment strategies of nematode resistance genes to promote plant resistance efficiency and durability. BMC Plant Biology *14*, 53. 10.1186/1471-2229-14-53.

34. Zhao, J.-Z., Cao, J., Li, Y., Collins, H.L., Roush, R.T., Earle, E.D., and Shelton, A.M. (2003). Transgenic plants expressing two Bacillus thuringiensis toxins delay insect resistance evolution. Nat Biotechnol *21*, 1493–1497. 10.1038/nbt907.

35. Gould, F. (2003). Bt -resistance management—theory meets data. Nat Biotechnol *21*, 1450–1451. 10.1038/nbt1203-1450.

36. Zhao, J.-Z., Cao, J., Collins, H.L., Bates, S.L., Roush, R.T., Earle, E.D., and Shelton, A.M. (2005). Concurrent use of transgenic plants expressing a single and two Bacillus thuringiensis genes speeds insect adaptation to pyramided plants. PNAS *102*, 8426–8430. 10.1073/pnas.0409324102.

37. van Houte, S., Ekroth, A.K.E., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S., Boots, M., Paterson, S., Buckling, A., *et al.* (2016). The diversity-generating benefits of a prokaryotic adaptive immune system. Nature *532*, 385–388. 10.1038/nature17436.

38. Makarova, K.S., Wolf, Y.I., Snir, S., and Koonin, E.V. (2011). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol *193*, 6039–6056. 10.1128/JB.05535-11.

39. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. Science *359*, eaar4120. 10.1126/science.aar4120.

40. Rimbaud, L., Papaïx, J., Barrett, L.G., Burdon, J.J., and Thrall, P.H. (2018). Mosaics, mixtures, rotations or pyramiding: What is the optimal strategy to deploy major gene resistance? Evol Appl *11*, 1791–1810. 10.1111/eva.12681.

41. Samson, J.E., and Moineau, S. (2013). Bacteriophages in food fermentations: new frontiers in a continuous arms race. Annu Rev Food Sci Technol *4*, 347–368. 10.1146/annurev-food-030212-182541.

42. Larder, B.A., Kemp, S.D., and Harrigan, P.R. (1995). Potential mechanism for sustained antiretroviral efficacy of AZT-3TC combination therapy. Science *269*, 696–699. 10.1126/science.7542804.

43. Feder, A.F., Rhee, S.-Y., Holmes, S.P., Shafer, R.W., Petrov, D.A., and Pennings, P.S. (2016). More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. eLife *5*, e10670. 10.7554/eLife.10670.

44. Feder, A.F., Harper, K.N., Brumme, C.J., and Pennings, P.S. (2021). Understanding patterns of HIV multi-drug resistance through models of temporal and spatial drug heterogeneity. eLife *10*, e69032. 10.7554/eLife.69032.

45. Tepekule, B., Uecker, H., Derungs, I., Frenoy, A., and Bonhoeffer, S. (2017). Modeling antibiotic treatment in hospitals: A systematic approach shows benefits of combination therapy over cycling, mixing, and mono-drug therapies. PLoS Comput. Biol. *13*, e1005745. 10.1371/journal.pcbi.1005745.

46. Tyers, M., and Wright, G.D. (2019). Drug combinations: a strategy to extend the life of antibiotics in the 21st century. Nat. Rev. Microbiol. *17*, 141–155. 10.1038/s41579-018-0141-x.

47. Angst, D.C., Tepekule, B., Sun, L., Bogos, B., and Bonhoeffer, S. (2021). Comparing treatment strategies to reduce antibiotic resistance in an in vitro epidemiological setting. Proc Natl Acad Sci U S A *118*, e2023467118. 10.1073/pnas.2023467118.

48. Bosman, A., and Mendis, K.N. (2007). A major transition in malaria treatment: the adoption and deployment of artemisinin-based combination therapies. Am. J. Trop. Med. Hyg. *77*, 193–197.

49. Hastings, I. (2011). How artemisinin-containing combination therapies slow the spread of antimalarial drug resistance. Trends Parasitol. *27*, 67–72. 10.1016/j.pt.2010.09.005.

50. Antao, T., and Hastings, I. (2012). Policy options for deploying anti-malarial drugs in endemic countries: a population genetics approach. Malaria Journal *11*, 422. 10.1186/1475-2875-11-422.

51. Boni, M.F., Smith, D.L., and Laxminarayan, R. (2008). Benefits of using multiple first-line therapies against malaria. Proceedings of the National Academy of Sciences *105*, 14216–14221. 10.1073/pnas.0804628105.

52. Abedon, S.T., Danis-Wlodarczyk, K.M., and Wozniak, D.J. (2021). Phage Cocktail Development for Bacteriophage Therapy: Toward Improving Spectrum of Activity Breadth and Depth. Pharmaceuticals (Basel) *14*, 1019. 10.3390/ph14101019.

53. Fabijan, A.P., Iredell, J., Danis-Wlodarczyk, K., Kebriaei, R., and Abedon, S.T. (2023). Translating phage therapy into the clinic: Recent accomplishments but continuing challenges. PLOS Biology *21*, e3002119. 10.1371/journal.pbio.3002119.

54. Kennedy, D.A., and Read, A.F. (2017). Why does drug resistance readily evolve but vaccine resistance does not? Proc. Biol. Sci. *284*. 10.1098/rspb.2016.2562.

55. McLeod, D.V., Wahl, L.M., and Mideo, N. (2021). Mosaic vaccination: How distributing different vaccines across a population could improve epidemic control. Evolution Letters *5*, 458–471. 10.1002/evl3.252.

56. Cook, J.K., Orbell, S.J., Woods, M.A., and Huggins, M.B. (1999). Breadth of protection of the respiratory tract provided by different live-attenuated infectious bronchitis vaccines against challenge with infectious bronchitis viruses of heterologous serotypes. Avian Pathol *28*, 477–485. 10.1080/03079459994506.

57. Baum, A., Fulton, B.O., Wloga, E., Copin, R., Pascal, K.E., Russo, V., Giordano, S., Lanza, K., Negron, N., Ni, M., *et al.* (2020). Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science, eabd0831. 10.1126/science.abd0831.

58. Kennedy, D.A., and Read, A.F. (2020). Monitor for COVID-19 vaccine resistance evolution during clinical trials. PLoS Biol. *18*, e3001000. 10.1371/journal.pbio.3001000.

59. Pozzetto, B., Legros, V., Djebali, S., Barateau, V., Guibert, N., Villard, M., Peyrot, L., Allatif, O., Fassier, J.-B., Massardier-Pilonchéry, A., *et al.* (2021). Immunogenicity and efficacy of heterologous ChAdOx1–BNT162b2 vaccination. Nature *600*, 701–706. 10.1038/s41586-021-04120-y.

60. Song, S., Zhou, B., Cheng, L., Liu, W., Fan, Q., Ge, X., Peng, H., Fu, Y.-X., Ju, B., and Zhang, Z. (2022). Sequential immunization with SARS-CoV-2 RBD vaccine induces potent and broad neutralization against variants in mice. Virology Journal *19*, 2. 10.1186/s12985-021-01737-3.

61. Altmann, D.M., and Boyton, R.J. (2022). COVID-19 vaccination: The road ahead. Science *375*, 1127–1132. 10.1126/science.abn1755.

62. Chivukula, S., Plitnik, T., Tibbitts, T., Karve, S., Dias, A., Zhang, D., Goldman, R., Gopani, H., Khanmohammed, A., Sarode, A., *et al.* (2021). Development of multivalent mRNA vaccine candidates for seasonal or pandemic influenza. npj Vaccines *6*, 1–15. 10.1038/s41541-021-00420-6.

63. Nelson, R.R. (1978). Genetics of Horizontal Resistance to Plant Diseases. Annual Review of Phytopathology *16*, 359–378. 10.1146/annurev.py.16.090178.002043.

64. Day, T., and Gandon, S. (2012). The evolutionary epidemiology of multilocus drug resistance. Evolution *66*, 1582–1597. 10.1111/j.1558-5646.2011.01533.x.

65. Nagaraja, P., Alexander, H.K., Bonhoeffer, S., and Dixit, N.M. (2016). Influence of recombination on acquisition and reversion of immune escape and compensatory mutations in HIV-1. Epidemics *14*, 11–25. 10.1016/j.epidem.2015.09.001.

66. Moreno-Gamez, S., Hill, A.L., Rosenbloom, D.I.S., Petrov, D.A., Nowak, M.A., and Pennings, P.S. (2015). Imperfect drug penetration leads to spatial monotherapy and rapid evolution of multidrug resistance. Proceedings of the National Academy of Sciences *112*, E2874–E2883. 10.1073/pnas.1424184112.

67. Débarre, F., Lenormand, T., and Gandon, S. (2009). Evolutionary Epidemiology of Drug-Resistance in Space. PLOS Computational Biology *5*, e1000337. 10.1371/journal.pcbi.1000337.

68. McLeod, D.V., and Gandon, S. (2020). Understanding the evolution of multiple drug resistance in structured populations. bioRxiv, 2020.07.31.230896. 10.1101/2020.07.31.230896.

69. Kepler, T.B., and Perelson, A.S. (1998). Drug concentration heterogeneity facilitates the evolution of drug resistance. Proc Natl Acad Sci U S A *95*, 11514–11519.

70. Lévesque, C., Duplessis, M., Labonté, J., Labrie, S., Fremaux, C., Tremblay, D., and Moineau, S. (2005). Genomic organization and molecular analysis of virulent bacteriophage 2972 infecting an exopolysaccharide-producing Streptococcus thermophilus strain. Appl Environ Microbiol *71*, 4057–4068. 10.1128/AEM.71.7.4057-4068.2005.

71. Hynes, A.P., Lemay, M.-L., Trudel, L., Deveau, H., Frenette, M., Tremblay, D.M., and Moineau, S. (2017). Detecting natural adaptation of the Streptococcus thermophilus CRISPR-Cas systems in research and classroom settings. Nat Protoc *12*, 547–565. 10.1038/nprot.2016.186.

Supplementary Information

In the following we give additional details on:
(1) the experimental model and the experimental protocol used to monitor the adaptation of phages introduced in a heterogeneous host population.
(2) the derivation of the probability of evolutionary emergence when the composition of the host population is assumed to be constant. We also derived an approximation of the probability of evolutionary emergence that allowed us to discuss more directly the influence of the inoculum size and the composition of the host population on our experimental results.

1. Experiments

S. thermophilus DGCC 7710 and wild-type phage 2972 [70] were obtained from the Félix d'Hérelle Reference Center for Bacterial Viruses (http://www.phage.ulaval.ca). Phage-resistant strains A, B, and AB were generated using a published protocol [22,71]. Briefly, each of the two protospacers of interest (Table S2) was separately cloned into the plasmid vector pNZ123, which encode a chloramphenicol resistance marker. The two sequence-confirmed recombinant plasmids were separately electroporated into the wild-type strain DGCC 7710, plated on M17 medium supplemented with 5 ug/ml of chloramphenicol, and incubated at 42˚C. The two resulting recombinant strains was separately infected with phage 2972, plated on M17 supplemented with 10 mM $CaCl_2$ and incubated at 42˚C for two days. The resulting colonies representing naturally-occurring bacteriophage-insensitive mutants (BIMs) were checked by PCR and sequencing for the acquisition of the appropriate spacer (from the protospacer cloned into the vector) into the CRISPR1 (CR1) array. The functionality of the new immunity in strains A and B was confirmed by a phage resistance assay [71]. Then, the recombinant plasmid containing the protospacer B was electroporated into strain A as above. Of note, strain A had lost the previous recombinant plasmid during the phage infection assay [71]. The resulting recombinant strain was challenged with a phage-escaping mutant isolated on strain A as described elsewhere [24]. The CR1 of a resulting BIM was analyzed for the acquisition of the second spacer and to confirm strain AB. Finally, phage-escaping mutants on strain B and on strain AB were also isolated and confirmed as previously described [24].

2. Statistics

We used a binomial GLM (Generalized Linear Model) to analyse the effect of the different experimental treatments and the inoculum size (predictor variables) on the probability to observe phages on different types of bacteria (i.e. bacteria that carry 0, 1 or 2 resistances). We used 4 treatment types T: Mixing (A+B), Combining A (A+AB), Combining B (B+AB) and

Pyramiding (AB). We used 10 levels of inoculum size V: $10^{-2}, 10^{-1} \ldots, 10^{6}, 10^{7}$. Our statistical models link the logit of the probability P of observing phages on susceptible bacteria (model M1) on single resistance bacteria (model M2) and on double resistant bacteria (model M3) to the effect of treatment T and inoculum size V as follows:

$$logit(P) \quad a_T \log(V) + b_T$$

where a_T is the slope that accounts for the effect of the inoculum size V and b_T is the intercept corresponding to a fixed treatment effect. The binomial GLM was fitted to the collected data using the R statistical software with the glm() function. For each model, we test if there is an interaction between T and V by computing the AIC of the simplified model (same effect of V in all treatments: $a_T = a$) or the model with and interaction (effect of V depends on the treatment: distinct a_T).

In a first model (M1), we study P_E the probability of emergence of phages which can infect the susceptible bacteria (i.e. detection of phages on a lawn of susceptible bacteria). We found that the model with interaction had a lower AIC than the simplified model (simplified model: AIC = 372.74; interaction model: AIC = 318.54), indicating that the slope is not constant between treatments. Both predictor variables had a significant effect (Treatment: Type II Anova: LR Chi-square =81.1, df=3, $P<2.2\times10^{-16}$; Inoculum size: Type II Anova: LR Chi-square =3744.2, df=1, $P<2.2\times10^{-16}$). We show the estimates of the intercepts and slopes for all treatments in **Table S5** and the corresponding fitted curves in **Figure S3**.

In a second model (M2), we study P_{EE} the probability of escape mutant evolution (i.e. detection of phages on a lawn of singly-resistant bacteria). The simplified model had a lower AIC (AIC=139.25) than the model with interaction (AIC=144.46), indicating that the effect of the inoculum on the probability of escape mutant evolution is independent of the treatment. We found a significant effect on the probability of escape of both the treatment variable (Type II Anova: LR Chi-square =219.46, df=3, $P<2.2\times10^{-16}$) and the inoculum variable (Type II Anova: LR Chi-square =3008.63 df=1, $P<2.2\times10^{-16}$). The estimates of the intercepts and slopes for all treatments are shown in **Table S6** and the corresponding fitted curves are in **Figure 5**.

In a third model (M3), we study P_{EE_2} the probability of double escape mutant evolution (i.e. detection of phages on a lawn of douby-resistant bacteria). The simplified model had a lower AIC (AIC=97.99) than the model with interaction (AIC=103.20), indicating that the effect of the inoculum on the probability of double escape mutant evolution is independent of the treatment. We found a significant effect on the probability of escape of both the treatment

variable (Type II Anova: LR Chi-square =68.97, df=3, $P=7.1\times10^{-15}$) and the inoculum variable (Type II Anova: LR Chi-square =2308.83, df=1, $P<2.2\times10^{-16}$). We show the estimates of the intercepts and slopes for all treatments in **Table S7** and the corresponding fitted curves in **Figure S4**.

3. Theory

3.1 Probability of evolutionary emergence: fixed composition of the host population

We are interested in the fate (extinction or not) of a single virus particle with genotype $i \in \{\varnothing, A, B, AB\}$ inoculated into a large host population with a proportion f_R of resistant hosts. Because we introduce a virus particle, its probability of extinction Q_i is distinct from the probability of extinction q_i of the virus when it is already infecting a host. Indeed, a virus may get extinct even before it manages to infect a single host if the virus particle adsorbs to a resistant host. The relationship between the two quantities is given by:

$$Q_i = (1-f_R)\left(P_i^{\varnothing} q_i + \left(1-P_i^{\varnothing}\right)\right) + f_R \sum_{h \in \{\varnothing, A, B, AB\}} f_h \left(P_i^h q_i + \left(1-P_i^h\right)\right) \quad \text{(S1)}$$

where P_i^h refer to the ability of genotype i to infect a host genotype $h \in \{\varnothing, A, B, AB\}$) given in **Table S2**, and f_h is the frequency of hosts of genotype h among the resistant hosts. For instance, if we focus on the wild-type pathogen we get: $Q_{\varnothing} = (1-f_R) q_{\varnothing} + f_R$.

Because the inoculum is sampled from a population of virus particles where some of them may already carry preexisting mutations (where p_i is the frequency of genotypes with i preexisting escape mutations) the ultimate probability of extinction after introducing a single virus sampled from this population is equal to:

$$Q = \sum_i p_i Q_i \quad \text{(S2)}$$

Next, we computed the probability of extinction q_i which depends both on the pathogen genotype i and the composition of the host population. The pathogen that carries both escape mutation can infect all the hosts and its probability of extinction q_{AB} is simply the extinction probability of a birth-death process (birth rate: $b_{AB} = b(1-c)^2$, death rate: d) which yields:

$$q_{AB}=\begin{cases}\dfrac{d}{b(1-c)^2} & if\ b(1-c)^2>d\\ 1 & if\ b(1-c)^2<d\end{cases} \quad \text{(S3)}$$

Next, to obtain q_A, q_B and $q_\varnothing$, we focused on the probability $q_i(t)$ at time t that a pathogen with genotype i in an infected host, will ultimately go extinct. In a small interval of time dt five different events may take place: (i) the pathogen may spread to a new host without additional escape mutations; (ii) after a single mutation event, the infected host may transmit a pathogen with an additional escape mutation to a new host ; (iii) if $i=0$, a double mutation may occur and the infected host may transmit a pathogen with genotype AB to a new host ; (iv) the infected host (and the pathogen in the host) may die ; (v) nothing may happen during the interval of time dt. Collecting these different terms allowed us to write down recursions for the probability $q_i(t)$, at time t, as a function of the probability $q_i(t+dt)$ and $q_j(t+dt)$ (where j refers to the pathogen genotypes produced by pathogen genotype after acquiring 1 or 2 escape mutations), at time $t+dt$. Under the assumption that the pathogen never reaches a high prevalence, the composition of the host population remains constant and the probabilities $q_i(t)$ are invariant with time. We can thus set $q_i(t)=q_i(t+dt)$ to obtain a recursion equation that allowed us to derive q_i from q_j. Hence, we first can derive q_A and q_B from q_{AB} using this recursion and, in a second step, $q_\varnothing$ from q_A, q_B and q_{AB}. Ultimately, we obtained the probability of emergence of an inoculum of V free virus particles sampled in a population with some preexisting mutations (using equation (S2)):

$$P_E^V=1-(Q)^V \quad \text{(S4)}$$

We computed this probability of emergence for different composition of the host population in a Mathematica 13.2.1 notebook available upon request.

3.2 Probability of evolutionary emergence: depletion of the susceptible hosts

Here, we use another approach to describe the dynamics taking place in our experiments. This dynamic can be summarized by a succession of three main steps that may eventually lead to evolutionary emergence:

- First, when a wild-type virus is introduced in the host population, this virus cannot escape extinction because the proportion of resistant host is so high that the basic reproduction $R_\varnothing=R_0(1-f_R)$ of the wild-type (genotype $\varnothing$) is below 1 and the wild-type genotype is doomed to go extinct. Yet, the wild-type virus may still be able to infect some of the

susceptible hosts before going extinct. The expected number of new infections induced after the inoculation of a virus particle of the wild-type genotype is equal to $N=(1-f_R)\sum_{i=0}^{\infty} R_{\emptyset}{}^{i}=(1-f_R)(1-R_{\emptyset})^{-1}$ where the $(1-f_R)$ term accounts for the probability that the introduced viral particle lands on a susceptible host, and i refers to the position in the epidemic chain that derives from the first case.

- Second, each new infection by the wild-type may generate new escape mutants $i \in \{A, B, AB\}$ and this will occur on average $R_{\emptyset}\sum_i \mu_i$ times per host infected by the wild-type.
- Finally, each of these escape mutants may have the ability to induce an epidemic (provided $R_i>1$) and the probability of each escape mutant to emerge is given by q_i (see above section for the derivation of q_A, q_B and q_{AB}.

Taking into account these three different steps, we can express the probability of evolution emergence after introducing a single virus particle of the wild-type:

$$P_{EE}^{1} \approx N\, R_{\emptyset} \sum_{i \in \{A,B,AB\}} \mu_i (1-q_i) \tag{S5}$$

This approximation can be used to obtain the probability of evolutionary emergence after the introduction of an inoculum of V wild-type virus particles:

$$P_{EE}=1-\left(1-P_{EE}^{1}\right)^{V} \approx V\, P_{EE}^{1} \tag{S6}$$

These expressions can be used to discuss the effect of the composition of the host population. In particular, for simplicity, we can use the following assumptions regarding the mutation rates: $\mu_A=\mu_B=\mu$ and $\mu_{AB}=\mu^2$. In this case, we have:

$$q_A=\begin{cases} \dfrac{d}{b(1-c)\left(1-f_R(1-f_A)\right)} & if\ b(1-c)\left(1-f_R(1-f_A)\right)>d \\ 1 & if\ b(1-c)\left(1-f_R(1-f_A)\right)<d \end{cases} \tag{S7}$$

$$q_B=\begin{cases} \dfrac{d}{b(1-c)\left(1-f_R(1-f_B)\right)} & if\ b(1-c)\left(1-f_R(1-f_B)\right)>d \\ 1 & if\ b(1-c)\left(1-f_R(1-f_B)\right)<d \end{cases}$$

$$q_{AB}=\begin{cases}\dfrac{d}{b(1-c)^2} & if\ b(1-c)^2>d\\ 1 & if\ b(1-c)^2<d\end{cases}$$

Mixing:

Using (S6) and (S7) we obtain:

$$P_{EE}^{M}\approx V\mu\left(\frac{d}{b(1-c)\left(1-f_R\left(1-f_A\right)\right)}+\frac{d}{b(1-c)\left(1-f_R\left(1-f_B\right)\right)}\right)+O\left(\mu^2\right) \quad \text{(S8a)}$$

In the symmetric case where $f_A=f_B=1/2$ this yields:

$$P_{EE}^{M}\approx 2V\mu\left(1-\frac{d}{b(1-c)(1-f_R/2)}\right)+O\left(\mu^2\right) \quad \text{(S8b)}$$

Combining:

$$P_{EE}^{C}\approx V\mu\left(1-\frac{d}{b(1-c)\left(1-f_Af_R\right)}\right)+O\left(\mu^2\right) \quad \text{(S9)}$$

Pyramiding:

$$P_{EE}^{P}\approx O\left(\mu^2\right) \quad \text{(S10)}$$

Data and Code Availability

Data will be fully available in dryad.

Acknowledgements

We thank Geneviève Rousseau for discussion. S.M. acknowledges funding from the Natural Sciences and Engineering Research Council of Canada (Discovery program). S.M. holds a T1 Canada Research Chair in Bacteriophages. S.G. acknowledges funding from the CNRS and from the Agence Nationale de la Recherche ANR (ANR-17-CE35-0012).

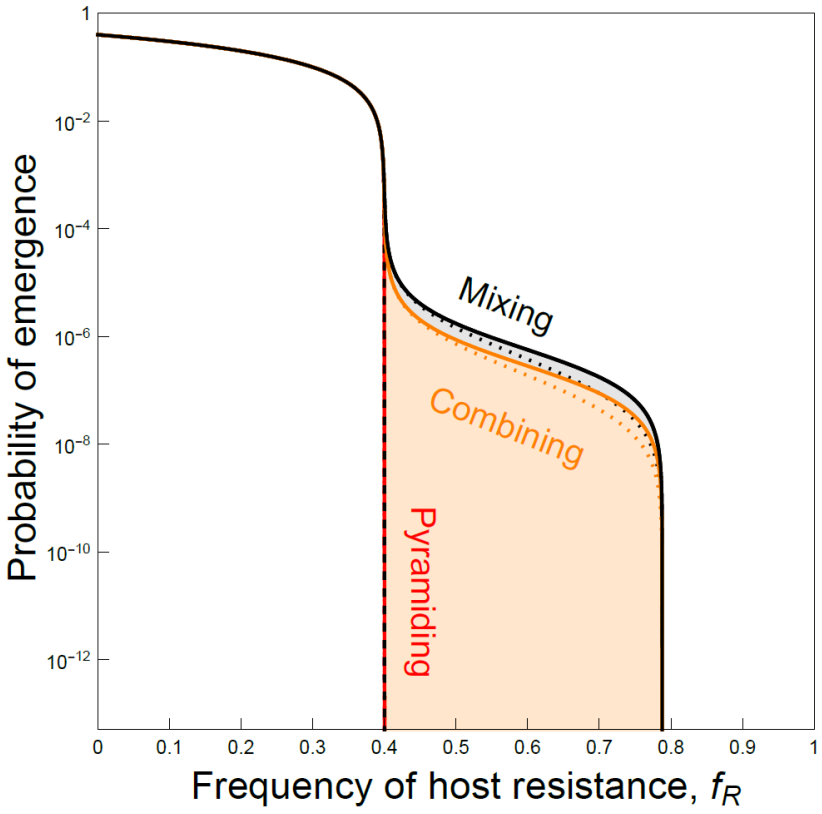


Figure S1: Pathogen emergence under the *mixing* (black), *combining* (orange), or *pyramiding* (red) scenarios when double escape mutations are lethal. We used the same parameters as in **Figure 2** except that double escape mutants were assumed to be lethal so that $q_{AB}=1$. This scenario refers to an extreme form of negative epistasis. This figure shows that evolutionary emergence is still possible in the *mixing* and *combining* treatments when the proportion of resistant bacteria is below a threshold value ($f_R<2\left(1-1/\left(R_0\left(1-c\right)\right)\right)\approx 0.79$) but it becomes impossible in the *pyramiding* treatment.

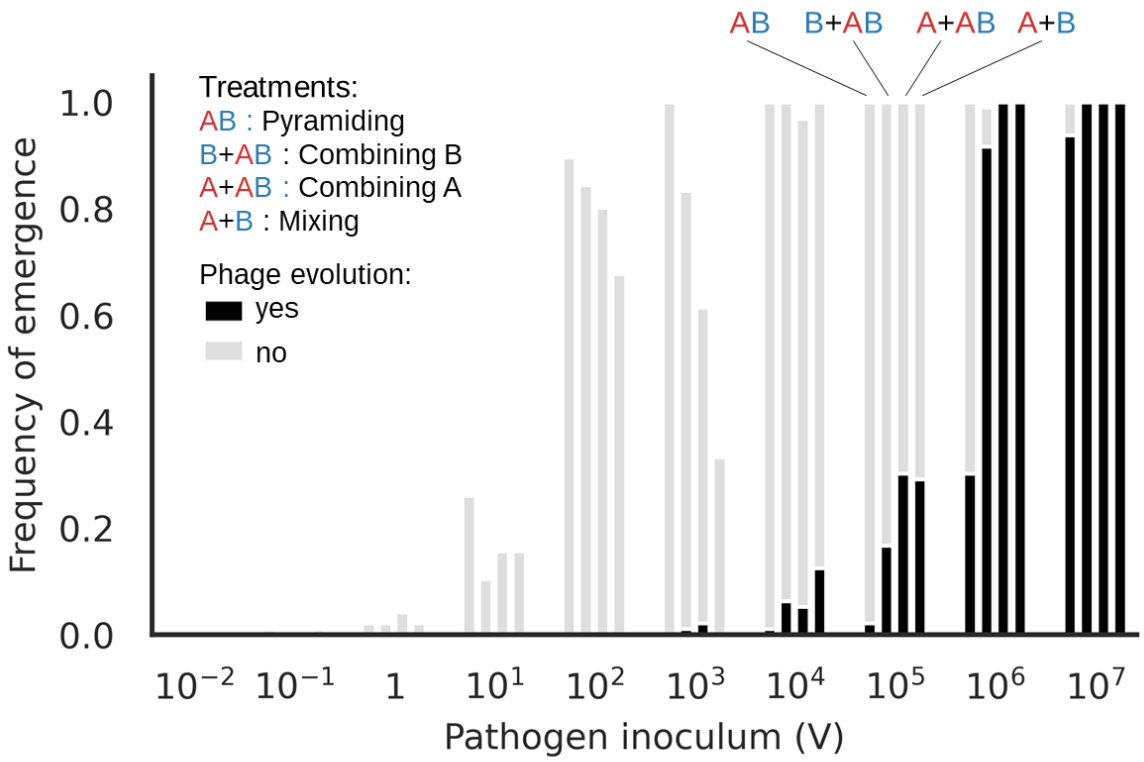


Figure S2: Probability of phage emergence for different inoculum dose (V) and for different host composition treatments (mixing, combining, and pyramiding). In gray, we plot the proportion of populations (among the 96 experimental replicates) in which we recovered viruses capable of infecting the wild-type host. In black, we indicate the fraction of those populations, in which we detected at least one escape mutation (see **Figure 4**).

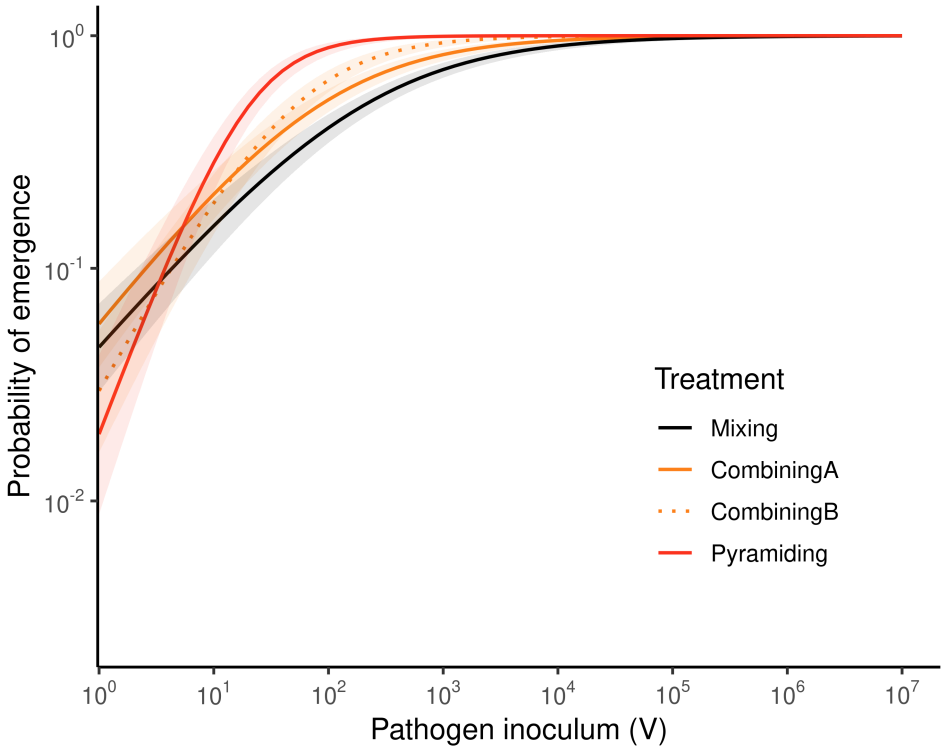


Figure S3: Logistic regression model for the probability of phage emergence for different inoculum dose (V) and for different host composition treatments (mixing, combining, and pyramiding). The estimates for the slope and intercept parameters of each treatment are presented in **Table S5.**

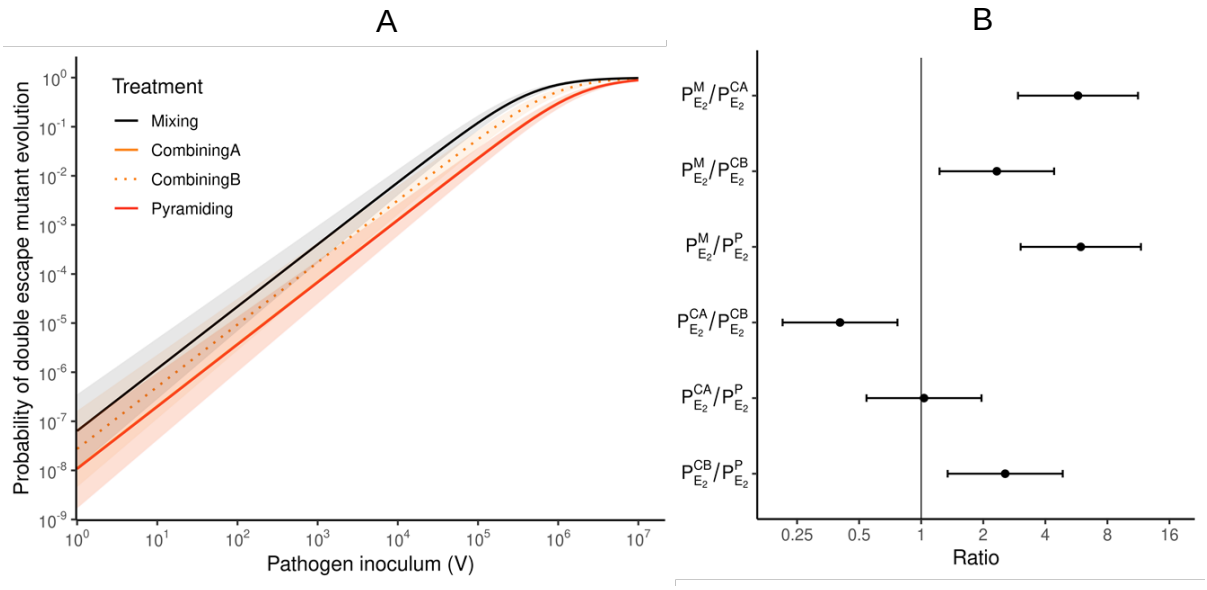


Figure S4: Probability of double escape mutant evolution is highest in the *mixing* treatment. (A) We plot here the estimation of the probability of evolutionary emergence (i.e., the probability to evolve two escape mutations) against the inoculum size V and the resistance treatment. The model can be written as $logit(P_{E2}^{T})\ a_T \log(V)+b_T$, where the slope parameter is the same for all treatments (see **supplementary information** and **Table S7**). The lines indicate the prediction of the statistical model for the different treatments and the shaded areas show 95% confidence interval. (B) We compare the estimated values of b_T for all pairs of treatment and we plot $e^{b_{T1}-b_{T2}} \approx P_{EE}^{T1}/P_{EE}^{T2}$. The error bars show 95% confidence interval.

Table S1: Target sites (protospacer) in the lytic phage 2972 genome. The PAM is indicated in bold and the spacer sequence (the sequence introduced in the CRISPR array of the bacteria) is indicated in red.

Name	Protospacer sequence (spacer + **PAM**, 5'-3')	Position	Function	Targeted phage gene
A	TTATCTGATTTTTTCCCCTTGATTTCGGGGAT**AGAA**	16226-16255	Tail protein	*orf18*
B	TCGTTTTCAGTCATTGGTGGTTTGTCAGCGAA**AGAA**	29988-30017	Replication protein	*orf37*

Table S2: Specificity of the different resistance genotypes of four strains of *S. thermophilus* against four variants of the lytic phage 2972. Table indicates the susceptibility (black) or the resistance (white) of different bacterial clones against different phages. The strain DGCC 7710 is susceptible to all phages. Single-resistance genotypes are resistant to all phages except those that carry a mutation in their genomic sequence targeted by CRISPR. The double-resistance genotype is resistant to all the phages except those that carry a mutation in each of the two target sequences. The number indicates the ability P_i^j of the pathogen genotype i to infect a host genotype j (where $i, j \in \{\varnothing, A, B, AB\}$).

		Bacteria clones			
		DGCC 7710	A	B	AB
Virus strains	2972-Ø	1	0	0	0
	2972-A	1	1	0	0
	2972-B	1	0	1	0
	2972-AB	1	1	1	1

Table S3: Frequency of preexisting escape mutations against the different resistant bacteria in the population of phage 2972 used to inoculate the populations.

Name	Frequency of preexisting escape mutations [95% confidence interval]
A	$3.67\ 10^{-6}$ [$2.91\ 10^{-6}$, $4.43\ 10^{-6}$]
B	$1.69\ 10^{-6}$ [$6.37\ 10^{-7}$, $2.75\ 10^{-6}$]
AB	0

Table S4: Rate of escape mutations against resistance A and B estimated in Chabas *et al.* [31].

Name	Probability of escape mutations [95% confidence interval]
A	$1.2\ 10^{-6}$ [$6.2\ 10^{-7}$, $1.7\ 10^{-6}$]
B	$7.1\ 10^{-7}$ [$2.9\ 10^{-7}$, $1.1\ 10^{-6}$]

Table S5: GLM estimates for the probability of emergence for the different treatments at different inoculum for model M1. The estimates for intercept (b_T) and slope (a_T) parameters presented in methods is shown with the corresponding standard error (see **supplementary information**).

	Parameter	Estimate	Standard error
Intercept	b_M	-3.24	0.23
	b_{CA}	-2.79	0.23
	b_{CB}	-3.48	0.32
	b_P	-3.92	0.41
Slope	a_M	1.32	0.09
	a_{CA}	1.46	0.10
	a_{CB}	2.04	0.16
	a_P	3.00	0.29

Table S6: GLM estimates for the probability of evolutionary emergence for the different treatments at different inoculum for model M2. The estimates for intercept (b_T) and slope (a_T) parameters presented in methods is shown with the corresponding standard error for the simplified model with no interaction between treatment and inoculum: $a=a_T$ (see **supplementary information**).

	Parameter	Estimate	Standard error
Intercept	b_M	-15.24	0.79
	b_{CA}	-15.37	0.79
	b_{CB}	-16.03	0.83
	b_P	-18.69	0.95
Slope	a	3.00	0.15

Table S7: GLM estimates for the probability of evolutionary emergence of a double mutant for the different treatments at different inoculum for model M3. The estimates for intercept (b_T) and slope (a_T) parameters presented in methods is shown with the corresponding standard error for the simplified model with no interaction between treatment and inoculum: $a=a_T$ (see **supplementary information**).

	Parameter	Estimate	Standard error
Intercept	b_M	-16.57	0.87
	b_{CA}	-18.32	0.95
	b_{CB}	-17.41	0.91
	b_P	-18.35	0.95
Slope	a	2.92	0.15

Chapter 3:

Competition and coevolution drive the evolution and the diversification of CRISPR immunity

Published in Nature Ecology and Evolution
in August 2022

nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-022-01841-9

Check for updates

Competition and coevolution drive the evolution and the diversification of CRISPR immunity

Martin Guillemet[1], Hélène Chabas[1,2], Antoine Nicot[1], François Gatchich[1], Enrique Ortega-Abboud[1], Cornelia Buus[3], Lotte Hindhede[3], Geneviève M. Rousseau[4,5], Thomas Bataillon[3], Sylvain Moineau[4,5,6] and Sylvain Gandon[1]✉

The diversity of resistance challenges the ability of pathogens to spread and to exploit host populations. Yet, how this host diversity evolves over time remains unclear because it depends on the interplay between intraspecific competition among host genotypes and coevolution with pathogens. Here we study experimentally the effect of coevolving phage populations on the diversification of bacterial CRISPR immunity across space and time. We demonstrate that the negative-frequency-dependent selection generated by coevolution is a powerful force that maintains host resistance diversity and selects for new resistance mutations in the host. We also find that host evolution is driven by asymmetries in competitive abilities among different host genotypes. Even if the fittest host genotypes are targeted preferentially by the evolving phages, they often escape extinctions through the acquisition of new CRISPR immunity. Together, these fluctuating selective pressures maintain diversity, but not by preserving the pre-existing host composition. Instead, we repeatedly observe the introduction of new resistance genotypes stemming from the fittest hosts in each population. These results highlight the importance of competition on the transient dynamics of host–pathogen coevolution.

Coevolution is thought to be a powerful evolutionary force at the origin of biological diversity[1–3]. Negative-frequency-dependent selection generated by coevolution can promote the emergence and the maintenance of genetic diversity in interacting species[2,4,5]. On the other hand, maintenance of genotype diversity is also affected by intrinsic differences in competitive abilities among genotypes. If this asymmetric competition is strong it can lead to the exclusion of less competitive genotypes and to a drop in diversity. The interplay between coevolution and competition has been explored theoretically with models based on the 'kill-the-winner' hypothesis that explicitly accounts for the influence of phages on diverse host communities[6–8]. This framework, however, is meant to describe the ecological dynamics of interacting host species. Several experimental studies have explored the influence of phages on interspecific competition among different bacterial species[9–12] but intraspecific competition is harder to monitor. Studying the interplay between competition among host genotypes and coevolution with pathogens is particularly challenging within a host species because it requires detailed knowledge of the genetic determinants of the specificity of the host–virus interaction to track the dynamics of different host genotypes[13].

Here, we track the coevolutionary dynamics of CRISPR immunity of the bacterial species *Streptococcus thermophilus* with its lytic phage 2972. This model system offers unique opportunities to explore the microevolutionary processes driven by competition among different bacteria and antagonistic coevolution between bacteria and their viral pathogens. In *S. thermophilus*, coevolution with phage 2972 is mainly driven by two (type II-A) CRISPR–Cas loci (CR1 and CR3), which allow the bacteria to incorporate 30-bp DNA sequences (spacers) from the genome of an infecting phage in the CRISPR array[14–17]. After transcription, each spacer RNA is used as a guide by Cas9 to target and cleave the corresponding target sequence (the protospacer) in the phage genome, thereby halting virus replication and reducing its titre. Phages can escape CRISPR immunity via mutations in the protospacers that avert recognition by the Cas complex. These mutations have been shown to be particularly effective at escaping immunity when they are located at specific positions in the protospacers such as the PAM (protospacer-adjacent motif) or the seed[18,19]. Crucially, the sequencing of the CRISPR array of the populations of bacteria and the whole-genome sequencing of the populations of phages allowed us to fully characterize the specificity of the infection network, without any phenotypic assays. Here we focus on the CRISPR array of the CR1 locus that has been shown to be the most active of the CRISPR loci of *S. thermophilus* against phage 2972 (ref. [15]).

To study how host diversity affects the dynamics of CRISPR immunity, we designed a short-term coevolution experiment (pictured in Fig. 1) where we followed the evolution of CRISPR immunity in the absence of phages (treatment A), in the presence of an initially monomorphic population of phages (treatment B), in the presence of an initially polymorphic population of phages (treatment C). We started each culture with a mix of 17 different bacterial strains in equal frequencies: one strain was fully sensitive to the wild-type lytic phage 2972 (strain DGCC 7710) and each of the remaining 16 strains carried a distinct single-spacer resistance in the CRISPR array at the CR1 locus (Supplementary Table 1). These strains were obtained from a previous study after exposing the susceptible strain DGCC 7710 to the phage 2972, leading to the

[1]CEFE, CNRS, Univ Montpellier, EPHE, IRD, Montpellier, France. [2]Institute of Integrative Biology, Department for Environmental System Science, ETH Zurich, Zurich, Switzerland. [3]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [4]Département de biochimie, microbiologie, et bio-informatique, Faculté des sciences et de génie, Université Laval, Québec City, Canada. [5]Groupe de recherche en écologie buccale, Faculté de médecine dentaire, Université Laval, Québec City, Canada. [6]Félix d'Hérelle Reference Center for Bacterial Viruses, Faculté de médecine dentaire, Université Laval, Québec City, Canada. ✉e-mail: sylvain.gandon@cefe.cnrs.fr

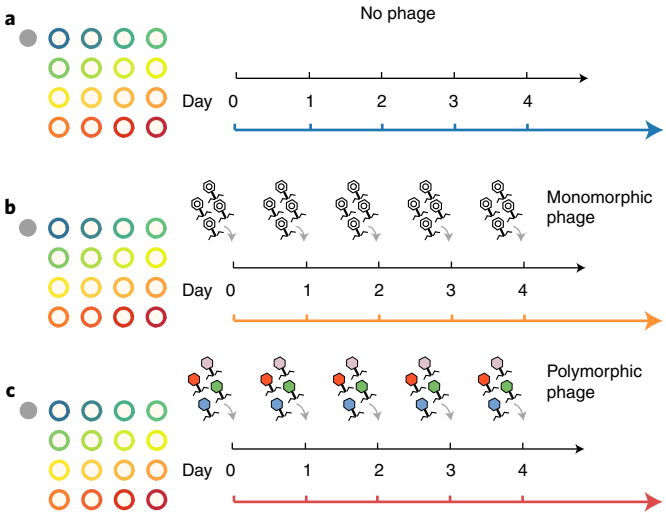


Fig. 1 | The three treatments of our coevolutionary experiment. Bacterial cultures were inoculated with a mix of 17 different strains in equal frequencies: one strain (filled grey circle) was susceptible to the wild-type lytic phage 2972 and the remaining 16 strains (empty coloured circles) carrying a distinct single-spacer resistance in the CRISPR 1 (CR1) locus. **a**, The daily transfers of 1% of the bacterial culture with no exposure to phages, treatment A. **b**, The daily transfers of 1% of the bacterial culture with inoculation of 10^5 phages at each transfer sampled from a monomorphic population of the wild-type phage, treatment B. **c**, The daily transfers of 1% of the bacterial culture with inoculation of 10^5 phages at each transfer sampled from a polymorphic phage population, treatment C. This polymorphic phage population is a mix of 16 escape variants that were previously selected to escape each of the 16 CR1 resistances of the polymorphic population of bacteria.

spontaneous acquisition of a single additional spacer targeting distinct protospacers in the phage genome[20]. Crucially, the bacteria may have acquired additional mutations in the bacterial genome outside the CRISPR locus during this selection procedure. We carried out whole-genome sequencing of all the strains to identify these mutations (Supplementary Table 3). A previous study demonstrated that the ability of phage 2972 to escape CRISPR immunity differs among the 16 resistant strains[20].

For each of the three experimental treatments, we transferred 1% of each replicate culture to a fresh medium for four consecutive days. In the absence of phages (treatment A), the change in the relative frequency of the different host genotypes informed us about the competitive abilities of the 17 bacterial strains. This treatment allowed us to evaluate the ability to maintain diversity on the CRISPR locus in the absence of selection for resistance (Extended Data Fig. 1). If some strains are more competitive, they are expected to outgrow the others and induce a rapid drop of diversity. The two other treatments allowed us to follow the interplay between competition and antagonistic interactions with phages on the evolution of the bacteria. At the beginning of each transfer, we added 10^5 phages from a monomorphic or a polymorphic phage populations (treatments B and C, respectively). The monomorphic phage population was obtained from the amplification of the wild-type phage 2972 that infects only the sensitive host strain (about 6% of the host population at the onset of our experiment). In the polymorphic phage population, we used a mix of 16 escape phage variants (phage cocktail) that were previously selected to escape each of the 16 CRISPR CR1 resistances of the polymorphic population of bacteria[20] (Supplementary Table 2). This recurrent inoculation of phages at each transfer was used to maintain a minimal amount of phage in treatments B and C. As pointed out below, this immigration of phages did not prevent phage adaptation and coevolution with the host.

To monitor the demography and evolution of bacteria we used spacers as barcodes and sequenced the 5′-end of the CRISPR array of the CR1 locus of the bacteria (Methods). This sequencing strategy allowed us to identify the emergence and the spread of additional resistant strains with new spacers in the CRISPR array[21]. To monitor the evolution of the phage populations we used whole-genome sequencing in the treatments exposed to the virus (treatments B and C) to identify new mutations and estimate their frequencies.

Results and discussion. *Phage diversity drives infection dynamics.* The treatments had great effects on both the bacteria and the phage densities (Fig. 2). The monomorphic phage treatment had a limited impact on bacterial growth the first day but led to a massive phage epidemic on the second day, marked by a drop in host density and an increase in the viral pathogen density. In contrast, the polymorphic phage treatment immediately led to a viral outbreak on the first day. Yet, under all phage treatments the bacterial populations eventually recovered and by day 4 they reached a density close to the no-phage treatment (Fig. 2a).

Evolution and diversification of CRISPR immunity. To monitor the evolutionary dynamics of bacteria, we tracked the diversity of CRISPR immunity at the CR1 locus and estimated the frequency h_i of each resistance genotype i in the population. We computed the effective number of host genotypes[22] across time for each replicate (Fig. 3). The effective number of host genotypes dropped very fast in the treatment without phage and remained very low until the end of the experiment. Experimental treatments had a significant effect on the mean effective number of host genotypes (Methods) (day 1 $F_{2,20}=1{,}431$, $P=2.6\times10^{-22}$; day 4 $F_{2,20}=7.80$, $P=3.1\times10^{-3}$). Compared to the treatment without phage at day 1 (effective number of genotype and 95% confidence interval (CI) 8.90 (8.72, 9.10)), exposure to a monomorphic phage population initially led to a faster drop in diversity (6.91 (6.74, 7.08)), but exposure to a polymorphic phage treatment maintained a high level of diversity (14.88 (14.63, 15.12)). Both phage treatments led to the maintenance of more diversity at the end of the experiment than the treatment without phage (day 4, no phage 1.96 (1.92, 1.99), monomorphic phage 3.66 (2.79, 4.52) and polymorphic phage 5.33 (3.78, 6.98)). The maintenance of diversity in host populations exposed to phages supports the idea that coevolution can drive the diversification of host populations[1,7,23,24]. The variation in the dynamics of diversity among replicate populations exposed to phages illustrates the impact of demographic stochasticity on this coevolutionary dynamics, particularly after demographic bottlenecks caused by viral epidemics.

Next, to better understand what drives the dynamics of CRISPR diversity we examined the competition between the different bacterial strains using modified Muller plots that provide a description of both the changes in density and in the genetic composition for each replicate population of bacteria (Fig. 4). All the replicates followed very similar dynamics in the treatment without phage (Fig. 4): one of the strain (indicated in red, strain 31725) outcompeted the other strains and nearly reached fixation by day 2, but another strain (indicated in green, strain 16236) increased in frequency towards the end of the experiment. These results indicate the main differences in competitive abilities among strains. The fitter strain (in red) is not the phage-sensitive wild-type strain but one of the 16 resistant strains (Extended Data Fig. 1). Whole-genome sequencing of the 17 strains used at the beginning of the experiment revealed the existence of other mutations across the bacterial genome outside the CRISPR locus (Supplementary Table 3). These mutations were acquired during the selection process that led to the natural acquisition of a new spacer on the CR1 locus[25]. For instance,

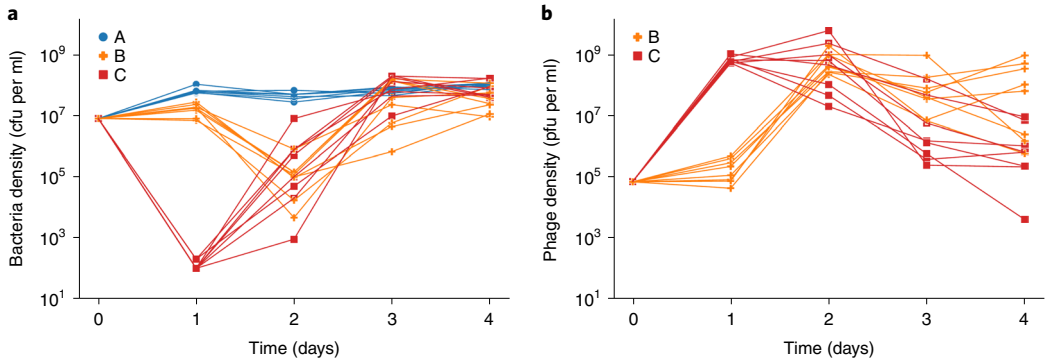


Fig. 2 | Demography of bacteria and phages. a,**b**, The density of bacteria (**a**) and density of phages (**b**) in the three experimental treatments. All replicates are shown, seven for the control (treatment A) and eight for the two phage treatments (treatments B and C). Blue points show the data in the absence of phages, while orange and red show the data for the monomorphic and polymorphic phage treatments, respectively.

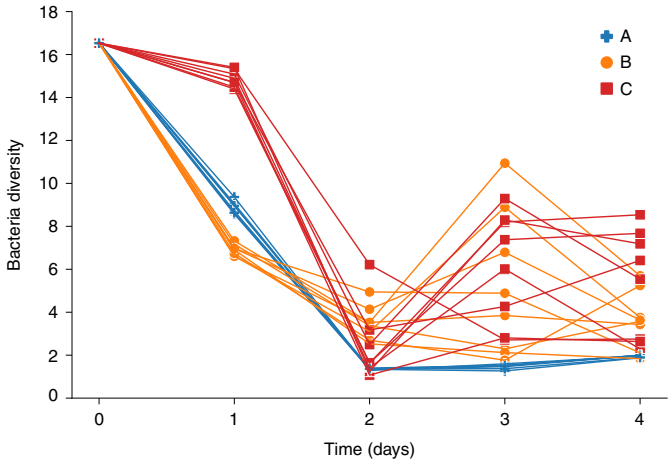


Fig. 3 | Dynamics of the diversity of CRISPR immunity. Dynamics of CRISPR locus diversity computed with the effective number of host genotypes: $1/(\sum_{i=1}^{n} h_i^2)$. Blue points show the data in the absence of phages, orange and red show the data for the monomorphic and the polymorphic phage treatments, respectively.

the 'red' strain has eight unique non-synonymous mutations in different genes. By contrast, the sequencing of the 'green' strain revealed only two unique synonymous mutations in different genes. The competitive ability of the 'green' strain is also more puzzling because this strain was initially less fit and only increased in frequency towards the end of the experiment. A more detailed analysis of the contribution of each of these mutations on the competitive ability of the strains falls beyond the scope of this study. But these highly consistent measures of fitness among replicates in the treatment without phage allowed us to study how competition affects the coevolutionary dynamics in the populations exposed to phages.

Figure 4 shows how phages affect both the density of bacteria and the evolution of CRISPR immunity. As expected from Fig. 3, the presence of phages maintains a higher number of strains. More specifically, we observe the emergence of several new resistant strains that carry up to three additional spacers in the CRISPR array, which are indicated by dark colours in Fig. 4. In all treatments exposed to phages, almost all the bacterial populations end up being dominated by lineages that are descendants of the two most competitive strains identified in the absence of phages (Extended Data Fig. 2). In other words, the increase in diversity observed at the end of the experiment in the populations exposed to phages (Fig. 3) is not due to the initial diversity being restored, but to new resistance genotypes that arose via the acquisition of new spacers in the CRISPR array of the winners of the competition among bacterial strains (Extended Data Fig. 3). It is unlikely that the per capita rate of acquisition of new spacers differs among the different bacterial strains. Indeed, none of the mutations found in these strains were in the genes known to control the adaptation step of type II-A CRISPR–Cas system (that is, *cas1*, *cas2*, *csn2*, *cas9*). Variation in the densities of bacteria provides a more parsimonious explanation for the faster acquisition of new spacers in the winners of the competition. Since the winners of the competition were more abundant, they were also more likely to acquire new spacers.

The comparison among replicate populations revealed very different dynamics in the presence of phages. To study this variation, we measured the amount of genetic differentiation among replicate populations within each treatment (Extended Data Fig. 4). Complementary measures of host differentiation (F_{ST} and D) allowed us to quantify the changes in population composition due to drift and selection among replicates (Methods). As expected, differentiation remained very low in the treatment without phages because all replicates displayed very similar dynamics. In contrast, exposure to phages led to the acquisition of distinct spacers in different replicates, which led to a rapid increase in differentiation among host populations. This is particularly noticeable right after the massive demographic bottleneck that took place after the first day in the treatment exposed to a polymorphic phage population.

Another way to demonstrate the influence of phages on bacterial evolution is to detect the presence of negative-frequency-dependent selection. As expected, in the absence of phages the change in strain frequency between time t and $t+1$ is independent of strain frequency at time t (Fig. 5a). Exposure to phages, however, yields a strongly negative relationship between these two quantities (the presence of phages has a highly significant effect on the slope of the regression line in both the monomorphic and the polymorphic phage treatments, Methods), which indicates that more frequent strains tend to be selected against because they are preferentially targeted by phages (Fig. 5b,c). All these results confirm the expected impact of viral pathogens on the diversification of host resistance[23,26] and highlight the relevance of the kill-the-winner hypothesis[6,7].

Phage coevolution across space and time. The sequencing of the phage populations revealed the emergence and the spread of many mutations across the phage genome (Extended Data Fig. 5). Most of these mutations were located in the protospacer regions targeted by CRISPR immunity and particularly in the PAM or the seed of

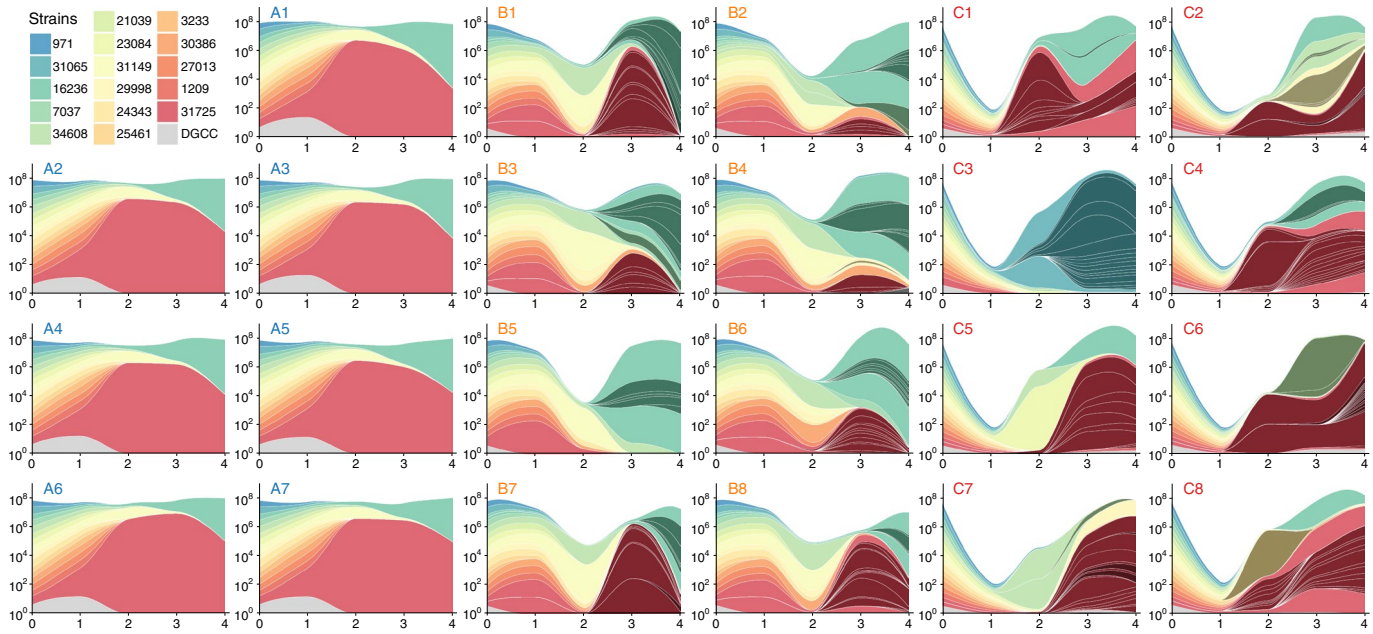


Fig. 4 | Host populations resist phages through the diversification of the CR1 locus. Modified Muller plots show the dynamics of different host genotypes in each replicate of the three experimental treatments as indicated above each graph (A for the no-phage control, B for the monomorphic phage treatment and C for the polymorphic phage treatment). The total height for each day shows the bacterial density (in cfu per ml) on a log scale, and the different colours show the proportion of the different host genotypes at each time point on a linear scale. The 17 strains that were added on day 0 (including the wild-type in grey) are shown in the legend. The blue-to-red colour scale ranks the strains according to their initial fitness as detailed in Extended Data Fig. 1. A darker coloured strain seen in later days stemming from inside one of the 17 original strains shows the acquisition of a new spacer. An even darker colour represents strains with two additional spacers (two new spacers is the maximum represented here). When there are several parallel acquisitions of new spacers, the new strains are separated with white lines. The lines are smoothed between each day.

protospacers (Extended Data Fig. 5). These mutations are expected to be strongly beneficial as they allow the virus to escape CRISPR immunity[18]. Knowing the genetic specificity of CRISPR immunity allows us to assign phenotypic effects to these mutations without any additional experimental measures. We combined sequencing data from the bacteria and the phages to compute the mean fitness $\overline{w}$ of each phage population using:

$$\overline{w} = \sum_{i=1}^{n} h_i p_i \quad (1)$$

where n is the total number of host genotypes, h_i is the frequency of host genotype i and p_i is the frequency of phage variants that can infect genotype i. Here, the mean fitness measures the mean fraction of the host population available to a randomly sampled virus in the phage population. This in silico measure of phage mean fitness provides a powerful way to estimate phage adaptation to contemporaneous host populations (when phage and bacteria frequencies are sampled in the same replicate and at the same point in time) but also across space and time[3,23].

Measures of phage adaptations across all time points revealed a striking pattern where levels of phage adaptation are maximal against host populations from the recent past (Fig. 6). In contrast, the degree of phage adaptation drops very rapidly against bacteria from the future in both phage treatments. This pattern is precisely the one expected under the rapid coevolution dynamics that are predicted to emerge in coevolutionary models[3,27–29]. The particularly rapid drop of phage mean fitness when matched against bacteria from the future shows how quickly bacteria are able to develop new resistance to the phages. This is consistent with the intrinsic asymmetry inherent to CRISPR specificity: bacteria have access to hundreds of different protospacers from the phage genome[18] allowing them to raise a diverse and distributed immune defence to the phage population at once[30]. In contrast, only mutations in the targeted viral genomic region (Extended Data Fig. 5) can provide an effective way for phages to escape CRISPR immunity, and only against one resistance (one spacer) at a time.

Measures of mean fitness across space allowed us to compute phage local adaptation to determine whether the phage is more adapted to sympatric (same replicate) than to allopatric (different replicate) host populations (Methods)[27,31]. Figure 6 shows the buildup of local adaptation across time in the two phage treatments. Local adaptation remains very low in the treatment with the monomorphic phage population. In contrast, we detect a strong pattern of phage local adaptation in the treatment with a polymorphic phage population. In particular, phage local adaptation is extremely strong ($0.53(\pm 0.18)$) at day 2, which coincides with the time at which host differentiation is maximal. Indeed, phage local adaptation can occur only when the composition of sympatric host populations differs substantially from allopatric host populations.

The dynamics of differentiation varied between the two phage treatments. Even if we detect significant differentiation among replicate populations in the treatment with the monomorphic phage population, the Q_{ST} that measures phenotypic differentiation (Methods) remains very low (Extended Data Fig. 6) because most escape mutations occur on the same protospacer (Extended Data Fig. 5). Indeed, the high number of escape mutations we observe (in the monomorphic phage treatment) in the terminal region of the phage genome (around 31 kb, Extended Data Fig. 5) can be explained by the presence of the protospacer associated with the 'red' resistant host (strain 31725). Selection to escape this resistance was very intense in this treatment (because this resistance strain was the most frequent) and each of these mutations correspond to different

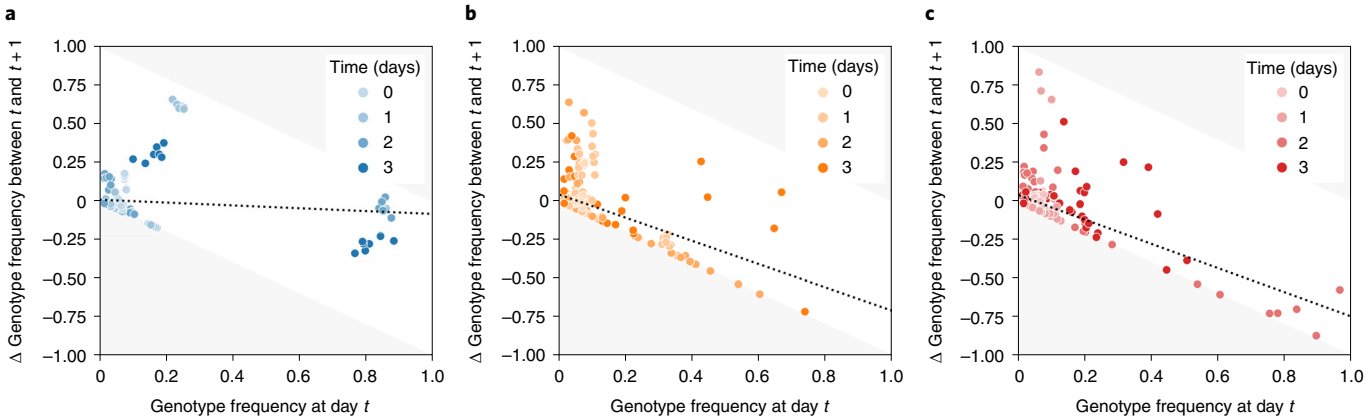


Fig. 5 | Phages induce negative frequency-dependent selection. a–c, Evidence for negative frequency-dependent selection in the two phage treatments (**b** and **c**) compared to the control (**a**). In every replicate, the change in frequency Δ of a genotype between day *t* and *t*+1 is plotted according to its frequency at day *t*. A linear regression is plotted in each panel to highlight the negative frequency-dependent selection (or lack thereof in the control). A significantly negative slope indicates that more frequent genotypes are counter-selected and tend to decrease in frequency the following day (Methods). The light grey area refers to unfeasible changes in frequency. Blue points show the data in the absence of phages, orange and red dots show the data for the monomorphic and polymorphic phage treatments, respectively.

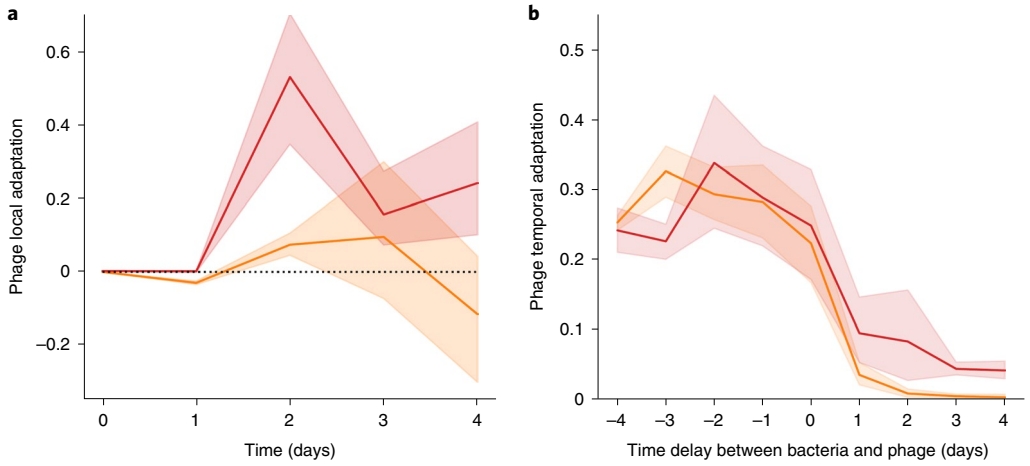


Fig. 6 | Phage adaptation across space and time. a,**b**, Evidence for local (**a**) and temporal adaptation (**b**) to CRISPR immunity in the two phage treatments. Local adaptation measures the amount of phage adaptation to sympatric host populations (same replicate) relative to allopatric host populations (different replicates) for each treatment at different points in time (Methods). Temporal adaptation measures, for a given replicate, the mean fitness of the phage population to host sampled at different points in time. A positive (negative) time delay between bacteria and phage indicates that the bacteria is sampled in the future (past) relative to the sampling point of the phage (Methods). Orange and red show the average level of phage adaptation for the monomorphic and polymorphic phage treatments, respectively. The shaded areas show the 95% confidence interval computed using the Jackknife method (Methods). The horizontal dotted line in **a** shows the position of the zero, where there is no local adaptation.

solutions leading to the same escape phenotype. In contrast, in the treatment with a polymorphic phage population, both F_{ST} and Q_{ST} are increasing on day 2 after the divergence of bacterial populations (Extended Data Fig. 6) and the distribution of escape mutations is more evenly distributed across the phage genome (Extended Data Fig. 5). Note, however, that the speed of phage adaptation seems too low to catch up with the build up of CRISPR immunity. Initial diversity in the polymorphic treatment yields faster adaptation but the number of phage mutations in protospacers stops increasing by day 1 (Extended Data Fig. 7). In the monomorphic treatment, the saturation in the number of mutations in protospacers is delayed (Extended Data Fig. 7). The drop in local adaptation with time is consistent with the overall drop in phage density we observed in most phage populations (Fig. 2b). This suggests that the phages are losing the coevolutionary arms race with their hosts, which is in line with previous studies showing that CRISPR immunity often yields phage extinction in this system[32–34]. Besides, we found some evidence that evolution of new resistance may also be due to the second active CRISPR–Cas system (CR3) in this host in which we detected spacer acquisition starting at day 3 or 4 (Supplementary Table 4). Accounting for evolution at the CR3 locus when estimating phage fitness magnifies the drop of mean fitness of phage populations (Extended Data Fig. 8). The recurrent introduction of ancestral phages at the beginning of each transfer was used to avoid phage extinctions in treatments B and C. This recurrent immigration probably had a limited impact on the coevolutionary dynamics in our short-term experiment because immigrants were unable to infect bacteria that rapidly acquired new resistances. Yet, migration

could have more impact on long-term persistence as it would allow phages to recolonize susceptible host populations.

Host competition governs the coevolution-driven diversification. We can track the dynamics of phage adaptation across space and time but can we predict the speed at which the phage escapes the phage-resistant strains? The speed of adaptation is governed (1) by the rate of mutation, which has been shown to vary among protospacers in a previous experiment[20], (2) by the strength of selection associated with the ability to escape CRISPR immunity against a specific protospacer and (3) by the fitness cost of these escape mutations. Because the fitness cost of these mutations has been shown to be a poor predictor of the durability of CRISPR–Cas immunity[20] we focus on the first two points. In the treatment with a polymorphic phage population, the rate of mutation is not limiting because the mutations against the 16 original spacers are pre-existing. In this phage treatment, as expected, we do not find a correlation between the speed of phage adaptation and the rate of escape mutation for different protospacers (Pearson's $r=-0.26$, $P=0.41$) (Extended Data Fig. 9). In contrast, the speed of adaptation is governed by the competitive ability of the different resistant strains (Pearson's $r=0.83$, $P=7.5\times10^{-5}$). Indeed, this competitive ability is a good predictor of the abundance of each resistant strain and, consequently, a good predictor of the fitness benefit associated with the ability to exploit these resistant strains. We obtain a very similar pattern in the monomorphic phage treatment (no correlation with the mutation rate: Pearson's $r=-0.02$, $P=0.41$; strong correlation with competitive ability: Pearson's $r=0.95$, $P=3.4\times10^{-8}$). These results indicate that phage mutation is not limiting and phage adaptation is mostly driven by the more abundant (that is the more competitive) phage-resistant strains of bacteria.

Conclusion. Our short-term evolution experiment demonstrates that the coevolutionary battle taking place between bacteria and phages is a potent evolutionary force driving the rapid diversification of interacting populations. The presence of phages generates strong negative-frequency-dependent selection, which prevents the loss of diversity of CRISPR immunity. This is consistent with the kill-the-winner hypothesis[6,7] that states viruses can maintain the host diversity. Similar conclusions were reached from studies that explored the interplay between interspecific competition and coevolution with phages[10,12]. But here, we could track the emergence of new resistance mutations (new spacers in the CRISPR array) and these new mutations are not equally distributed among the bacterial strains present initially. Indeed, we see that the initial host diversity vanished rapidly (Extended Data Fig. 3) and in all but one replicate population exposed to phages, the bacterial population at day 4 is dominated by strains that descend from the most competitive strains (the 'winners') identified in the control (Fig. 4 and Extended Data Fig. 2). To understand these results, it is important to recall that host adaptation results from both the selection imposed by phages at the CRISPR locus and the selection imposed on the rest of the bacterial genome. The recurrent bottlenecks in the host population size induced by phage infections may lead to a faster fixation of new mutations. Even if these additional mutations are expected to be often deleterious[35], their effects on fitness will vary and introduce variation in competitive abilities among strains[35,36]. In the absence of phages, fitter host genotypes consistently outcompete other strains. In the presence of the phage, viral adaptation targets preferentially more abundant and competitive strains. But the evolution of CRISPR immunity allows the winners of the intraspecific competition to strike back after phage adaptation. Ultimately, this explains why diversity is generally maintained and originates from the descendants of the winners in populations exposed to phages. This feedback of competition on coevolutionary dynamics can also be discussed in the light of the recent 'royal family model'[37]. In a classic version of the kill-the-winner framework, the most frequent host strain is preferentially targeted by the evolving population of pathogens and is driven to low frequency. Next, another host strain rises to high frequency and the cycle repeats. In the royal family model, intrinsic asymmetries in competitive abilities indicate that the newly rising host genotypes are likely to descend from the previously dominating genotypes. Our experimental results squarely fit within this framework as we can readily identify a royal family in the bacteria population that often derives from the more competitive strains (the red and green strains in Fig. 4). Note, however, that our experiment also features the rise of a new royal family (strain 31065) in one population after a particularly strong demographic bottleneck (replicate C3 in Fig. 4 and Extended Data Fig. 2). Hence, the stochastic acquisition of new resistance may open up new opportunities for previously dominated strains of bacteria. As expected from the royal family model this evolutionary dynamics within the population of bacteria implies that there is also a royal family of phages, which is particularly adapted to the royal family of bacteria (Extended Data Fig. 10). We stress that our short-term experiment focuses on a very specific scenarios where (1) the initial diversity in the host population was manipulated artificially with equal frequency among different strains and no multiresistance to the phage and (2) the initial diversity of the phage population was also manipulated experimentally (treatment B versus C). Yet, the distributions of CRISPR immunity and phage diversity are expected to build up naturally after a phage epidemic and the network structure of strain diversity may be very different from the one used in our experiment[38]. Our work should be viewed as an attempt to monitor coevolutionary dynamics experimentally and the relevance of the royal family model remains to be investigated in a more natural setting.

This short-term experiment demonstrates that ecological and evolutionary processes can take place on a similar time scale. A better understanding of the coevolution between CRISPR immunity and phages requires a more comprehensive theoretical framework considering the mutations involved in the interaction as well as in the rest of the genome. Current models of host–parasite coevolution neglect possible asymmetries in competitive abilities among host genotypes carrying the same number of resistance genes. However, our experiment shows that the accumulation of mutations in loci not involved in interactions with the phages can lead to a drop in the immune diversity after a local extinction of the phage population. This drop in resistance diversity is likely to facilitate the evolutionary emergence of the phages when new viruses are introduced in the population[24,39,40]. We expect that this process may alter dramatically the coevolutionary dynamics studied with numerical simulations in ref. [38]. The collapse of the diversity of CRISPR immunity in the absence of phages (or when phages are very rare) would shorten the duration of periods where the host controls the phage population and would speed up the coevolutionary dynamics between phages and CRISPR immunity. At the larger spatial scale, this succession of local phage extinction and rapid recolonization could ensure the long-term coexistence of bacteria and phages in spatially structured environments.

Methods

Bacteria and bacteriophage strains. *S. thermophilus* DGCC 7710 and phage 2972 (ref. [41]) were obtained from the Félix d'Hérelle Reference Center for Bacterial Viruses (www.phage.ulaval.ca). Sixteen derivative phage-resistant strains, each with an unique CRISPR spacer were generated previously[20] and sequenced to look for mutations outside the CRISPR loci (Supplementary Table 1). Similarly, 16 phages carrying mutation to escape the resistance of these individual spacers were isolated after selection on each resistant bacteria (see the list of protospacer sequences in Supplementary Table 2)[20].

Experimental procedure. Before the experiment, the 17 bacterial strains were mixed and grown during 6 h in LM17 + $CaCl_2$ (37 g l^{-1} of M17 (Oxoid) supplemented with 5 g l^{-1} of lactose and 10 mM of $CaCl_2$). Then, the bacterial

mix was transferred 1:100 into 10 ml of fresh LM17 + $CaCl_2$ (no-phage treatment, seven replicates), infected with 10^5 wild-type 2972 phages (monomorphic phage treatment, eight replicates) or infected with the mix of 10^5 phages (polymorphic phage treatment, eight replicates), then incubated at 42 °C. We used only seven replicates in the control: because we were limited by the total number of replicates, we could sequence using the Nextera XT 96 samples prep kit (below). Every day (after 18 h of incubation), 1% of the cultures were transferred into 10 ml of LM17 + $CaCl_2$ and 10^5 phages were inoculated from the same population of phage (monomorphic or polymorphic) used at the beginning of the experiment. Following each transfer, the bacteria and phages from each replicate were separated by filtration (0.2 µm) and titrated as described in ref. [20]. To guarantee that there was enough DNA for sequencing, the phages were reamplified once on susceptible host bacteria (that is, DGCC 7710) over 5 h (after full lysis of the bacteria), then DNA was extracted using the ZYMO Quick-DNA Miniprep plus kit. Note that this amplification step may have introduced some bias in phage mutation frequencies if some phage genotypes were more fit than others in an environment with only susceptible hosts.

Bacteria sequencing. The CRISPR–Cas CR1 locus was amplified through PCR (primers 5′-3′: AGTAAGGATTGACAAGGACAGT; CCAATAGCTCCTCGTCATT) from the different populations from the three different treatments, the different replicates and the different time points. These PCR products were tagged using Nextera XT 96 samples prep kit and pooled before sequencing with Illumina MiSeq. The spacers were extracted from the sequences by searching for the flanking repeats allowing for a maximum of one mismatch. The spacers were then matched with their protospacers on the phage genome using Blast v.2.8.1 (ref. [42]) and the protospacer database presented in the next section. After these steps, an average sequencing depth of around 95,700 was obtained. A minimum identical word size of 10, and a 70% identity threshold was used. The top result of the search, if any, was used to replace the name of the spacer by the middle position of the protospacer in the phage genome. A frequency cut-off of 1% was used to optimize the quality of our dataset. The resulting frequencies of genotypes over time in each replicate are available in the Supplementary Information. We found that in the treatment with the monomorphic phage population there has been significantly more acquisition of spacers that were already present in the original 16 bacterial strains than the other 677 potential spacers (Chi-square test, $\chi^2 = 12.17$, degrees of freedom 1, $P = 4.8 \times 10^{-4}$). This means that the spacers already present in the mix were acquired preferentially, which may be because of DNA transfer among bacteria. The CRISPR–Cas CR3 locus was amplified through PCR (primers 5′-3′: GGTGACAGTCACATCTTGTCTAAAACG; GCTGGATATTCGTATAACATGTC) and migrated on 1.5% agarose gel to check for spacer acquisition. The samples with additional bands indicating the acquisition of an additional spacer are given in Supplementary Table 4.

Phage sequencing. The phage DNA samples were sequenced (Illumina MiSeq) with 150-bp paired-end reads. Trimmomatic[43] was used to clean and trim the sequencing reads yielding an average sequencing depth of around 650, before mapping them on the reference genome using Bowtie2 (ref. [44]). The software FreeBayes[45] was then used to detect single-nucleotide polymorphism and the phage reference genome[41] was updated to include the single-nucleotide polymorphism with a frequency >0.45 in the initial mix to distinguish these pre-existing mutations from the ones that arose during the experiment. The read mapping and the single-nucleotide polymorphism detection were done a second time using this updated genome as reference. The resulting frequencies of phage mutations over time in each replicate are available in the Supplementary Information. To detect the protospacers in the phage genome, we looked for the CR1 specific PAM sequence 'GGAA' or 'AGAA' in both strands of this reference genome and found 693 occurrences (281 and 412, respectively, for the two PAMs).

Fitness and adaptation estimates. We computed the mean phage fitness in a certain host population with equation (1). The frequencies of matching spacers and protospacer mutations are provided in the Supplementary Information. Our short-read sequencing data for the phages does not give linkage information between mutations so we need a linkage hypothesis to compute p_i from the frequencies of escape mutation derived from whole-genome sequencing of phage populations. When the host resistance genotype i carried more than a single spacer we assumed that the genotype frequency of the phage variant able to infect host resistance genotype i was the product of the frequencies of the mutations on all the protospacers targeted by this set of spacers (that is, we assumed linkage equilibrium among these mutations). To check the robustness of our results we computed phage fitness under the alternative assumption that escape mutations are fully linked (by setting to the frequency of phage mutations providing escape to the last spacer in the CRISPR locus). We observed a maximum of 2.7% difference between the measures of mean fitness of the phage in sympatric (same replicate) and contemporaneous (same time point) host populations under the two alternative assumptions for linkage. Hence, since linkage seems to have a limited effect in our analysis, all the results computed are derived under the assumption of no linkage.

Phage local adaptation was obtained for each replicate r at time t by computing the mean fitness of the phage on contemporaneous hosts (same time point t) from the same replicate r and by subtracting the mean fitness of the phage on contemporaneous hosts (same time point t) from all other replicates:

$$LA(r,t) = \sum_{i=1}^{n} h_i(r,t)\, p_i(r,t) - \frac{1}{n_r - 1} \sum_{j \neq r} \sum_{i=1}^{n} h_i(j,t)\, p_i(j,t) \qquad (2)$$

where $h_i(r,t)$ and $p_i(r,t)$ are the frequencies of host and phage genotypes in replicate r at time t and n_r is the number of replicates per treatment. Figure 6a shows phage local adaptation for different values of t after averaging over the $n_r = 8$ replicates for the monomorphic and the polymorphic phage treatments. The shaded areas present the 95% confidence interval after bootstrapping over replicates.

Phage temporal adaptation (TA) was obtained for each replicate r at time t by computing the mean fitness of the phage on hosts from the same replicate r but sampled at a different time point in the past or in the future (τ measures the time delay between bacteria and the phage: when $\tau > 0$ bacteria come from the future, when $\tau < 0$ bacteria come from the past). This measure was averaged over time t:

$$TA(\tau, r) = \frac{1}{n_t - |\tau|} \sum_{t=\max(0,-\tau)}^{\min(n_t, n_t-\tau)} \sum_{i=1}^{n} h_i(r, t+\tau)\, p_i(r,t) \qquad (3)$$

where n_t is the number of time points in the experiments, here $n_t = 5$ (that is, 0 to 4). Note that when we average over t we have to account for the fact that the number of elements we use for this calculation varies with τ. For instance, if $\tau = 0$ there are $n_t = 5$ points we can use (that is, the diagonal in Extended Data Fig. 8). In contrast, if $\tau = -4$ there is only one point (that is, the lower right corner in Extended Data Fig. 8). Hence, the number of elements in the sum over time in equation (3) is equal to $n_t - |\tau|$. In Fig. 6b we present the phage temporal adaptation for different values of τ after averaging over the $n_r = 8$ replicates for the monomorphic and the polymorphic phage treatments. The shaded areas present the 95% confidence interval after bootstrapping over replicates.

Differentiation measures. Jost's D for bacteria was computed on the CR1 locus according to Jost[46] with equation:

$$D = \frac{H_T - H_S}{1 - H_S} \frac{n_r}{n_r - 1} \qquad (4)$$

with n_r the number of replicates, H_T the mean heterozygosity of the pooled replicates and H_S the mean within-replicate heterozygosity, considering each different set of spacers a different genotype. Phage F_{ST}, Q_{ST} and bacteria F_{ST} was computed according to Weir and Cockerham[47] to take into account unequal sample sizes among treatments. For the Q_{ST}, which measures phenotypic rather than genetic differentiation, we pooled together phage mutations that led to the same phage phenotype, for example two mutations in the same protospacer, as a single phenotype. F_{ST} is the most usual measure of genetic differentiation, but D was computed too in Extended Data Fig. 4 as it better accounts for the change in the total number of resistances. Indeed, contrary to Jost's D, the value of the F_{ST} is heavily constrained by the range of genotype frequencies and particularly by the highest frequencies[48]. This property explains why the F_{ST} drops after day 2 in treatment C while D remains very high (Extended Data Fig. 4).

Statistical analysis. The 95% confidence intervals displayed on Fig. 6, Extended Data Figs. 4 and 6 were computed using a bootstrap approach, by resampling the data from the different replicates within a treatment 1,000 times.

In Fig. 3, the effect of treatment on bacteria diversity (that is, the effective number of host genotypes[22]: $1/(\sum_{i=1}^{n} h_i^2)$) was assessed for each day using an analysis of variance on the linear model: effective nb. of genotypes ~ treatment.

The linear regressions and the associated statistics for Fig. 5 and Extended Data Fig. 8 were computed using the SciPy[49] and statsmodel[50] Python packages. In Fig. 5, the statistical significance of the results was assessed by comparing separately each phage treatment to the treatment without phages. For each phage treatment, we built the following linear model: $\Delta h_i(t) \sim h_i(t) \times$ treatment, with $\Delta h_i(t) = h_i(t+1) - h_i(t)$, including the data from that treatment and the treatment without phages.

To demonstrate the presence of negative-frequency-dependent selection we tested the interaction term in the linear model (this measures the effect of phage infections on the effect of $h_i(t)$ on $\Delta h_i(t)$). This analysis confirmed the presence of negative-frequency-dependent selection induced by phages: the P values associated with the interaction term were 1×10^{-192} and 3×10^{-267} for the monomorphic and polymorphic phage treatments, respectively.

For all differentiation estimates (Extended Data Figs. 4 and 6), confidence intervals were generated with the Jackknife approach. This was done by computing the measures n_r times, each time leaving a different replicate out of the calculation[51]. The analysis and plotting were carried out using R v.3.6.3 (ref. [52]) and Python v.3.8.5 (ref. [53]).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The sequences of both phages and bacteria of this study have been deposited on National Center for Biotechnology Information under the BioProject of accession number PRJNA843584. Additional data such as the density measurements and the minimal dataset are available at https://zenodo.org/record/6646716.

Code availability

All codes used to process, analyse the data and make the figures are available at https://github.com/martingui/crispr_competition_coevolution.

Received: 22 November 2021; Accepted: 28 June 2022;
Published online: 15 August 2022

References

1. Ehrlich, P. R. & Raven, P. H. Butterflies and plants: a study in coevolution. *Evolution* **18**, 586–608 (1964).
2. Thompson, J. N. *The Coevolutionary Process* (Univ. Chicago Press, 2009).
3. Koskella, B. & Brockhurst, M. A. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.* **38**, 916–931 (2014).
4. Frank, S. Models of plant-pathogen coevolution. *Trends Genet.* **8**, 213–219 (1992).
5. Nuismer, S. *Introduction to Coevolutionary Theory* (Macmillan Higher Education, 2017).
6. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).
7. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanograph.* **45**, 1320–1328 (2000).
8. Winter, C., Bouvier, T., Weinbauer, M. G. & Thingstad, T. F. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the 'killing the winner' hypothesis revisited. *Microbiol. Mol. Biol. Rev.* **74**, 42–57 (2010).
9. Harcombe, W. & Bull, J. Impact of phages on two-species bacterial communities. *Appl. Environ. Microbiol.* **71**, 5254–5259 (2005).
10. Brockhurst, M. A., Fenton, A., Roulston, B. & Rainey, P. B. The impact of phages on interspecific competition in experimental populations of bacteria. *BMC Ecology* **6**, 19 (2006).
11. Alseth, E. O. et al. Bacterial biodiversity drives the evolution of CRISPR-based phage resistance. *Nature* **574**, 549–552 (2019).
12. Gómez, P. & Buckling, A. Bacteria-phage antagonistic coevolution in soil. *Science* **332**, 106–109 (2011).
13. Brockhurst, M. A. & Koskella, B. Experimental coevolution of species interactions. *Trends Ecol. Evol.* **28**, 367–375 (2013).
14. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
15. Horvath, P. et al. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1401–1412 (2008).
16. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317–327 (2010).
17. Hynes, A. P. et al. Detecting natural adaptation of the *Streptococcus thermophilus* CRISPR-Cas systems in research and classroom settings. *Nat. Protoc.* **12**, 547–565 (2017).
18. Deveau, H. et al. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).
19. Martel, B. & Moineau, S. CRISPR-Cas: an efficient tool for genome engineering of virulent bacteriophages. *Nucleic Acids Res.* **42**, 9504–9513 (2014).
20. Chabas, H. et al. Variability in the durability of CRISPR-cas immunity. *Philos. Trans. R. Soc. B.* **374**, 20180097 (2019).
21. Philippe, C. et al. A truncated anti-CRISPR protein prevents spacer acquisition but not interference. *Nat. Commun.* **13**, 1–8 (2022).
22. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Natl Acad. Sci. USA* **70**, 3321–3323 (1973).
23. Betts, A., Gray, C., Zelek, M., MacLean, R. & King, K. High parasite diversity accelerates host adaptation and diversification. *Science* **360**, 907–911 (2018).
24. van Houte, S. et al. The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* **532**, 385–388 (2016).
25. Barrangou, R. et al. Genomic impact of CRISPR immunization against bacteriophages. *Biochem. Soc. Trans.* **41**, 1383–1391 (2013).
26. Koskella, B. & Lively, C. M. Evidence for negative frequency-dependent selection during experimental coevolution of a freshwater snail and a sterilizing trematode. *Evolution: Int. J. Org. Evol.* **63**, 2213–2221 (2009).
27. Blanquart, F. & Gandon, S. Time-shift experiments and patterns of adaptation across time and space. *Ecol. Lett.* **16**, 31–38 (2013).
28. Gandon, S., Buckling, A., Decaestecker, E. & Day, T. Host–parasite coevolution and patterns of adaptation across time and space. *J. Evol. Biol.* **21**, 1861–1866 (2008).
29. Nourmohammad, A., Otwinowski, J. & Plotkin, J. B. Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. *PLoS Genet.* **12**, e1006171 (2016).
30. Childs, L. M., England, W. E., Young, M. J., Weitz, J. S. & Whitaker, R. J. CRISPR-induced distributed immunity in microbial populations. *PloS ONE* **9**, e101710 (2014).
31. Blanquart, F., Kaltz, O., Nuismer, S. L. & Gandon, S. A practical guide to measuring local adaptation. *Ecol. Lett.* **16**, 1195–1205 (2013).
32. Common, J., Walker-Sünderhauf, D., van Houte, S. & Westra, E. R. Diversity in CRISPR-based immunity protects susceptible genotypes by restricting phage spread and evolution. *J. Evol. Biol.* **33**, 1097–1108 (2020).
33. Common, J., Morley, D., Westra, E. R. & van Houte, S. CRISPR-cas immunity leads to a coevolutionary arms race between *Streptococcus thermophilus* and lytic phage. *Philo. Trans. R. Soc. B.* **374**, 20180098 (2019).
34. Paez-Espino, D. et al. CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *mBio* **6**, e00262–15 (2015).
35. Kassen, R. & Bataillon, T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nat. Genet.* **38**, 484–488 (2006).
36. Bataillon, T., Zhang, T. & Kassen, R. Cost of adaptation and fitness effects of beneficial mutations in *Pseudomonas fluorescens*. *Genetics* **189**, 939–949 (2011).
37. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat. Microbiology* **3**, 754–766 (2018).
38. Pilosof, S. et al. The network structure and eco-evolutionary dynamics of CRISPR-induced immune diversification. *Nat. Ecol. Evol.* **4**, 1650–1660 (2020).
39. King, K. & Lively, C. Does genetic diversity limit disease spread in natural host populations? *Heredity* **109**, 199–203 (2012).
40. Chabas, H. et al. Evolutionary emergence of infectious diseases in heterogeneous host populations. *PLoS Biol.* **16**, e2006738 (2018).
41. Lévesque, C. et al. Genomic organization and molecular analysis of virulent bacteriophage 2972 infecting an exopolysaccharide-producing *Streptococcus thermophilus* strain. *Appl. Environ. Microbiol.* **71**, 4057–4068 (2005).
42. Camacho, C. et al. Blast+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
43. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nat. Meth.* **9**, 357–359 (2012).
45. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at *arXiv* https://arxiv.org/abs/1207.3907 (2012).
46. Jost, L. Gst and its relatives do not measure differentiation. *Mol. Ecol.* **17**, 4015–4026 (2008).
47. Weir, B. S. & Cockerham, C. C. Estimating f-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
48. Jakobsson, M., Edge, M. D. & Rosenberg, N. A. The relationship between *f*st and the frequency of the most frequent allele. *Genetics* **193**, 515–528 (2013).
49. Virtanen, P. et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Meth.* **17**, 261–272 (2020).
50. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 92–96 (2010).
51. Efron, B. in *Breakthroughs in Statistics* (eds Kotz, S. & Johnson, N. L.) 569–593 (Springer, 1992).
52. R Core Team, R. C. et al. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014); https://www.R-project.org/
53. Van Rossum, G., Drake, F. L. et al. *Python Reference Manual* (Univ. Indiana, 2000).

Acknowledgements

Sequencing data were obtained through the genotyping and sequencing facilities of the Institut des Sciences de l'Evolution-Montpellier and Labex Centre Méditerranéen Environnement Biodiversité. We thank D. Tremblay, P.-L. Plante and G. Pageau for technical assistance during the sequencing of the bacterial strains. S.M. acknowledges funding from the Natural Sciences and Engineering Research Council of Canada (Discovery program). S.M. holds a T1 Canada Research Chair in Bacteriophages. H.C. was supported by an ETH Zurich Postdoctoral Fellowship. S.G. acknowledges support from a grant on 'Phylodynamics for experimentally evolving viruses' funded by the CNRS-MITI (Mission pour les Initiatives Transverses et Interdisciplinaires) and from the grant no. ANR-17-CE35-0012 from the Agence National de la Recherche.

Author contributions

S.G., A.N. and H.C. designed the experimental protocol. A.N. carried out the experiment and F.G. carried out phage sequencing. G.M.R. conducted the supplementary experiments for bacterial genomics. E.O.-A. helped with bioinformatics treatment of sequencing data and S.M. participated in the analysis. M.G., H.C., T.B., C.B. and

L.H. analysed the data. M.G. and S.G. wrote the manuscript. H.C., T.B. and S.M. revised the manuscript.

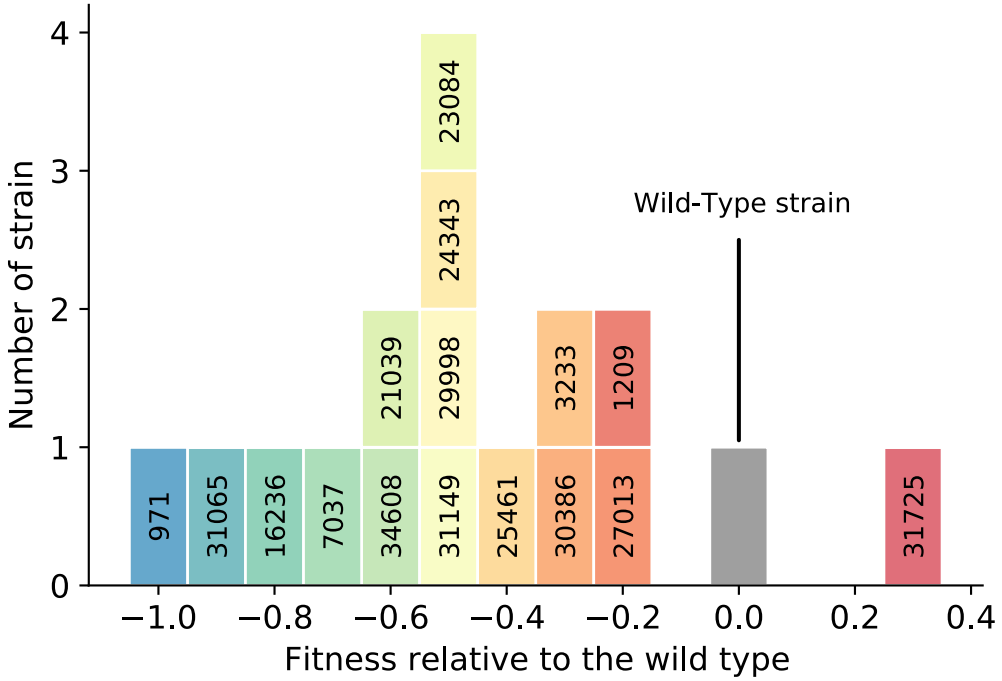
Competing interests
The authors declare no competing interests.

Additional information
Extended data are available for this paper at https://doi.org/10.1038/s41559-022-01841-9.

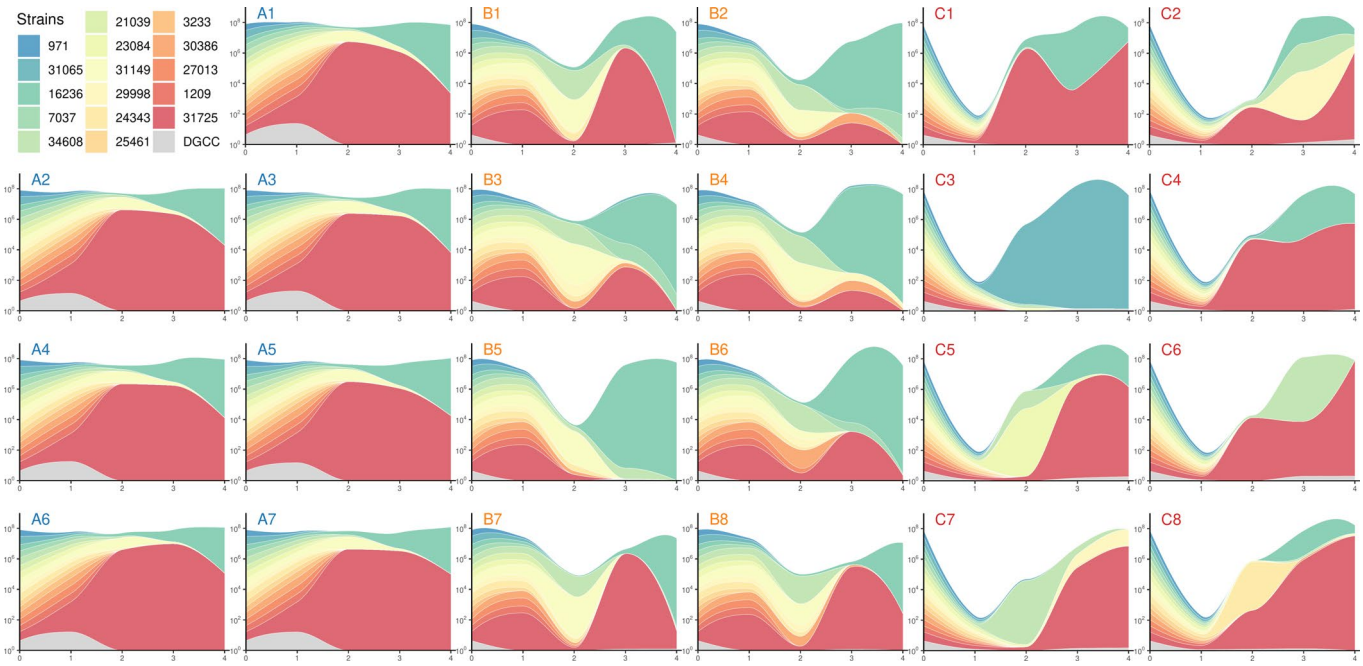
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-022-01841-9.

Correspondence and requests for materials should be addressed to Sylvain Gandon.

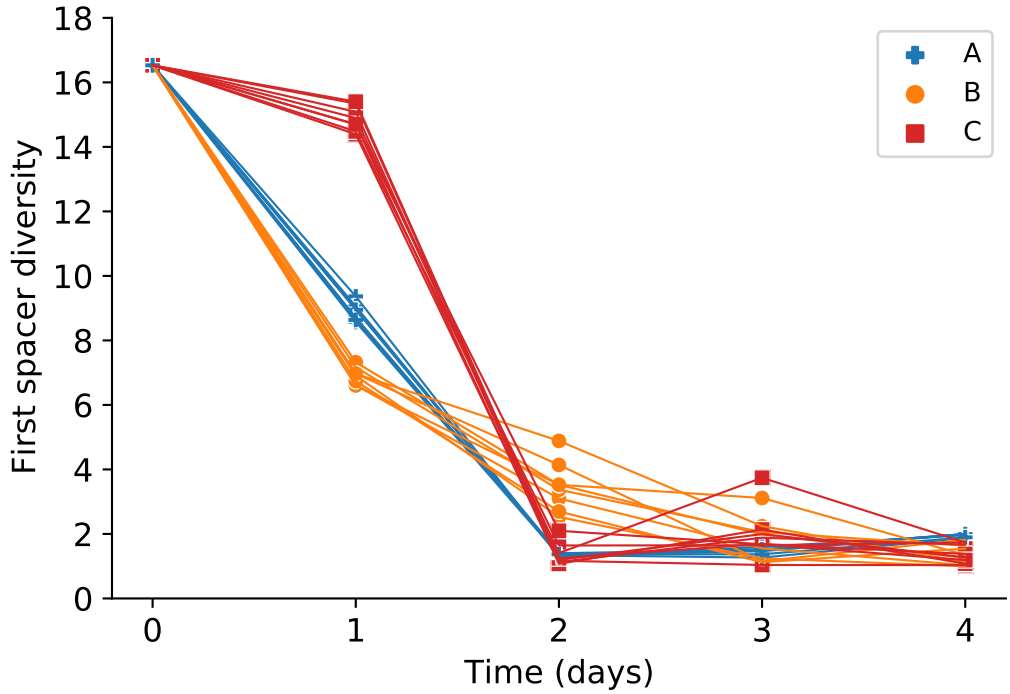
Peer review information *Nature Ecology & Evolution* thanks Adela Luján, Rachel Whitaker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



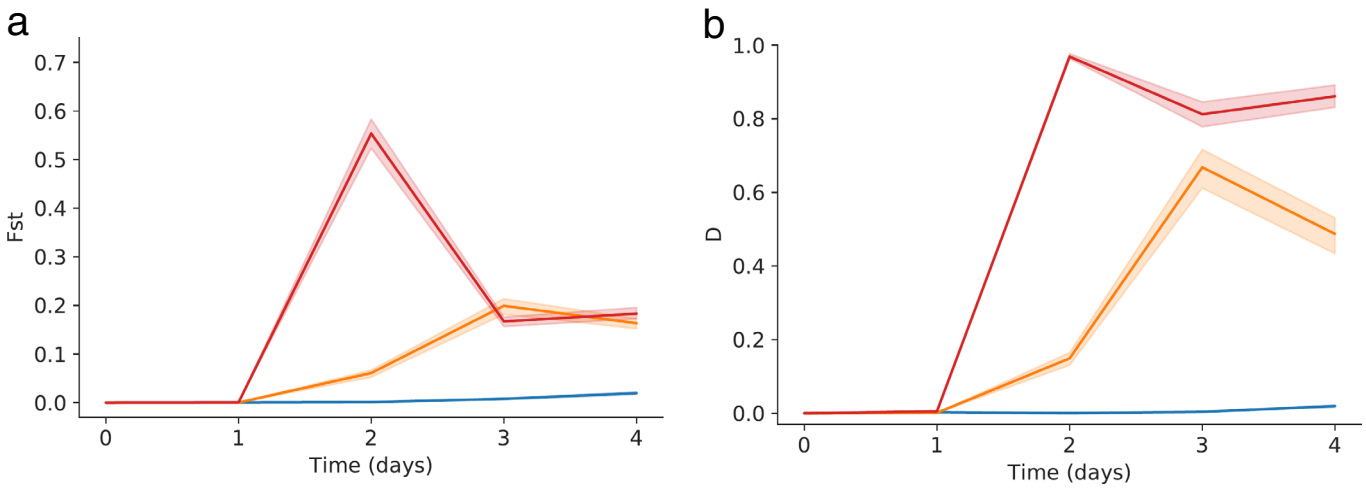
Extended Data Fig. 1 | Fitness distribution of the 16 resistant and the wild-type bacterial strains in the absence of the phage. The wild-type bacteria is shown in grey and the colors indicate the relative fitness of each of the 16 resistant strains. We used the same color code as the one used in Fig. 4. The fitness of strain i (relative to the wild-type wt) is computed with W_i-W_{wt}, where $W_i = \log_{10}(\frac{f_1(1-f_0)}{f_0(1-f_1)})$, f_0 and f_1 are the frequencies of strain i at day 0 and day 1, respectively. Hence, a positive (negative) value means that the strain grows faster (slower) than the wild-type at the beginning of the competition (in the first day of the experiment in treatment A).



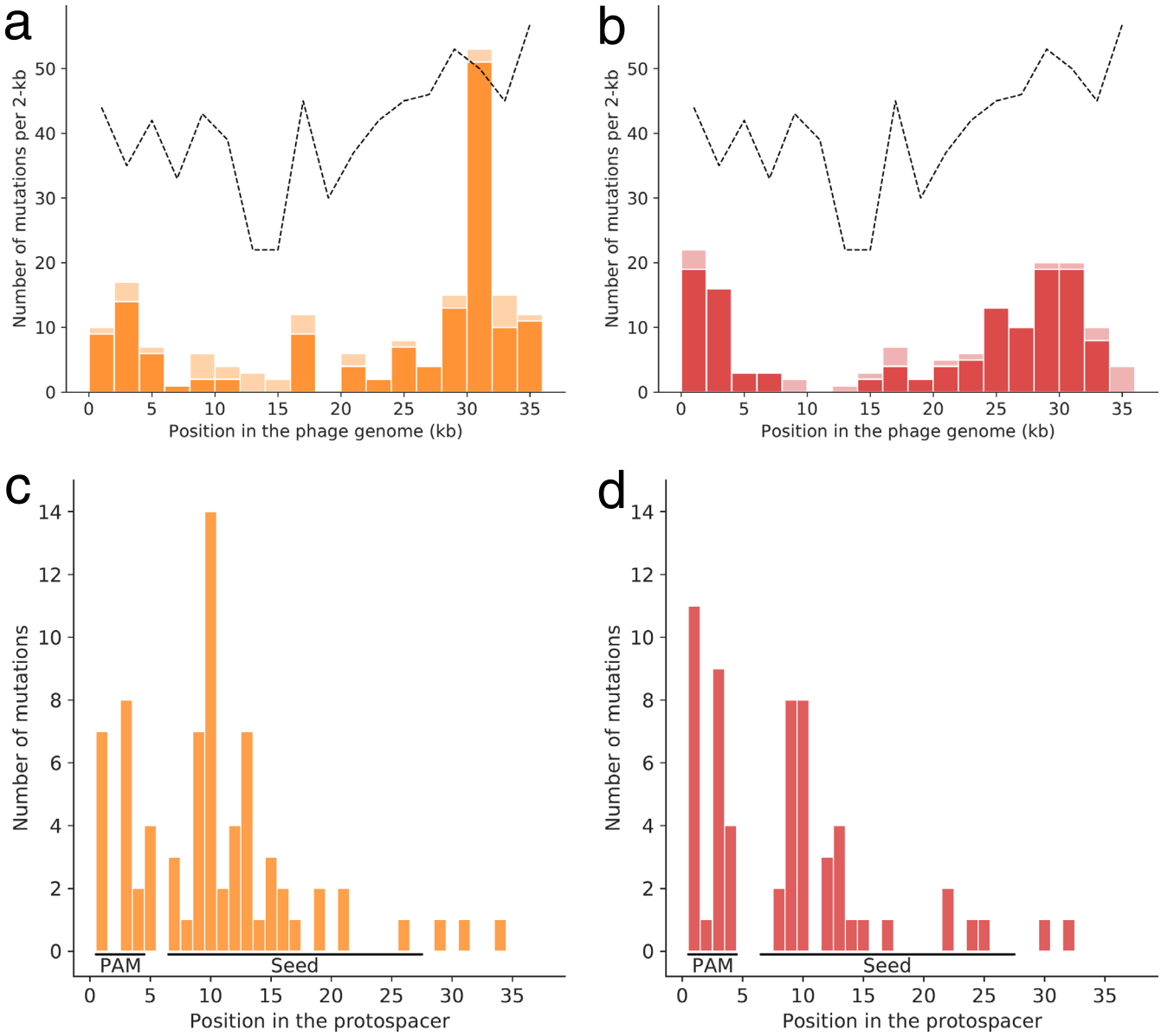
Extended Data Fig. 2 | Modified Muller plots of the bacterial populations based on the first spacer at the CR1 locus. Above each graph is the name of the replicate ('A' for the no phage control, 'B' for the monomorphic phage treatment and 'C' for the polymorphic phage treatment). The total height for each day shows the bacterial density (in cfu/ml) on a log scale, and the different colors show the proportion of the strains at each time point on a linear scale. The 17 strains that were added on day 0 (including the phage sensitive strain in grey) are shown in the legend (top-left corner). The blue-to-red color scale ranks the strains according to their initial fitness as detailed in Extended Data Figure 1. We used the same color code as the one used in Fig. 4. The lines are smoothed between each day.



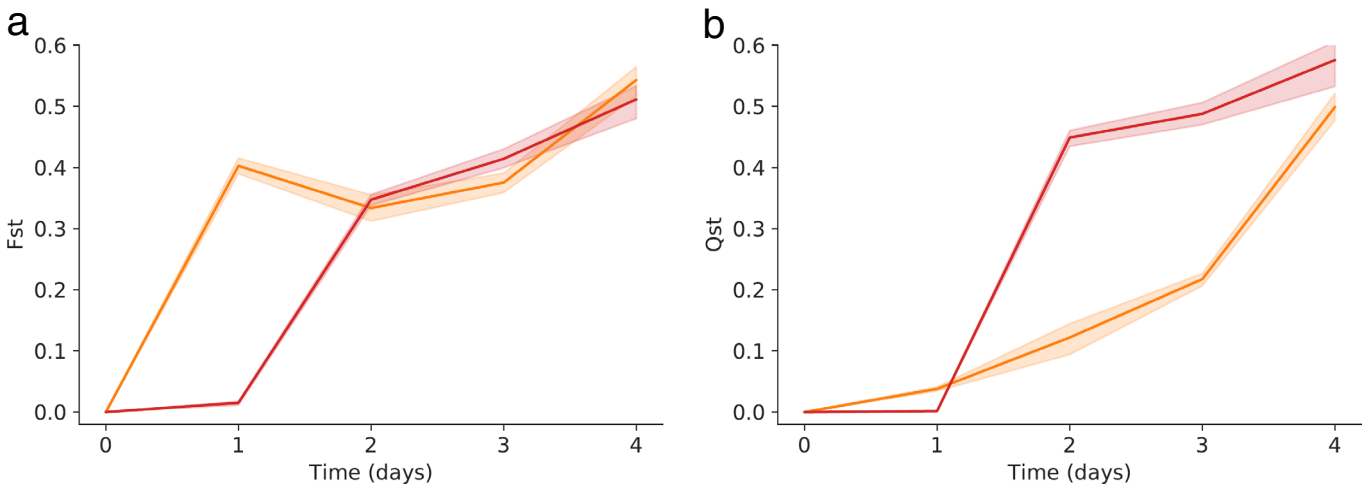
Extended Data Fig. 3 | Diversity of the first spacer of resistance in the bacterial population at the CR1 locus. The diversity is computed as the effective number of host genotypes using only the first spacer from the CR1 locus (compare with Fig. 3 where we used the whole array of new spacers on CR1). Blue points show the data in the absence of phages, orange and red show data for the monomorphic and polymorphic phage treatments, respectively.



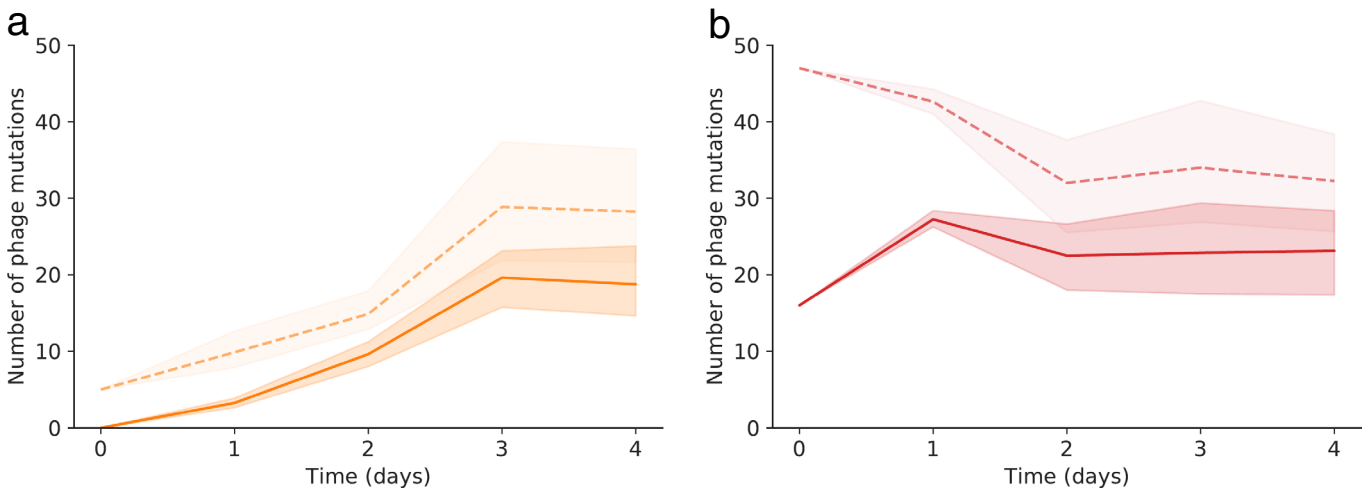
Extended Data Fig. 4 | Measure of the differentiation of bacterial population between replicates of the same treatment with (a) F_{ST} and (b) Jost's D (see Methods). As discussed by Jost[46], the D statistics may be a more relevant measure of differentiation when the total number of allele varies (see Methods). Blue curves show the values of differentiation in the absence of phages (treatment A), orange and red curves show the values of differentiation in the monomorphic (B) or the polymorphic (C) phage treatments, respectively. The shaded areas show the bootstrap 95% confidence interval and the center of the bands show the mean value.



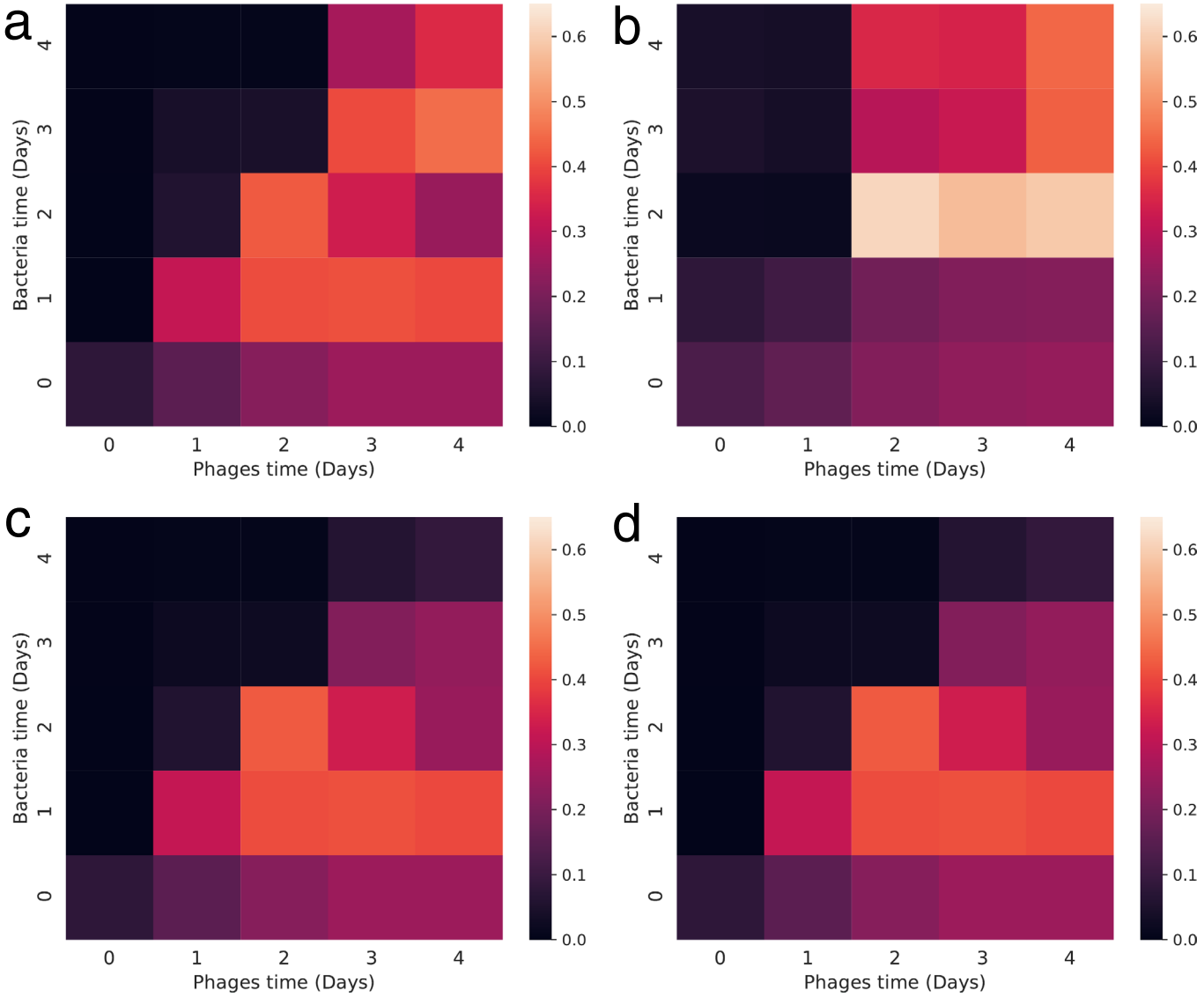
Extended Data Fig. 5 | Phage mutations in (a,c) the monomorphic and (b,d) the polymorphic phage treatments. The histograms (a,b) show the number of mutations per region of 2-kb in the phage genome. The light colors show mutations that are not located in a protospacer. The black dashed line shows the density of PAM in the genome. The positions of the phage mutations falling inside a protospacer are shown in panels (c,d). The mutations falling into two overlapping protospacers were discarded. The PAM and the seed region of the protospacer are shown.



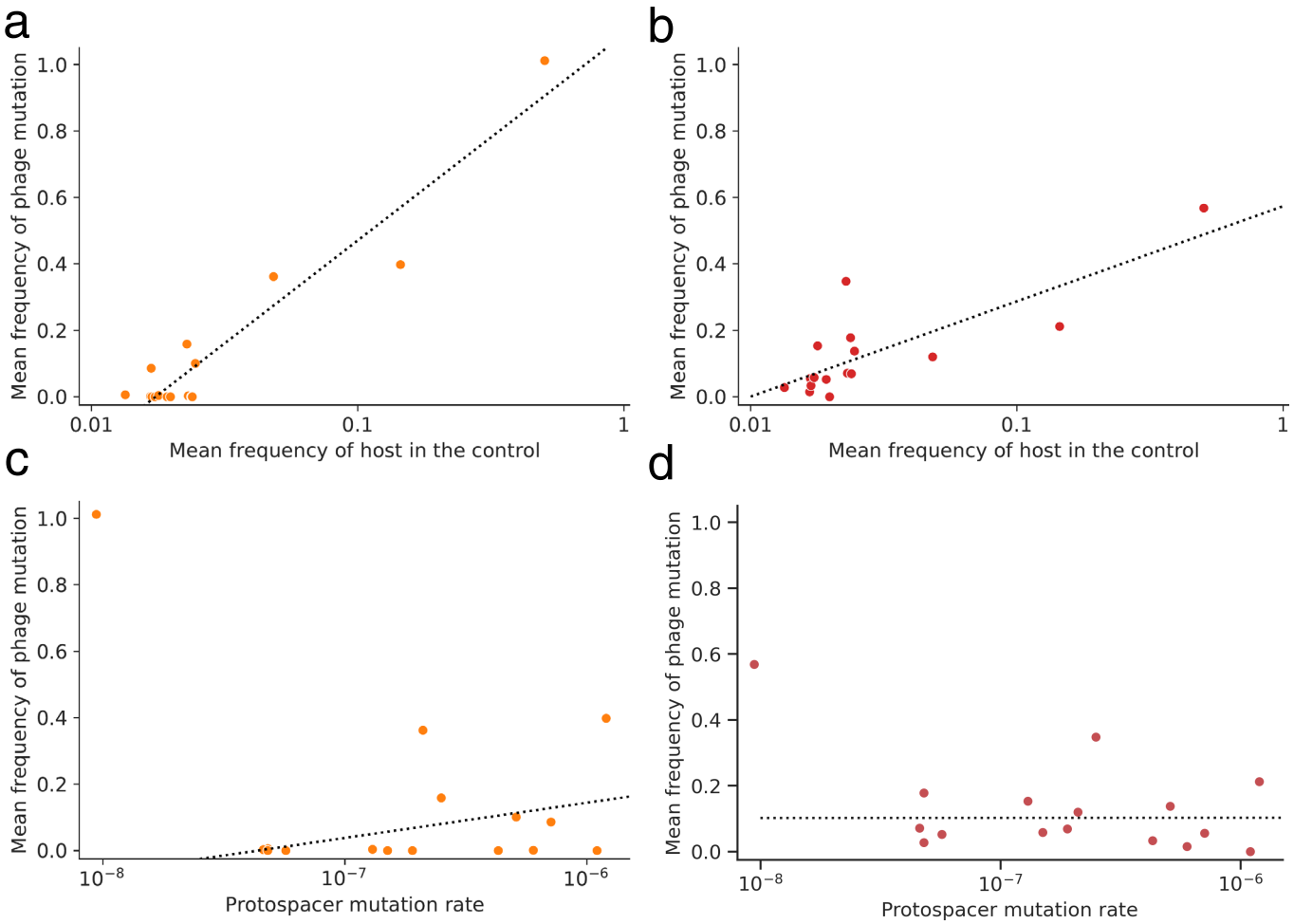
Extended Data Fig. 6 | Measure of phage differentiation among replicate populations of the same treatment using (a) F_{ST} and (b) Q_{ST} (see Methods). Orange and red curves show the level of differentiation for the monomorphic (treatment B) and the polymorphic (treatment C) phage treatments, respectively. The shaded areas show the bootstrap 95% confidence interval and the center of the bands show the mean value.



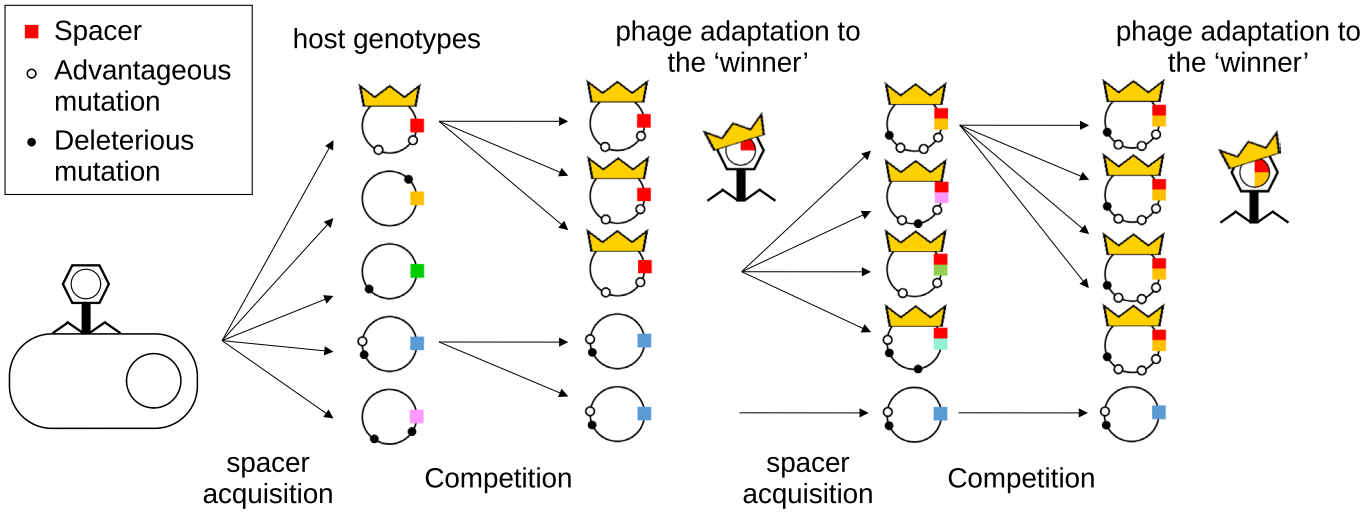
Extended Data Fig. 7 | Number of phage mutations through time in (a) the monomorphic and (b) polymorphic phage treatments. The plain line shows the mutation in protospacers, the dashed line shows all of the mutations. Only mutations with frequencies over 0.025 are kept. The shaded areas show the bootstrap 95% confidence interval and the center of the bands show the mean value.



Extended Data Fig. 8 | Phage fitness when confronting *in silico* phages and bacteria of each time points from the same replicate in the monomorphic (a,c) and polymorphic (b,d) phage treatments. The fitness was computed using equation (1). In panels c and d we try to correct the signal from the CR3 locus. To do this we selected all bacterial genotypes i with a frequency above 0.1 while the corresponding escape mutation i in the phage is at a frequency higher than 0.5. The fact that these host genotypes keep growing (that is their frequency remain > 0.1) even in the presence of escape phages indicates that these host genotypes probably carry an additional resistance on the CR3 locus (see also Table S4). If these host genotypes are resistant to these phages we can correct the measure of mean fitness using $h_i\ p_i = 0$ for these host genotypes and this yields figures (c) and (d). Note that this correction only affects measures of phage adaptation at late time points in the experiments (consistent with the emergence of CR3 resistance at the end of the experiment, Table S4).



Extended Data Fig. 9 | Phage mutation frequencies correlate with host frequency in the control but not with the protospacer mutation rate. There is one point for the protospacers targeted by each of the 16 different resistant strains. We show the mean frequency of escape mutations in each of the 16 protospacers (averaged over days 1 to 4 and over the eight replicates) against (a,b) the mean frequency of the corresponding host strain in the treatment without phages (averaged over days 1 to 4 and over the eight replicates) or (c,d) the protospacer mutation rates estimated by Chabas et al.[23]. The results are shown for the monomorphic phage treatment (a,c) and the polymorphic phage treatment (b,d). Log-linear regression lines (dashed lines) highlight the influence of strain frequencies on the frequencies of escape mutations in the phage population. In panels (c,d), the point on the upper left side was left out of the regression as it may be considered as an outlier (but this point is not left out of the Pearson's r calculation given in the main text).



Extended Data Fig. 10 | The 'royal family' model provides a conceptual framework to describe the coevolutionary dynamics in our experiment. First, selection imposed by phages leads to a diversification of CRISPR immunity. The competitive fitness of distinct resistant strains differ because they carry a variable number of beneficial and deleterious mutations (white and black dots on the bacterial chromosome, respectively). The resistant strain that carries the fewest number of deleterious mutations and the highest number of beneficial mutations is more competitive (that is, the winner in the 'kill-the-winner' hypothesis) and constitutes the 'royal family' (most future bacteria will derive from this strain). Second, the phage will preferentially adapt to this abundant strain. The acquisition of escape mutations in the phage genome will impose negative-frequency-dependent selection and will contribute to the maintenance of CRISPR diversity. Third, the 'royal family' strain will acquire new spacers and become abundant again. Competition will take place, phages will adapt to the 'royal family' again and this coevolutionary cycle will continue. Spacers and their corresponding escape mutations in the phage are indicated with the same colors. The 'royal families' of bacteria and phages are represented with a crown symbol.

Discussion

Life-history evolution

In the first chapter we have built a model that couples within-host epidemiological dynamics with the evolutionary dynamics of the pathogen. We adapted Fisher's Geometric Model (FGM) of adaptation to link pathogen phenotypes to transmission rate. We are able to derive the evolutionary dynamics of the distribution of phenotypes and of the mean transmission rate in a regime of Weak Selection Strong Mutation (WSSM), which we show can be relaxed to allow for stronger mutational effects. We use this modeling framework to investigate the possibility of lethal mutagenesis: driving viral populations to extinction through an increase in mutation rate. This higher mutation rate increases the influx of lethal mutations, which we interpret as an additional mortality term, and of non-lethal mutations, which have a negative mean effect on transmission rate and thus impart a mutational load: a cost on fitness of the variance in transmission rates in the pathogen population. In this work we show how mutations can lead to pathogen extinction but also to fitter pathogens. Contrary to many studies on lethal mutagenesis, we also consider beneficial mutations that increase transmission rate. We also incorporate the demographic feedback from the epidemiological dynamics, and show that much like the dynamics of the density of infected cells, the evolutionary dynamics is also scaled by the density of susceptible cells. This feedback could have a strong effect in scenarios of lethal mutagenesis, for when the infected cell population decreases, the susceptible population increases, which helps the pathogen population evade extinction.

Perspectives

Demographic stochasticity and evolutionary rescue

In the first chapter, we used a deterministic model to study the evolution of transmission rate and the possibility of lethal mutagenesis, i.e. driving the pathogen population to extinction through an increased mutation rate. In such a deterministic framework, we can follow the transient evolution of transmission but we can only detect extinction at the equilibrium state of the system. However demographic stochasticity could have a significant impact on our results. When studying the possibility of extinction, the most critical situations are when the pathogen population is very low, which coincides with situations where stochasticity will have the most significant effect. In our model, the pathogen population could have a very small equilibrium density without going extinct. This also prevents our model from being used in an evolutionary rescue setting. If the pathogen population is initially maladapted with a negative mean fitness, it will always reach an evolutionary equilibrium around the optimum of the FGM with any positive (potentially very low) value of mutation rate, no matter how low the density will get transiently. In these highlighted cases, demographic stochasticity would cause these populations to go extinct. This means that the deterministic critical mutation rates for lethal mutagenesis that we derive can be considered as upper bounds for the actual values.

To adress this issue, Anciaux et al. computed the fitness and demographic trajectories of a (non-pathogen) replicating population deterministically, and subsequently added demographic stochasticity with a noise term to the trajectory of the population density (Anciaux et al., 2019). In our work, we do not obtain explicit trajectories for the density of the infected population as it depends on the density of susceptible cells. To explore the effect of stochasticity, we are developing an individual-based simulation model based on a tau-leaping approach (Gillespie, 2001), where both the demographic dynamics and the mutations are modeled stochastically using Poisson distributions. With this approach, we hope to study the effect of epidemiological feedback on the probability of pathogen extinction or evolutionary rescue. We expect that this feedback will help the pathogen to avoid extinction in both scenarios of rescue or lethal mutagenesis, as when the pathogen population is decreasing, the susceptible population (which can be considered the resource of the pathogen) increases, which constitutes an additional rescue mechanism.

Joint evolution of transmission rate and virulence

We have studied adaptation through the increase in mean fitness of the population. In the experimental projects we have seen how this increase in fitness was due to the optimisation of several qualitative traits: the ability to escape the resistance from different distinct spacers in the host population. In the first chapter we studied theoretically the evolution of one single quantitative trait: the transmission rate. As described in the introduction, the evolution of pathogen transmission rate is often coupled with the evolution of another quantitative trait, the virulence, which is the additional host mortality caused by the infection. This question has often been described at equilibrium to find the Evolutionary Stable Strategy (ESS) defined as a strategy (here a certain value of transmission rate and virulence) that cannot be invaded by any other strategies (see Appendix C). It is possible to describe the dynamics of the evolution of both of these traits using a Price equation (Day, Parsons, et al., 2020; Price, 1970). Using such an equation, the change in transmission rate and virulence is dependent on the variance of each trait as well as their covariance. Nevertheless, we are not aware of any study in which the dynamics of these variances and covariance are explicitly modeled, and so the complete evolutionary dynamics of the joint distribution of transmission rates and virulence is not resolved.

We developed a Partial Derivative Equation (PDE) approach to model the dynamics of the distribution of transmission rates based on the work of Martin and Roques, 2016. We show in Appendix A how this approach can be extended to model the distributions of both of these traits at the same time, by introducing in our phenotype to life-history traits landscape an optimum for virulence (where it is minimized), distinct from the optimum of transmission rate (where it is maximised). With this framework, there is an emerging trade-off between the two traits as phenotypes cannot be simultaneously at the two optima. In our derivations, we recover a Price equation and also provide expressions for the dynamics of the variance and covariance, which allows us to draw a complete picture of the joint dynamics of adaptation of transmission rate and virulence. We find with our model that maladapted pathogens will first adapt quickly as the covariance between transmission and virulence is initially negative, but decreases and eventually becomes positive with adaptation.

In our modeling framework, we also couple the evolutionary with the epidemiological dynamics. This point is crucial as the density of susceptible hosts S can shape the direction of adaptation towards an optimisation of transmission rate or

virulence. The selective pressure on transmission rate is scaled by S whereas the selective pressure on virulence is independent from this density of hosts. Taking into account the effect of the epidemiological dynamics is particularly important for rapidly evolving pathogens such as viruses because evolutionary and epidemiological changes happen on the same timescale.

Modeling other quantitative traits

This modeling approach could be used to model the evolution of other traits as long as their contribution on fitness is linear. For instance it would be possible to study the joint evolution of a quantitative vaccine escape trait with virulence (which can be linked to transmission rate). The consequences of vaccination on the evolutionary dynamics of both vaccine escape (Gupta, Ferguson, and Anderson, 1997; Lipsitch, 1997; McLean, 1995; Restif and Grenfell, 2007) and virulence (André and Gandon, 2006; Gandon, M. Mackinnon, et al., 2003; Gandon, M. J. Mackinnon, et al., 2001; Van Boven et al., 2005) have been extensively studied individually. Yet the joint vaccine-induced evolution of these traits is rarely considered (Bernhauerová, 2016). Recently, this question has been tackled using a multi-locus adaptation model (McLeod and Gandon, 2022) which allowed the authors to investigate the effect of potential epistasis between the fitness effect of virulence and escape alleles. Although our model does not incorporate such epistatic interactions, it could be used to study the trajectories of vaccine-induced adaptation of escape and virulence as quantitative traits with these different vaccine types. A feature of our framework is that we can distinguish between the effects of direct and indirect selection on these traits. These two forces appear when we use a Price equation to model the evolution: a trait will change with a term depending on its variance (direct selection) and another term depending on the covariance with the second trait (indirect selection).

A major limitation of our work is the use of a certain mutational regime with Weak Selection and Strong Mutation (WSSM). This regime is adapted for traits that evolve in many multiple steps, and is for example unsuited for the modeling of qualitative trait like protospacer mutations which immediately grant escape to host resistance.

Another limitation is the shape of the distribution of phenotypes in the population, which we model with a single Gaussian distribution as a consequence of the WSSM regime. In particular this means that the distribution is unimodal. A possi-

ble perspective would be to extend this framework to model several sub-populations, which would each have a Gaussian distribution of phenotypes, thus resulting in a potentially multi-modal distribution at the scale of the whole population. Such an approach has been developed in a quantitative genetics framework, though without an explicit modeling of mutations, by (Lion, Sasaki, and Boots, 2022). With this approach which they call "oligomorphic dynamics", they track the dynamics of several distributions, which for example allows them to observe a whole population splitting into two diverging sub-populations when the fitness landscape causes disrupting selection. Such an approach could allow to track the emergence of several co-existing strategies when the environment is heterogeneous, for instance with different susceptible populations, vaccinated or not, with varied doses of drugs etc.

Escaping host resistance

In the second chapter, we have studied the probability of pathogen emergence according to the strategy of deployment of resistance on the host population. There are known results on the effect of diversity in limiting pathogen spread, however here we focused on the depth of resistance. It is also possible for hosts to be multi-resistant, meaning that pathogens need several escape mutations to successfully infect them. We have studied three strategies of resistance deployment in particular:
- A Mixing strategy where half of the resistant hosts carry the resistance A, and the other half carry the resistance B (which requires a different escape mutation to be escaped)
- A Pyramiding strategy where all hosts are double resistant AB, meaning that two distinct escape mutations are required for pathogens for successful infection.
- A Combining strategy where half of the resistant hosts are single resistant A or B, and the other half is double resistant AB
First, we find that larger inoculums increase the probability of emergence in two ways: by increasing the number of inoculated pre-existing mutants, and by causing a larger initial epidemic on the susceptible hosts, which allows for more replications and possibly more mutants. We make clear analytic predictions showing that pyramiding is the most effective strategy in order to prevent the emergence of pathogens as the necessary step of acquiring both escape mutations is limiting. In contrast, the mixing strategy is the most prone to pathogen emergence through adaptation. In this case, single escape mutations directly provide great fitness benefits and can spread in the population. The intermediate combining treatment is also intermedi-

ate in term of limiting emergence as only one of the two escape mutations (against the single-resistant host) is associated with a strong fitness benefit, but it provides a stepping stone which facilitates the acquisition of the other escape mutation to infect the double-resistant hosts.

We tested these predictions with the experimental system of CRISPR-resistant *Streptococcus thermophilus* and its virulent phage 2972. We confirmed the prediction on the effect of inoculum size and the hierarchy of treatments that the probability of emergence was higher in the mixing treatment, intermediate for one combining treatment and the lowest in the pyramiding treatment. However we found that the other combining treatment was not significantly different from the mixing treatment in terms of probability of evolutionary emergence. This could be due to a difference in mutation rate between the two protospacers of interest.

Perspectives

The dynamics of escape mutations with an heterogeneous host population

In Appendix B we describe an evolution experiment with the same bacteria and phage system. In this work, we wanted to monitor the evolutionary dynamics of the escape mutations after the initial emergence. We tested the effect on the evolutionary dynamics of the selection coefficient associated with the escape mutations, which we manipulated through differences in the initial frequencies of different hosts. We found that a higher frequency of hosts was associated with an increased frequency of the corresponding escape mutations early in the experiment, but this effect vanished in the later days. We also tested the effect of escape mutation rate by comparing the frequency of escape mutations in two groups of protospacers that differ in mutation rate. Contrary to our expectations, we did not find a significant effect of mutation rate early in the experiment, however we found an effect in the later days. We expected that mutation rate would be limiting early, but that once mutants arose in the population, it would not be impactful at all. We observed that phages carrying multiple escape mutations appeared early in the experiment, but that by the end of the experiment, the frequency of escape mutations were still increasing and we did not reach a point where all phages can infect all hosts. Therefore, we can understand that mutation rate can play a role later in the experiment as it can speed up the acquisition of additional escape mutations in phages already infecting other resistant hosts.

Interestingly, we find contrasting results with Chapter 2 where we predict that in the Mixing treatment, the probability of emergence is minimized when $f_A = 1/2$, i.e. when the two resistant hosts are present in equal frequencies. In Appendix B, we study the dynamics of escape frequencies after the emergence. In this case, we find that the overall frequency of escape mutations is higher when the resistant hosts are in equal frequencies (treatment B). Therefore heterogeneity in resistant hosts frequency could favor pathogen emergence, but limit the subsequent spread of escape mutants when emergence is achieved.

Multi-locus evolution of escape

We have studied the evolution of phage escape against a diversity of CRISPR-resistant hosts in two separate (co-)evolution experiments. We have seen in these conditions that phages acquired multiple escape mutations to be able to infect different hosts, but also multi-resistant hosts which had acquired additional spacers. We could observe the emergence of phages with several escape mutations as the total frequency of escape mutations could go above 1. In the coevolution experiment, we also observed the increase in frequency of escape mutation corresponding to the additional spacers of multi-resistant hosts. In this case, it is very likely that these 'secondary' escape mutations were present in the same phages that carried the 'primary' escape mutations. In other words, if a bacteria with spacer A1 acquired an additional spacer A2 and both escape mutations A1 and A2 are found in the phage population, they are very likely to be found in the same phages (at least in the case where there are no bacteria with only the A2 spacer). Thus we could find strong indications which supported hypotheses of linkage between escape mutations targeting spacers from the same bacterial CRISPR genotype. However in the case where there are distinct hosts with different spacers (eg. A and B), and the two corresponding escape mutations in the phage populations for example with each a frequency of 50% , we lack the information to differentiate between the scenarios:
- 50% of phages with both A and B escape mutations and 50% of phages without any of the two
- 50% of phages with escape mutation A and 50% of phages with escape mutation B
- all possible intermediate cases

This limitation stems from the fact that we used Illumina short reads to monitor evolutionary dynamics. This technology is well suited for the sequencing of the CRISPR locus of the bacterial hosts which contains all the spacers and thus all the

information relative to resistance, however in the phage, the protospacers targeted by CRISPR are scattered throughout the genome (34 kb for phage 2972 which we used). This leads to escape mutations in different protospacers being sequenced independently from each other in different reads, which does not keep the linkage information between mutations, i.e. whether they were found in the same phage genome.

To circumvent these limitations and study more rigorously multi-locus adaptation, we want to use a long read sequencing technology so that we could get sequencing reads that span over several protospacers, thus potentially containing escape mutation against different spacers. This would allow us to get frequencies of escape genotypes instead of simply independent mutations or small haplotypes. We would like to use the PacBio HiFi sequencing method, which can be used to produce reads between 15 and 20kb with an accuracy above 99.5% (Hon et al., 2020). However we face technical difficulties to get enough genetic material of good quality to sequence and consider several possibilities:
- Directly sequence the phage DNA extraction
- Use restriction enzymes to produce fragments of known lengths, one of which contains all protospacers
-Amplify the region of interest with all the protospacers using long-range PCR

We would like to use this sequencing technology to monitor the evolutionary dynamics in the same CRISPR-resistant bacteria and phage system. In Appendix B, we tried to limit host adaptation to follow the the dynamics of adaptation of phage escape in a controlled environment as constant as possible. In this project, we observed that the system had not reached an equilibrium and the frequency of escape mutations was still increasing by the end of the experiment. To get a complete picture of phage adaptation, we would like to carry out a longer experiment to hopefully reach a point where the phage population remains constant in terms of escape genotypes, where potentially the phage population is homogeneous and all phages harbor escape mutations against all present hosts, which is what seem to predict the results presented in Appendix B.

Coevolution

In the third chapter we studied a similar experiment but this time allowing for co-evolution of the bacteriophages and the bacterial hosts. We monitored both the evolution of phage escape mutations and the CRISPR locus of the bacteria. In all treatments, we used a mix of 16 resistant bacterial strains that differed in their CRISPR locus, and a fully susceptible wild-type strain. We observed in a treatment without phages that bacterial diversity was quickly lost due to intrinsic fitness differences between the different strains. Particularly, we found two strains which were more competitive and completely took over the the population in the 4 days of the experiment in all replicates. With such repeatable host dynamics in the absence of phages, we used our experiment to study the reciprocal effects of host competition on pathogen adaptation and vice versa. We found that through negative frequency-dependent selection (NFDS), phages limited the loss of host diversity. However this NFDS did not lead to a kill-the winner scenario (Thingstad, 2000; Weinbauer, 2004). What we found was that the most fit strains identified in the control outgrew the others initially that was quickly followed by the increase in frequency of the corresponding escape mutation in the phage population. Yet before going extinct, these initially more competitive strains repeatedly acquired new spacers of resistance and grew back in frequency. What we observed regarding the diversity of hosts was therefore due to phages generating diversity at the CRISPR locus of the already dominant strains, rather than a conservation of the initial diversity.

Using our system we are also able to track the dynamics of the mean fitness of the phage population, using both the frequency of host resistance genotypes and the frequency of escape mutations. We also compute the fitness of the phage population when confronted to contemporary hosts, or hosts from past and future time points thus mimicking time-shift experiments. We find that phages are the most fit against hosts from the near past, but this fitness quickly drops against hosts from the future. This highlights the adaptation of the phages to escape the spacers present in the host population, as well as the adaptation of the hosts that quickly acquire new spacers to resist the phages. We also find evidence of local adaptation of the bacteriophages by comparing the fitness of phage populations against hosts from the same replicate (sympatric) or against hosts from others replicates (allopatric).

Perspectives

CRISPR coevolution and stochasticity

In our bacteria-bacteriophages experiments, we have been able to draw general conclusions on the evolutionary dynamics of escape mutations, with or without limiting coevolution. We found general behaviours but we observed different trajectories in the different replicates. In particular, we specifically find differentiation of both phage and bacterial population from different replicates of the coevolution experiment. In contrast, we observed very repeatable dynamics driven by between-host competition in the control without phages. Adding phages generates on both the host and pathogen population a strong selective pressure for spacer acquisition on one side, and protospacer escape mutation on the other. These mutational events are essentially stochastic and rare yet can completely change the evolutionary and demographic trajectory of a host-pathogen system. In the treatments with phages, we observe in all but one replicate that the descendants of the most fit strains detected in the control still win the competition and make up most of the host population after a few days. In just one replicate (C3), we find that these most competitive strains apparently went extinct before the acquisition of new CRISPR spacers that could have granted resistance. We find that in this replicate the host population is completely dominated after a few days by a strain which was not particularly successful in any other replicate. This observation and the large differentiation we compute stem from stochastic events (or lack thereof), highlighting the relevance of stochasticity in understanding but also modeling these types of host-pathogen systems.

To explore the evolutionary dynamics of CRISPR coevolution we established a collaboration with visiting PhD student Armun Liaghat and Pr. Mercedes Pascual from the University of Chicago. They developed a stochastic model using a Gillespie approach for this system, close to the model presented in (Pilosof et al., 2020), which could be used to explore the long-term dynamics of coevolution which are more difficult to obtain experimentally. With our experimental data, we have a thorough description of the composition of both the host and pathogen populations through time, which can help to better adjust the model and its parameters to fit observations. Following our work on competition, we also introduced variability in host fitness in this model, which could be used to explore the long-term dynamics of coevolution which we did not obtain experimentally. Consistent with what we expected, we find using this model that an increased heterogeneity in host intrinsic

fitness was associated with an increase in the probability of viral escape because competitive asymmetry reduced host diversity. Additionally, we could study the impact of the value of certain critical parameters, for instance the rate of acquisition of new spacers: we observed in one out of sixteen replicates that the most fit strains did not acquire new spacers before extinction and so with a slightly lower rate of acquisition, the dynamics we observe might have been completely different without a lasting domination of the host lineages with the highest intrinsic fitness.

Conclusion

In this thesis we have studied different aspects of the dynamics of viral adaptation, with a variety of approaches, both theoretical and experimental. If the different chapters seem very distinct, like the theoretical model on the evolution of transmission rate and the coevolution experiment with CRISPR-resistant bacteria, in these projects the evolution of viruses is driven by the same forces, which we can sum up with the equation:

$$\Delta r = \Delta r_{ns} + \Delta r_m + \Delta r_{ec} \tag{27}$$

The dynamics of malthusian fitness, or growth rate, is driven by natural selection (Δr_{ns}), by the direct effect of mutation (Δr_m) and by changes in the environment (Δr_{ec}).

We have seen how natural selection can drive the evolution of quantitative traits towards an optimum, or select for qualitative traits like resistance escape. Natural selection operates on the variance in the population, which is generated by mutations. These mutations could provide escape to host resistance, or increase transmission rate but could also lead to extinction due to their being deleterious on average. Finally we have studied how biotic environmental change could impact viral evolution: through the density of susceptible cells, changes in the frequency of different hosts and even the appearance of new resistant hosts in coevolutionary scenarios.

We have showcased with both theoretical and experimental approaches the interplay between epidemiology and evolutionary dynamics, highlighting how these processes could happen on the same timescales. This thesis shows that all these processes must be jointly taken into account to better understand viral evolution

and potentially design better therapeutic approaches or disease management policies.

Evolutionary dynamics of life-history traits and applications for disease management

In this thesis we have modeled the evolutionary dynamics of transmission rate (and virulence) in a framework which allows for feedback between evolution and epidemiological dynamics. This feedback has been largely ignored with the more classic framework of R_0 maximisation. In the Adaptive Dynamics framework, ecological feedback are taken into account but an assumption was made that the time scale of evolutionary change was much higher than that of ecological change, and so the latter were considered immediate (Dieckmann, 2002). Additionally, in this framework, mutations are considered rare events and are therefore not modeled explicitly, and selection is fueled by a standing variance.

We have witnessed with the SARS-CoV-2 pandemic an example that this separation of timescale is not always well justified. We observed evolutionary changes with the appearance and subsequent increase in frequency of several new variants during the early phases of the pandemic, thus before any epidemiological equilibrium could have been reached as for instance the proportion of immunized people was still relatively low. This highlights the fact that epidemiological and evolutionary changes must be studied in concert at least for epidemics the scale of the SARS-CoV-2 pandemic. Another aspect that is lacking in understanding the evolution of this virus is the lack of prediction for the evolution of transmission rate or virulence. Without a specific mutational model and phenotypic landscape, it proved difficult to predict whether the initial variants would be associated with higher or lower transmission rate and/or virulence. Our approach using FGM allows us to propose such predictions, which could explain why initially a maladapted virus could evolve to both increase transmission rate and reduce virulence, before a trade-off is reached which limits the optimisation of both these of these traits simultaneously.

CRISPR host pathogen system: a model for epidemiology

In this thesis we have used the experimental system of *Streptococcus thermophilus* and its virulent phage 2972. With this system we have explored the coevolutionary dynamics of CRISPR resistance and escape. This system is worthy of interest on its

own as this bacteria is massively used in the dairy fermentation industry (Samson and Moineau, 2013). With the increasingly considered possibility of phage therapy – which consists in treating (multi drug-resistant) pathogenic bacteria with a selected cocktail of phages – it becomes increasingly important to study CRISPR and more generally bacteria-bacteriophage coevolution.

We advocate that this experimental system is also a suitable model for the study of the evolution of pathogen escape of host resistance in general. Bacteria can be resistant with one or potentially more CRISPR spacers and we know exactly the genetic determinism of escape mutations in the phages. We showed how it could be used to explore the efficacy of certain strategies of resistance deployment in the host population in limiting pathogen emergence. We believe that these conclusions could hold for a variety of systems for which experimental data is hard to obtain like crop-pathogen systems or a vaccinated human population. We also show that after pathogen emergence, we can monitor the dynamics of escape pathogens with sequencing while controlling to a certain extent the composition of the host population. Such an approach could be used to study dynamic resistance deployment strategies such as the progressive rollout of potentially several vaccines, similar to the scenario of the SARS-CoV-2 pandemic.

Bibliography

Alizon, S., A. Hurford, N. Mideo, and M. Van Baalen (2009). "Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future". In: *Journal of evolutionary biology* 22.2, pp. 245–259 (cit. on p. 10).

Anciaux, Y., A. Lambert, O. Ronce, L. Roques, and G. Martin (2019). "Population persistence under high mutation rate: from evolutionary rescue to lethal mutagenesis". In: *Evolution* 73.8, pp. 1517–1532 (cit. on p. 144).

Andersen, K. G., A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry (2020). "The proximal origin of SARS-CoV-2". In: *Nature medicine* 26.4, pp. 450–452 (cit. on p. 5).

Anderson, R. M. and R. M. May (1982). "Coevolution of hosts and parasites". In: *Parasitology* 85.2, pp. 411–426 (cit. on pp. 9, 10).

— (1992). *Infectious diseases of humans: dynamics and control*. Oxford university press (cit. on p. 6).

Andino, R. and E. Domingo (2015). "Viral quasispecies". In: *Virology* 479, pp. 46–51 (cit. on p. 14).

André, J.-B. and T. Day (2005). "The effect of disease life history on the evolutionary emergence of novel pathogens". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1575, pp. 1949–1956 (cit. on p. 23).

André, J.-B. and S. Gandon (2006). "Vaccination, within-host dynamics, and virulence evolution". In: *Evolution* 60.1, pp. 13–23 (cit. on p. 146).

Bernhauerová, V. (2016). "Vaccine-driven evolution of parasite virulence and immune evasion in age-structured population: the case of pertussis". In: *Theoretical Ecology* 9, pp. 431–442 (cit. on p. 146).

Bernheim, A. and R. Sorek (2020). "The pan-immune system of bacteria: antiviral defence as a community resource". In: *Nature Reviews Microbiology* 18.2, pp. 113–119 (cit. on p. 26).

Blanquart, F. and S. Gandon (2013). "Time-shift experiments and patterns of adaptation across time and space". In: *Ecology letters* 16.1, pp. 31–38 (cit. on p. 33).

Bonneaud, C. and B. Longdon (2020). "Emerging pathogen evolution: Using evolutionary theory to understand the fate of novel infectious pathogens". In: *EMBO reports* 21.9, e51374 (cit. on p. 10).

Brockhurst, M. A., T. Chapman, K. C. King, J. E. Mank, S. Paterson, and G. D. Hurst (2014). "Running with the Red Queen: the role of biotic conflicts in evolution". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1797, p. 20141382 (cit. on p. 35).

Buckling, A. and P. B. Rainey (2002). "Antagonistic coevolution between a bacterium and a bacteriophage". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269.1494, pp. 931–936 (cit. on p. 33).

Bull, J. J., R. H. Heineman, and C. O. Wilke (2011). "The phenotype-fitness map in experimental evolution of phages". In: *PLoS One* 6.11, e27796 (cit. on p. 26).

Bull, J. J., R. Sanjuan, and C. O. Wilke (2007). "Theory of lethal mutagenesis for viruses". In: *Journal of virology* 81.6, pp. 2930–2939 (cit. on pp. 20, 238).

Bull, J., M. Badgett, D. Rokyta, and I. Molineux (2003). "Experimental evolution yields hundreds of mutations in a functional viral genome". In: *Journal of molecular evolution* 57, pp. 241–248 (cit. on p. 26).

Bull, J., M. Badgett, and H. Wichman (2000). "Big-benefit mutations in a bacteriophage inhibited with heat". In: *Molecular biology and evolution* 17.6, pp. 942–950 (cit. on p. 26).

Burch, C. L. and L. Chao (1999). "Evolution by small steps and rugged landscapes in the RNA virus ϕ6". In: *Genetics* 151.3, pp. 921–927 (cit. on p. 26).

— (2004). "Epistasis and its relationship to canalization in the RNA virus φ6". In: *Genetics* 167.2, pp. 559–567 (cit. on p. 17).

Burch, C. L., S. Guyader, D. Samarov, and H. Shen (2007). "Experimental estimate of the abundance and effects of nearly neutral mutations in the RNA virus ϕ6". In: *Genetics* 176.1, pp. 467–476 (cit. on pp. 16, 26).

Burger, R. (1991). "Moments, cumulants, and polygenic dynamics". In: *Journal of mathematical biology* 30, pp. 199–213 (cit. on pp. 18, 21).

Bürger, R. (2000). *The mathematical theory of selection, recombination, and mutation*. John Wiley & Sons (cit. on p. 21).

Chabas, H., S. Lion, A. Nicot, S. Meaden, S. van Houte, S. Moineau, L. M. Wahl, E. R. Westra, and S. Gandon (2018). "Evolutionary emergence of infectious diseases in heterogeneous host populations". In: *PLoS biology* 16.9, e2006738 (cit. on pp. 23, 24, 34).

Childs, L. M., W. E. England, M. J. Young, J. S. Weitz, and R. J. Whitaker (2014). "CRISPR-induced distributed immunity in microbial populations". In: *PloS one* 9.7, e101710 (cit. on p. 35).

Clokie, M. R., A. D. Millard, A. V. Letarov, and S. Heaphy (2011). "Phages in nature". In: *Bacteriophage* 1.1, pp. 31–45 (cit. on p. 29).

Colavecchio, A., B. Cadieux, A. Lo, and L. D. Goodridge (2017). "Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne pathogens of the Enterobacteriaceae family–a review". In: *Frontiers in microbiology* 8, p. 1108 (cit. on p. 29).

Common, J., D. Morley, E. R. Westra, and S. van Houte (2019). "CRISPR-Cas immunity leads to a coevolutionary arms race between Streptococcus thermophilus and lytic phage". In: *Philosophical Transactions of the Royal Society B* 374.1772, p. 20180098 (cit. on p. 35).

Common, J., D. Walker-Sünderhauf, S. van Houte, and E. R. Westra (2020). "Diversity in CRISPR-based immunity protects susceptible genotypes by restricting phage spread and evolution". In: *Journal of evolutionary biology* 33.8, pp. 1097–1108 (cit. on p. 34).

Cuevas, J. M., R. Geller, R. Garijo, J. López-Aldeguer, and R. Sanjuán (2015). "Extremely high mutation rate of HIV-1 in vivo". In: *PLoS biology* 13.9, e1002251 (cit. on p. 14).

Day, T. and S. Gandon (2006). "Insights from Price's equation into evolutionary epidemiology". In: *Disease evolution: models, concepts, and data analyses* 71, pp. 23–44 (cit. on p. 239).

Day, T., T. Parsons, A. Lambert, and S. Gandon (2020). "The Price equation and evolutionary epidemiology". In: *Philosophical Transactions of the Royal Society B* 375.1797, p. 20190357 (cit. on pp. 145, 239).

Del Portillo, A., J. Tripodi, V. Najfeld, D. Wodarz, D. N. Levy, and B. K. Chen (2011). "Multiploid inheritance of HIV-1 during cell-to-cell infection". In: *Journal of virology* 85.14, pp. 7169–7176 (cit. on p. 15).

Deveau, H., R. Barrangou, J. E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath, and S. Moineau (2008). "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus". In: *Journal of bacteriology* 190.4, pp. 1390–1400 (cit. on p. 27).

Dieckmann, U. (2002). “Adaptive dynamics of pathogen-host interactions”. In: (cit. on p. 154).

Diekmann, O., H. Heesterbeek, and T. Britton (2013). *Mathematical tools for understanding infectious disease dynamics*. Vol. 7. Princeton University Press (cit. on p. 7).

Domingo, E. and C. Perales (2019). “Viral quasispecies”. In: *PLoS genetics* 15.10, e1008271 (cit. on p. 14).

Domingo, E., J. Sheldon, and C. Perales (2012). “Viral quasispecies evolution”. In: *Microbiology and Molecular Biology Reviews* 76.2, pp. 159–216 (cit. on p. 14).

Doron, S., S. Melamed, G. Ofir, A. Leavitt, A. Lopatina, M. Keren, G. Amitai, and R. Sorek (2018). “Systematic discovery of antiphage defense systems in the microbial pangenome”. In: *Science* 359.6379, eaar4120 (cit. on p. 26).

Ehrlich, P. R. and P. H. Raven (1964). “Butterflies and plants: a study in coevolution”. In: *Evolution*, pp. 586–608 (cit. on p. 30).

Eigen, M. (1993). “Viral quasispecies”. In: *Scientific American* 269.1, pp. 42–49 (cit. on p. 14).

Ellwanger, J. H. and J. A. B. Chies (2021). “Zoonotic spillover: Understanding basic aspects for better prevention”. In: *Genetics and Molecular Biology* 44 (cit. on p. 5).

Fisher, R. A. (1999). *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press (cit. on pp. 11, 17).

Frank, S. A. and M. Slatkin (1992). “Fisher’s fundamental theorem of natural selection”. In: *Trends in Ecology & Evolution* 7.3, pp. 92–95 (cit. on p. 13).

Gaba, S. and D. Ebert (2009). “Time-shift experiments as a tool to study antagonistic coevolution”. In: *Trends in Ecology & Evolution* 24.4, pp. 226–232 (cit. on p. 31).

Gandon, S., A. Buckling, E. Decaestecker, and T. Day (2008). “Host–parasite coevolution and patterns of adaptation across time and space”. In: *Journal of evolutionary biology* 21.6, pp. 1861–1866 (cit. on pp. 31, 33).

Gandon, S., M. E. Hochberg, R. D. Holt, and T. Day (2013). “What limits the evolutionary emergence of pathogens?” In: *Philosophical transactions of the Royal Society B: biological sciences* 368.1610, p. 20120086 (cit. on p. 23).

Gandon, S., M. Mackinnon, S. Nee, and A. Read (2003). “Imperfect vaccination: some epidemiological and evolutionary consequences”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270.1520, pp. 1129–1136 (cit. on p. 146).

Gandon, S., M. J. Mackinnon, S. Nee, and A. F. Read (2001). "Imperfect vaccines and the evolution of pathogen virulence". In: *Nature* 414.6865, pp. 751–756 (cit. on p. 146).

Gavrilets, S. (1997). "Coevolutionary chase in exploiter–victim systems with polygenic characters". In: *Journal of theoretical biology* 186.4, pp. 527–534 (cit. on p. 35).

Gillespie, D. T. (2001). "Approximate accelerated stochastic simulation of chemically reacting systems". In: *The Journal of chemical physics* 115.4, pp. 1716–1733 (cit. on p. 144).

Gupta, S., N. M. Ferguson, and R. M. Anderson (1997). "Vaccination and the population structure of antigenically diverse pathogens that exchange genetic material". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 264.1387, pp. 1435–1443 (cit. on p. 146).

Halligan, D. L. and P. D. Keightley (2009). "Spontaneous mutation accumulation studies in evolutionary genetics". In: *Annual Review of Ecology, Evolution, and Systematics* 40, pp. 151–172 (cit. on p. 15).

Hampton, H. G., B. N. Watson, and P. C. Fineran (2020). "The arms race between bacteria and their phage foes". In: *Nature* 577.7790, pp. 327–336 (cit. on p. 26).

Hon, T., K. Mars, G. Young, Y.-C. Tsai, J. W. Karalius, J. M. Landolin, N. Maurer, D. Kudrna, M. A. Hardigan, C. C. Steiner, et al. (2020). "Highly accurate long-read HiFi sequencing data for five complex genomes". In: *Scientific data* 7.1, p. 399 (cit. on p. 150).

Houte, S. van, A. K. Ekroth, J. M. Broniewski, H. Chabas, B. Ashby, J. Bondy-Denomy, S. Gandon, M. Boots, S. Paterson, A. Buckling, et al. (2016). "The diversity-generating benefits of a prokaryotic adaptive immune system". In: *Nature* 532.7599, pp. 385–388 (cit. on p. 34).

Janzen, D. H. et al. (1980). "When is it coevolution". In: *Evolution* 34.3, pp. 611–612 (cit. on pp. 29, 30).

Jung, A., R. Maier, J.-P. Vartanian, G. Bocharov, V. Jung, U. Fischer, E. Meese, S. Wain-Hobson, and A. Meyerhans (2002). "Multiply infected spleen cells in HIV patients". In: *Nature* 418.6894, pp. 144–144 (cit. on p. 15).

King, K. and C. Lively (2012). "Does genetic diversity limit disease spread in natural host populations?" In: *Heredity* 109.4, pp. 199–203 (cit. on p. 24).

Kopp, M. and S. Gavrilets (2006). "Multilocus genetics and the coevolution of quantitative traits". In: *Evolution* 60.7, pp. 1321–1336 (cit. on p. 35).

Koskella, B. (2014). "Bacteria-phage interactions across time and space: merging local adaptation and time-shift experiments to understand phage evolution". In: *The American Naturalist* 184.S1, S9–S21 (cit. on p. 34).

Labrie, S. J., J. E. Samson, and S. Moineau (2010). "Bacteriophage resistance mechanisms". In: *Nature Reviews Microbiology* 8.5, pp. 317–327 (cit. on p. 26).

Lande, R. and S. J. Arnold (1983). "The measurement of selection on correlated characters". In: *Evolution*, pp. 1210–1226 (cit. on p. 18).

Lauring, A. S. and R. Andino (2010). "Quasispecies theory and the behavior of RNA viruses". In: *PLoS pathogens* 6.7, e1001005 (cit. on p. 14).

Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler (1991). "Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations". In: *The American Naturalist* 138.6, pp. 1315–1341 (cit. on p. 16).

Lion, S. and J. A. Metz (2018). "Beyond R0 maximisation: on pathogen evolution and environmental dimensions". In: *Trends in ecology & evolution* 33.6, pp. 458–473 (cit. on p. 11).

Lion, S., A. Sasaki, and M. Boots (2022). "Extending eco-evolutionary theory with oligomorphic dynamics". In: *bioRxiv*, pp. 2022–12 (cit. on p. 147).

Lipsitch, M. (1997). "Vaccination against colonizing bacteria with multiple serotypes". In: *Proceedings of the National Academy of Sciences* 94.12, pp. 6571–6576 (cit. on p. 146).

Lively, C. M. (2010). "The effect of host genetic diversity on disease spread". In: *The American Naturalist* 175.6, E149–E152 (cit. on p. 24).

Lynch, M., R. Bürger, D. Butcher, and W. Gabriel (1993). "The mutational meltdown in asexual populations". In: *Journal of Heredity* 84.5, pp. 339–344 (cit. on pp. 20, 238).

Lynch, M. and W. Gabriel (1990). "Mutation load and the survival of small populations". In: *Evolution* 44.7, pp. 1725–1737 (cit. on p. 20).

Makarova, K. S., Y. I. Wolf, J. Iranzo, S. A. Shmakov, O. S. Alkhnbashi, S. J. Brouns, E. Charpentier, D. Cheng, D. H. Haft, P. Horvath, et al. (2020). "Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants". In: *Nature Reviews Microbiology* 18.2, pp. 67–83 (cit. on p. 27).

Makarova, K. S., Y. I. Wolf, S. Snir, and E. V. Koonin (2011). "Defense islands in bacterial and archaeal genomes and prediction of novel defense systems". In: *Journal of bacteriology* 193.21, pp. 6039–6056 (cit. on p. 26).

Martin, G. and S. Gandon (2010). "Lethal mutagenesis and evolutionary epidemiology". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1548, pp. 1953–1963 (cit. on pp. 20, 21, 238).

Martin, G. and T. Lenormand (2015). "The fitness effect of mutations across environments: Fisher's geometrical model with multiple optima". In: *Evolution* 69.6, pp. 1433–1447 (cit. on p. 21).

Martin, G. and L. Roques (2016). "The nonstationary dynamics of fitness distributions: asexual model with epistasis and standing variation". In: *Genetics* 204.4, pp. 1541–1558 (cit. on pp. 18, 21, 145, 239).

Matuszewski, S., L. Ormond, C. Bank, and J. D. Jensen (2017). "Two sides of the same coin: A population genetics perspective on lethal mutagenesis and mutational meltdown". In: *Virus evolution* 3.1, vex004 (cit. on p. 20).

May, R. M. and R. M. Anderson (1983). "Epidemiology and genetics in the coevolution of parasites and hosts". In: *Proceedings of the Royal society of London. Series B. Biological sciences* 219.1216, pp. 281–313 (cit. on p. 9).

McDonald, B. A. and C. Linde (2002). "Pathogen population genetics, evolutionary potential, and durable resistance". In: *Annual review of phytopathology* 40.1, pp. 349–379 (cit. on p. 26).

McLean, A. R. (1995). "Vaccination, evolution and changes in the efficacy of vaccines: a theoretical framework". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 261.1362, pp. 389–393 (cit. on p. 146).

McLeod, D. V. and S. Gandon (2022). "Effects of epistasis and recombination between vaccine-escape and virulence alleles on the dynamics of pathogen adaptation". In: *Nature ecology & evolution* 6.6, pp. 786–793 (cit. on p. 146).

Méthot, P.-O. (2012). "Why do Parasites Harm Their Host? On the Origin and Legacy of Theobald Smith's" Law of Declining Virulence"—1900-1980". In: *History and Philosophy of the Life Sciences*, pp. 561–601 (cit. on p. 9).

Miralles, R., P. J. Gerrish, A. Moya, and S. F. Elena (1999). "Clonal interference and the evolution of RNA viruses". In: *Science* 285.5434, pp. 1745–1747 (cit. on p. 16).

Morley, D., J. M. Broniewski, E. R. Westra, A. Buckling, and S. van Houte (2017). "Host diversity limits the evolution of parasite local adaptation". In: *Molecular ecology* 26.7, pp. 1756–1763 (cit. on p. 34).

Muller, H. J. (1927). "Artificial transmutation of the gene". In: *Science* 66.1699, pp. 84–87 (cit. on p. 15).

Muller, H. J. (1964). "The relation of recombination to mutational advance". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1.1, pp. 2–9 (cit. on p. 20).

Nachman, M. W. and S. L. Crowell (2000). "Estimate of the mutation rate per nucleotide in humans". In: *Genetics* 156.1, pp. 297–304 (cit. on p. 14).

O'Brien, S. J. and J. F. Evermann (1988). "Interactive influence of infectious disease and genetic diversity in natural populations". In: *Trends in Ecology & Evolution* 3.10, pp. 254–259 (cit. on p. 24).

Orr, H. A. (2003). "The distribution of fitness effects among beneficial mutations". In: *Genetics* 163.4, pp. 1519–1526 (cit. on p. 16).

— (2005). "The genetic theory of adaptation: a brief history". In: *Nature Reviews Genetics* 6.2, pp. 119–127 (cit. on p. 16).

Paez-Espino, D., I. Sharon, W. Morovic, B. Stahl, B. C. Thomas, R. Barrangou, and J. F. Banfield (2015). "CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus". In: *MBio* 6.2, e00262–15 (cit. on p. 35).

Pilosof, S., S. A. Alcalá-Corona, T. Wang, T. Kim, S. Maslov, R. Whitaker, and M. Pascual (2020). "The network structure and eco-evolutionary dynamics of CRISPR-induced immune diversification". In: *Nature Ecology & Evolution* 4.12, pp. 1650–1660 (cit. on p. 152).

Price, G. R. (1970). "Selection and covariance". In: *Nature* 227, pp. 520–521 (cit. on p. 145).

— (1972). "Fisher's 'fundamental theorem'made clear". In: *Annals of human genetics* 36.2, pp. 129–140 (cit. on p. 13).

Restif, O. and B. T. Grenfell (2007). "Vaccination and the dynamics of immune evasion". In: *Journal of the Royal Society Interface* 4.12, pp. 143–153 (cit. on p. 146).

Samson, J. E. and S. Moineau (2013). "Bacteriophages in food fermentations: new frontiers in a continuous arms race". In: *Annual review of food science and technology* 4, pp. 347–368 (cit. on pp. 155, 246).

Sanjuán, R. and P. Domingo-Calap (2016). "Mechanisms of viral mutation". In: *Cellular and molecular life sciences* 73, pp. 4433–4448 (cit. on p. 14).

Sanjuán, R., A. Moya, and S. F. Elena (2004a). "The contribution of epistasis to the architecture of fitness in an RNA virus". In: *Proceedings of the National Academy of Sciences* 101.43, pp. 15376–15379 (cit. on p. 17).

— (2004b). "The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus". In: *Proceedings of the National Academy of Sciences* 101.22, pp. 8396–8401 (cit. on pp. 16, 17).

Smith, E. C., N. R. Sexton, and M. R. Denison (2014). "Thinking outside the triangle: replication fidelity of the largest RNA viruses". In: *Annual Review of Virology* 1, pp. 111–132 (cit. on p. 14).

Sobesky, R., C. Feray, F. Rimlinger, N. Derian, A. Dos Santos, A.-M. Roque-Afonso, D. Samuel, C. Bréchot, and V. Thiers (2007). "Distinct hepatitis C virus core and F protein quasispecies in tumoral and nontumoral hepatocytes isolated via microdissection". In: *Hepatology* 46.6, pp. 1704–1712 (cit. on p. 15).

Suttle, C. A. (2005). "Viruses in the sea". In: *Nature* 437.7057, pp. 356–361 (cit. on p. 26).

Taylor, L. H., S. M. Latham, and M. E. Woolhouse (2001). "Risk factors for human disease emergence". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356.1411, pp. 983–989 (cit. on p. 5).

Tenaillon, O. (2014). "The utility of Fisher's geometric model in evolutionary genetics". In: *Annual review of ecology, evolution, and systematics* 45, pp. 179–201 (cit. on pp. 18, 238).

Thingstad, T. F. (2000). "Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems". In: *Limnology and Oceanography* 45.6, pp. 1320–1328 (cit. on pp. 36, 151, 243).

Van Boven, M., F. R. Mooi, J. F. Schellekens, H. E. de Melker, and M. Kretzschmar (2005). "Pathogen adaptation under imperfect vaccination: implications for pertussis". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1572, pp. 1617–1624 (cit. on p. 146).

Van Valen, L. (1973). "A new evolutionary law". In: *Evol theory* 1, pp. 1–30 (cit. on p. 29).

Weinbauer, M. G. (2004). "Ecology of prokaryotic viruses". In: *FEMS Microbiology Reviews* 28.2, pp. 127–181 (cit. on pp. 36, 151, 243).

Westra, E. R., S. van Houte, S. Oyesiku-Blakemore, B. Makin, J. M. Broniewski, A. Best, J. Bondy-Denomy, A. Davidson, M. Boots, and A. Buckling (2015). "Parasite exposure drives selective evolution of constitutive versus inducible defense". In: *Current biology* 25.8, pp. 1043–1049 (cit. on p. 34).

Westra, E. R. and B. R. Levin (2020). "It is unclear how important CRISPR-Cas systems are for protecting natural populations of bacteria against infections by mobile genetic elements". In: *Proceedings of the National Academy of Sciences* 117.45, pp. 27777–27785 (cit. on p. 35).

Wilke, C. O. (2005). "Quasispecies theory in the context of population genetics". In: *BMC evolutionary biology* 5.1, pp. 1–8 (cit. on p. 15).

Willett, B. J., J. Grove, O. A. MacLean, C. Wilkie, G. De Lorenzo, W. Furnon, D. Cantoni, S. Scott, N. Logan, S. Ashraf, et al. (2022). "SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway". In: *Nature microbiology* 7.8, pp. 1161–1179 (cit. on p. 24).

Wiser, M. J., N. Ribeck, and R. E. Lenski (2013). "Long-term dynamics of adaptation in asexual populations". In: *Science* 342.6164, pp. 1364–1367 (cit. on p. 16).

Woolhouse, M. E. and S. Gowtage-Sequeria (2005). "Host range and emerging and reemerging pathogens". In: *Emerging infectious diseases* 11.12, p. 1842 (cit. on p. 5).

Woolhouse, M. E., J. P. Webster, E. Domingo, B. Charlesworth, and B. R. Levin (2002). "Biological and biomedical implications of the co-evolution of pathogens and their hosts". In: *Nature genetics* 32.4, pp. 569–577 (cit. on pp. 29, 32, 33).

Contributions

Chapter 1

I built the CGF-based model, analyzed the results and wrote the manuscript.

Chapter 2

I built the statistical model and analyzed the experimental data.

Chapter 3

I processed the raw sequencing data, analyzed the data and wrote the manuscript.

Appendix A

I built the model, and wrote the manuscript.

Appendix B

I processed the raw sequencing data, analyzed the data and wrote the manuscript.

Appendix C

I co-wrote the document.

Appendix

Appendix A: Joint evolutionary dynamics of transmission and virulence with epidemiological feedback

Martin Guillemet, Guillaume Martin, Sylvain Gandon

This appendix contains an introduction and a derivation of the model.

1 Introduction

Pathogen virulence, like any other life-history trait, can evolve in response to the action of natural selection. A classical example of virulence evolution is the evolution of myxomatosis that has been monitored after the introduction of this virus in Australia to control rabbit population (Fenner and Marshall, 1957). The study of the evolution of pathogen virulence has led to development of many mathematical models. Most of these models rely on the assumption that virulence is costly for the pathogen because host death reduces the duration of infection and, consequently, it reduces the time during which the pathogen can be transmitted. Yet, virulence can be selected if virulence is associated with other life history traits like higher transmission rates or lower recovery rates. In these situations, virulence can be selected *indirectly* via its link with other pathogen traits. A classical way to model this link is to assume a trade-off function that relates virulence and transmission (Anderson and May, 1982, 1991). Under this assumption, one can identify the Evolutionary Stable virulence strategy within the Adaptive Dynamics framework (AD book on infectious diseases). This framework assumes a separation of time scales between epidemiology and evolution and relies on the assumption that the genetic variance of the pathogen population is minimal.

With higher genetic variance, the speed of adaptation can be faster and in this case one can study the joint evolutionary and epidemiological dynamics during an epidemic. These models show that life-history evolution is governed by the gradient of selection and the genetic variance-covariance matrix G (Lande, 1982; Roff, 1993). Both selection and the G matrix are dynamical variables. In particular, the G matrix is expected to vary with the action of natural selection and the influx of mutations.

Here we developed a theoretical framework to study the joint evolution of pathogen virulence and transmission. We extend the model developed in

the first chapter of this thesis on the evolution of pathogen virulence to account for the fact that pathogen fitness depends also on pathogen virulence. This analysis allows us to model the evolution of the G matrix together with epidemiological dynamics and the dynamics of mean life-history traits. This analysis might be particularly relevant to understand the evolution of newly emerged pathogens. Those pathogens are expected to be far from their optimal virulence and their optimal transmission rate. One may thus expect no trade-off between these two traits and our model could provide a way to understand the transient phase of pathogen adaptation.

.

2 Model

We build a model describing the within-host dynamics of adaptation of a pathogen to a population of naive hosts. We suppose that pathogens can differ by their phenotype $\mathbf{x}$, upon which depend both the transmission rate and the virulence (additional mortality).

We use the framework of Fisher's Geometric Model (FGM), and we consider that a phenotype is a vector of n phenotypic quantitative traits. Note that a bold notation such as $\mathbf{x}$ refers to vectors. We assume that transmission rate $\beta_{\mathbf{x}}$ and virulence $\alpha_{\mathbf{x}}$ vary through a Gaussian transmission function with the distance of the phenotypes to the distinct respective optima of both traits $\mathbf{O}_\beta$ and $\mathbf{O}_\alpha$. Without loss of generality, we assume that the optimum phenotype in regard to transmission rate (where transmission rate is maximized) is at the origin $\mathbf{O}_\beta = \{0, 0, ..., 0\}$. The optimum phenotype in regard to virulence (where virulence is minimized) is at a distance D. For the sake of mathematical tractability, we approximate the Gaussian transmission function linking phenotype to life-history traits as a quadratic function. This approximation is adapted when distances to the optima are small, which no-

tably prevents negative values of transmission rate. The phenotype to life history-traits transmission functions are thus:

$$\beta_{\mathbf{x}} = \beta_0 - \frac{\|\mathbf{x}\|^2}{2s_\beta} \tag{1}$$

$$\alpha_{\mathbf{x}} = \alpha_0 + \frac{\|\mathbf{x} - \mathbf{O}_\alpha\|^2}{2s_\alpha} \tag{2}$$

Where s_β and s_α are parameters that govern the steepness of the landscape for respectively for transmission rate and virulence. To ease the calculations in building the model, we will use a new parameter for virulence $\delta = -\alpha$ so that fitness is an increasing function of δ, and the phenotype to life-history trait landscape has the same as that of transmission rate (concave, with a maximum at the optimal phenotype $\mathbf{O}_\delta = \mathbf{O}_\alpha$) which gives:

$$\delta_{\mathbf{x}} = \delta_0 - \frac{\|\mathbf{x} - \mathbf{O}_\delta\|^2}{2s_\delta} \tag{3}$$

with $\delta_0 = -\alpha_0$ and $s_\delta = s_\alpha$. The dynamics of the density of infected hosts of phenotype $\mathbf{x}$ are then:

$$\dot{I}_{\mathbf{x}} = I_{\mathbf{x}}\big(\beta_{\mathbf{x}} S + \delta_{\mathbf{x}} - d\big) \tag{4}$$

Note that we drop the dependence on time for clarity. By introducing the total number of infected cells $I = \int I_{\mathbf{x}} dx$, the frequency of of phenotype $\mathbf{x}$ in the infected population $p_{\mathbf{x}} = I_{\mathbf{x}}/I$, the mean transmission rate $\overline{\beta} = \int p_{\mathbf{x}} \beta_{\mathbf{x}} dx$ and $\overline{\delta} = \int p_{\mathbf{x}} \delta_{\mathbf{x}} dx$, we can write the following system to describe the epidemiological dynamics:

$$\begin{aligned} \dot{S} &= b - \overline{\beta} SI - Sd \\ \dot{I} &= I\big(\overline{\beta} S + \overline{\delta} - d\big) \end{aligned} \tag{5}$$

We make the the assumption that the underlying distribution of phenotypes $\mathbf{x}$ is gaussian and with an equal variance in all directions of the pheno-

typic space, i.e. it follows a multivariate normal distribution $\mathcal{N}(\overline{\mathbf{x}}(t), I_n V_x(t))$ where $\overline{\mathbf{x}}(t)$ is the mean phenotype and $V_x(t)$ is the phenotypic variance at time t. We can write the mean transmission rate and the mean virulence as a function of these phenotypic variables so that:

$$\begin{aligned} \overline{\beta}(t) &= \beta_{\overline{\mathbf{x}}}(t) - \frac{nV_x(t)}{2s_\beta} \\ \overline{\delta}(t) &= \delta_{\overline{\mathbf{x}}}(t) - \frac{nV_x(t)}{2s_\delta} \end{aligned} \tag{6}$$

Where $\beta_{\overline{\mathbf{x}}}(t)$ and $\delta_{\overline{\mathbf{x}}}(t)$ are respectively the transmission rate and virulence of the mean phenotype $\overline{\mathbf{x}}$. The mean life history traits in the population are thus function of the distances of the mean phenotype to the two optima through $\beta_{\overline{\mathbf{x}}}(t)$ and $\delta_{\overline{\mathbf{x}}}(t)$ and the phenotypic variance V_x.

In this work, we use Generating Functions to describe the distribution of life-history traits. We define the bivariate density Moment Generating Function (dMGF) $M_t(z_1, z_2)$ and density Cumulant Generating Function (dCGF) $C_t(z_1, z_2)$ for transmission and virulence:

$$\begin{aligned} M_t(z_1, z_2) &= \int I_{\mathbf{x}} e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n\mathbf{x} \\ C_t(z) &= Log\big(M_t(z_1, z_2)\big) \end{aligned} \tag{7}$$

These differ from usual CGF or MGF because they are computed using the density $I_{\mathbf{x}}$ instead of the frequency $p_{\mathbf{x}}$ of each phenotype. A direct consequence of this is that setting $z = 0$ yields:

$$\begin{aligned} M_t(0, 0) &= I \\ C_t(0, 0) &= \log(I) \end{aligned} \tag{8}$$

With the dMGF one can easily generate the moments and cumulants of the distribution of β and δ by taking the derivatives in z_1 and z_2 and setting these parameters to zero:

$$
\begin{aligned}
\partial_{z_1} M_t(z_1, z_2) &= \int I_{\mathbf{x}} \beta_{\mathbf{x}} e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n \mathbf{x} \\
\partial_{z_1} M_t(0, 0) &= \int I_{\mathbf{x}} \beta_{\mathbf{x}} d^n \mathbf{x} = I\overline{\beta} \\
\partial_{z_2} M_t(z_1, z_2) &= \int I_{\mathbf{x}} \delta_{\mathbf{x}} e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n \mathbf{x} \\
\partial_{z_2} M_t(0, 0) &= \int I_{\mathbf{x}} \delta_{\mathbf{x}} d^n \mathbf{x} = I\overline{\delta}
\end{aligned}
\tag{9}
$$

and similarly with the dCGF:

$$
\begin{aligned}
\partial_{z_1} C_t(z_1, z_2) &= \frac{\int I_{\mathbf{x}} \beta_{\mathbf{x}} e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n \mathbf{x}}{\int I_{\mathbf{x}} e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n \mathbf{x}} = \frac{\partial_{z_1} M_t(z_1, z_2)}{M_t(z_1, z_2)} \\
\partial_{z_1} C_t(0, 0) &= \frac{\int I_{\mathbf{x}} \beta_{\mathbf{x}} d^n \mathbf{x}}{I} = \overline{\beta} \\
\partial_{z_2} C_t(z_1, z_2) &= \frac{\int I_{\mathbf{x}} \delta_{\mathbf{x}} e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n \mathbf{x}}{\int I_{\mathbf{x}} e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n \mathbf{x}} = \frac{\partial_{z_2} M_t(z_1, z_2)}{M_t(z_1, z_2)} \\
\partial_{z_2} C_t(0, 0) &= \frac{\int I_{\mathbf{x}} \delta_{\mathbf{x}} d^n \mathbf{x}}{I} = \overline{\delta}
\end{aligned}
\tag{10}
$$

One can check that taking the kth derivative of the dCGF according to z_1 (resp. z_2) and setting $z_1 = z_2 = 0$ will yield the kth cumulant of the distribution of β (resp. δ), and particularly:

$$
\begin{aligned}
\partial_{z_1} \partial_{z_1} C_t(0, 0) &= V_\beta \\
\partial_{z_2} \partial_{z_2} C_t(0, 0) &= V_\delta
\end{aligned}
\tag{11}
$$

Where the notations V_β and V_δ refer to the variance of the distribution of both distributions. Using the dCGF of the joint distribution of transmission rate and virulence, it is also possible to differentiate according to the two parmeters z_1 and z_2 at the same time to access cumulants of the joint distribution of β and δ such that:

$$
\partial_{z_1, z_2} C_t(0, 0) = \overline{\beta\delta} - \overline{\beta}\,\overline{\delta} = \mathrm{Cov}(\beta, \delta) \tag{12}
$$

where $\mathrm{Cov}(\beta, \delta) = \mathbb{E}\left[(\beta - \overline{\beta})(\delta - \overline{\delta})\right]$ is the covariance of the distributions of β and δ.

Using equation (4), we can write the time partial derivative of the dCGF of the joint distribution such that:

$$\begin{aligned}\partial_t M_t(z_1, z_2) &= \int \dot{I}_{\mathbf{x}} e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n\mathbf{x} \\ &= \int I_{\mathbf{x}}\big(\beta_{\mathbf{x}} S + \delta_{\mathbf{x}} - d\big) e^{\beta_{\mathbf{x}} z_1 + \delta_{\mathbf{x}} z_2} d^n\mathbf{x} \\ &= S\partial_{z_1} M_t(z_1, z_2) + \partial_{z_2} M_t(z_1, z_2) - d M_t(z_1, z_2)\end{aligned} \tag{13}$$

From which follows the time partial derivative of the dCGF:

$$\begin{aligned}\partial_t C_t(z_1, z_2) = \frac{\partial_t M_t(z_1, z_2)}{M_t(z_1, z_2)} &= S\frac{\partial_{z_1} M_t(z_1, z_2)}{M_t(z)} + \frac{\partial_{z_2} M_t(z_1, z_2)}{M_t(z)} - d \\ &= S\,\partial_{z_1} C_t(z_1, z_2) + \partial_{z_2} C_t(z_1, z_2) - d\end{aligned} \tag{14}$$

This partial derivative equation completely describes the dynamics of the distribution of β and δ in the absence of mutations. Taking the partial derivative according to z_1 (resp. z_2) and setting $z_1 = z_2 = 0$, we can directly recover the dynamics of the mean transmission rate and virulence:

$$\begin{aligned}\partial_t \partial_{z_1} C_t(0,0) &= S\,\partial_{z_1}\partial_{z_1} C_t(0,0) + \partial_{z_2}\partial_{z_1} C_t(0,0) \\ \dot{\overline{\beta}} &= S\,V_\beta + \mathrm{Cov}(\beta, \delta)\end{aligned} \tag{15}$$

$$\begin{aligned}\partial_t \partial_{z_2} C_t(0,0) &= S\,\partial_{z_1}\partial_{z_2} C_t(0,0) + \partial_{z_2}\partial_{z_2} C_t(0,0) \\ \dot{\overline{\delta}} &= S\,\mathrm{Cov}(\beta, \delta) + V_\delta\end{aligned} \tag{16}$$

Similarly we can directly get a first expression by taking the appropriate partial derivatives and setting $z_1 = z_2 = 0$ in the dynamics PDE (36) such that:

$$\begin{aligned}\partial_{z_1}^2\partial_t C_t(0,0) &= S(t)\partial_{z_1}^3 C_t(0,0) + \partial_{z_1}^2\partial_{z_2} C_t(0,0)\\ \dot{V}_\beta(t) &= S\,\mathrm{Coskew}(\beta,\beta,\beta) + \mathrm{Coskew}(\beta,\delta,\delta)\end{aligned}$$

$$\begin{aligned}\partial_{z_2}^2\partial_t C_t(0,0) &= S(t)\partial_{z_1}\partial_{z_2}^2 C_t(0,0) + \partial_{z_2}^3 C_t(0,0)\\ \dot{V}_\delta(t) &= S\,\mathrm{Coskew}(\beta,\delta,\delta) + \mathrm{Coskew}(\delta,\delta,\delta)\end{aligned} \tag{17}$$

$$\begin{aligned}\partial_{z_1}\partial_{z_2}\partial_t C_t(0,0) &= S(t)\partial_{z_1}^2\partial_{z_2} C_t(0,0) + \partial_{z_1}\partial_{z_2}^2 C_t(0,0)\\ \dot{Cov}(\beta,\delta)(t) &= S\,\mathrm{Coskew}(\beta,\beta,\delta) + \mathrm{Coskew}(\delta,\beta,\delta)\end{aligned}$$

Where $\mathrm{Coskew}(X,Y,Z) = \mathbb{E}\left[(X-\overline{X})(Y-\overline{Y})(Z-\overline{Z})\right]$ is the coskewness of the distribution X, Y and Z.

2.1 Modeling mutations

The goal in this section is to compute the effect of mutations on the dCGF of life history traits, to complete the PDE (14). In an infected cell of phenotype x, a mutation of respective effects u_β and u_δ has the following effect on the dMGF of the joint distribution of β and δ:

$$\begin{aligned}\underset{mut}{\Delta} M_t\Big((z_1,z_2)|\big((u_\beta,u_\delta),\mathbf{x}\big)\Big) &= I\,U(1-f)\,\Delta t(e^{(\beta_\mathbf{x}+u_\beta)\,z_1+(\delta_\mathbf{x}+u_\delta)\,z_2} - e^{\beta_\mathbf{x}\,z_1+\delta_\mathbf{x}\,z_2})\\ &= I\,U(1-f)\,e^{\beta_\mathbf{x}\,z_1+\delta_\mathbf{x}\,z_2}(e^{u_\beta z_1+u_\delta z_2}-1)\\ &= I\,U(1-f)\,e^{\beta_\mathbf{x}\,z_1+\delta_\mathbf{x}\,z_2}(e^{u_\beta z_1+u_\delta z_2}-1)\end{aligned} \tag{18}$$

Taking expectations over the distribution of mutational effects s in background $\beta_\mathbf{x}$:

$$\underset{mut}{\Delta} M_t\Big((z_1,z_2)|\mathbf{x}\Big) = \int\int \underset{mut}{\Delta} M_t\Big((z_1,z_2)|\big((u_\beta,u_\delta),\mathbf{x}\big)\Big)\, f\big((u_\beta,u_\delta)|\mathbf{x}\big)\, du_\beta\, du_\delta \tag{19}$$

$$= I\,U(1-f)\,\Delta t\, e^{\beta_{\mathbf{x}}\, z_1+\delta_{\mathbf{x}}\, z_2}(M^u\big((z_1,z_2)|\mathbf{x}\big)-1) \tag{20}$$

with $f\big((u_\beta,u_\delta)|\mathbf{x}\big)$ the probability density function of the joint distribution of mutational effects on transmission rate and virulence in phenotype $\mathbf{x}$, and $M^u\big((z_1,z_2)|\mathbf{x}\big)$ the MGF of the distribution of mutational effects in background $\beta_{\mathbf{x}}$. Then taking expectations over all phenotypes $\mathbf{x}$:

$$\begin{aligned}\underset{mut}{\Delta} M_t(z) &= \int p_{\mathbf{x}} \underset{mut}{\Delta} M_t\Big((z_1,z_2)|\mathbf{x}\Big)\, d^n\mathbf{x} \\ &= I\,U(1-f)\Delta t \int p_{\mathbf{x}} e^{\beta_{\mathbf{x}}\, z_1+\delta_{\mathbf{x}}\, z_2}(M^u\big((z_1,z_2)|\mathbf{x}\big)-1) d^n\mathbf{x} \\ &= I\,U(1-f)\Delta t(\overline{e^{\beta_{\mathbf{x}} z_1+\delta_{\mathbf{x}} z_2} M^u\big((z_1,z_2)|\mathbf{x}\big)} - \overline{e^{\beta_{\mathbf{x}} z_1+\delta_{\mathbf{x}} z_2}})\end{aligned} \tag{21}$$

Where the overbar refers to the average over all phenotypes $\mathbf{x}$ at time t. In continuous time, as $\Delta t \to 0$, we use the fact that $\underset{mut}{\Delta} C_t(z) = \underset{mut}{\Delta} M_t(z)/M_t(z)$ to obtain :

$$\underset{mut}{\Delta} C_t(z) = U(1-f)\Delta t(\frac{\overline{e^{\beta_{\mathbf{x}} z_1+\delta_{\mathbf{x}} z_2} M^u\big((z_1,z_2)|\mathbf{x}\big)}}{\overline{e^{\beta_{\mathbf{x}} z_1+\delta_{\mathbf{x}} z_2}}} - 1) \tag{22}$$

To go further with the expression, we need to express the MGF of the distribution of mutational effects $M^u\big((z_1,z_2)|\mathbf{x}\big)$. Note that it is not dependent on the number of infected, and so this is a classic MGF and not a density MGF. In the next section we derive an expression of this MGF in with a phenotypic dimensionality $n > 1$.

2.1.1 CGF of the distribution of mutational effects

A mutation of effect $\mathbf{u}$ on phenotype $\mathbf{x}$ has an effect u_β on transmission rate and an effect u_δ on virulence of:

$$
\begin{aligned}
(u_\beta|\mathbf{x}) = \beta(\mathbf{x}+\mathbf{u}) - \beta(\mathbf{x}) &= -\frac{2\,\mathbf{x}.\mathbf{u} + \mathbf{u}.\mathbf{u}}{2s_\beta} \\
(u_\delta|\mathbf{x}) = \delta(\mathbf{x}+\mathbf{u}) - \delta(\mathbf{x}) &= -\frac{2\,(\mathbf{x}-\mathbf{O}_\delta).\mathbf{u} + \mathbf{u}.\mathbf{u}}{2s_\delta}
\end{aligned}
\tag{23}
$$

Using polar coordinates for the mutation $\mathbf{u}$, let $r = \|\mathbf{u}\|$ and $\theta_\beta = cos(\widehat{\mathbf{x}, \mathbf{u}})$ and $\theta_\delta = cos(\widehat{(\mathbf{x} - \mathbf{O}_\delta)}, \mathbf{u})$.

$$
\begin{cases}
\mathbf{u}.\mathbf{u} = r^2 \\
\mathbf{x}.\mathbf{u} = r\ \|\mathbf{x}\|\ \theta_\beta \\
(\mathbf{x}-\mathbf{O}_\delta).\mathbf{u} = r\ \|\mathbf{x}-\mathbf{O}_\delta\|\ \theta_\delta
\end{cases}
\tag{24}
$$

We want to compute the MGF of the distribution of mutation effects u which is

$$
\begin{aligned}
M^u\big((z_1,z_2)|\mathbf{x}\big) = \mathbb{E}\left(e^{z_1(u_\beta|\mathbf{x})+z_2(u_\delta|\mathbf{x})}\right) &= \mathbb{E}_{r,\theta}\left[e^{z_1\left(-\frac{r^2}{2s_\beta} - \frac{r\,\theta_\beta\,\|\mathbf{x}\|}{s_\beta}\right) + z_2\left(-\frac{r^2}{2s_\delta} - \frac{r\,\theta_\delta\,\|\mathbf{x}-\mathbf{O}_\delta\|}{s_\delta}\right)}\right] \\
&= \mathbb{E}_r\left[e^{-\frac{ar^2}{2}}\,\mathbb{E}_{\theta_\beta,\theta_\delta}\left[e^{\theta_\beta(\frac{z_1 r\|\mathbf{x}\|}{s_\beta}) + \theta_\delta(\frac{z_2 r\|\mathbf{x}-\mathbf{O}_\delta\|}{s_\delta})}\right]\right] \\
&= \mathbb{E}_r\left[e^{-\frac{ar^2}{2}}\,M_{\theta_\beta,\theta_\delta}(\frac{z_1 r\|\mathbf{x}\|}{s_\beta}, \frac{z_2 r\|\mathbf{x}-\mathbf{O}_\delta\|}{s_\delta})\right]
\end{aligned}
\tag{25}
$$

With $a = (\frac{z_1}{s_\beta} + \frac{z_2}{s_\delta})$ and $M_{\theta_\beta,\theta_\delta}$ is the Moment Generating Function of the joint distribution of the cosine of angles from phenotype $\mathbf{x}$ between a random mutation and the optima respectively $\mathbf{O}_\beta$ and $\mathbf{O}_\delta$. Equation (A1.7) from (Martin and Lenormand, 2015) gives the following form or this MGF:

$$
M_{\theta_\beta,\theta_\delta}(t_1,t_2) = {}_0F_1(\frac{n}{2}, \frac{t_1^2 + t_2^2 + 2\rho\, t_1 t_2}{4})
\tag{26}
$$

where ${}_0F_1$ is the confluent hypergeometric function and ρ is the cosine of the angle from phenotype $\mathbf{x}$ to the two optima $(\widehat{x\mathbf{O}_\beta, x\mathbf{O}_\delta})$ and θ_δ. Using the Law of cosines and equations (1) and (3), we get :

$$\rho = \frac{s_\beta(\beta_0 - \beta_{\mathbf{x}}) + s_\delta(\delta_0 - \delta_{\mathbf{x}}) - m_{\beta\delta}}{2\sqrt{s_\beta(\beta_0 - \beta_{\mathbf{x}})}\sqrt{s_\delta(\delta_0 - \delta_{\mathbf{x}})}} \tag{27}$$

with $m_{\beta\delta} = \|\mathbf{O}_\delta\|^2/2$ Now we can rewrite the last line of equation (25) using equation (26) to get:

$$\begin{aligned}
M^u\big((z_1, z_2)|\mathbf{x}\big) &= \mathbb{E}_r\left[e^{-\frac{ar^2}{2}}\,{}_0F_1\Big(\frac{n}{2}, \frac{r^2}{4}\big(\frac{z_1^2\|\mathbf{x}\|^2}{s_\beta^2} + \frac{z_2^2\|\mathbf{x} - \mathbf{O}_\delta\|^2}{s_\delta^2} + \frac{2\rho z_1 z_2\|\mathbf{x}\|\|\mathbf{x} - \mathbf{O}_\delta\|}{s_\beta s_\delta}\big)\Big)\right] \\
&= \mathbb{E}_r\left[e^{-\frac{ar^2}{2}}\,{}_0F_1\Big(\frac{n}{2}, \frac{r^2}{4}\big(\frac{z_1^2 2s_\beta(\beta_0 - \beta_{\mathbf{x}})}{s_\beta^2} + \frac{z_2^2 2s_\delta(\delta_0 - \delta_{\mathbf{x}})}{s_\delta^2}\right. \\
&\qquad \left. + \frac{2\rho z_1 z_2\sqrt{2s_\beta(\beta_0 - \beta_{\mathbf{x}})}\sqrt{2s_\delta(\delta_0 - \delta_{\mathbf{x}})}}{s_\beta s_\delta}\big)\Big)\right] \\
&= \mathbb{E}_r\left[e^{-\frac{ar^2}{2}}\,{}_0F_1\Big(\frac{n}{2}, \frac{b\,r^2}{2}\Big)\right]
\end{aligned} \tag{28}$$

with

$$\begin{aligned}
b &= \frac{z_1^2(\beta_0 - \beta_{\mathbf{x}})}{s_\beta} + \frac{z_2^2(\delta_0 - \delta_{\mathbf{x}})}{s_\delta} + \frac{2\rho z_1 z_2\sqrt{s_\beta(\beta_0 - \beta_{\mathbf{x}})}\sqrt{s_\delta(\delta_0 - \delta_{\mathbf{x}})}}{s_\beta s_\delta} \\
&= \frac{-z_1\, z_2\, m_{\beta\delta} + (z_2 s_\beta + z_1 s_\delta)\big(z_1(\beta_0 - \beta_{\mathbf{x}}) + z_2(\delta_0 - \delta_{\mathbf{x}})\big)}{s_\beta\, s_\delta}
\end{aligned} \tag{29}$$

155 The result of (28) is given by equation (A1.8) of (Martin and Lenormand, 2015):

$$M^u\big((z_1, z_2)|\mathbf{x}\big) = e^{\frac{b\lambda}{1+a\lambda}}(1 + a\lambda)^{-\frac{n}{2}} \tag{30}$$

which finally gives for the CGF of mutational effects:

$$M^u\big((z_1,z_2)|\mathbf{x}\big) =$$
$$\Big(1+\frac{\lambda z_1}{s_\beta}+\frac{\lambda z_2}{s_\delta}\Big)^{-\frac{n}{2}} Exp\left(\frac{\lambda\,\big(-z_1\,z_2\,m_{\beta\delta}+(z_2 s_\beta+z_1 s_\delta)\big(z_1(\beta_0-\beta_{\mathbf{x}})+z_2(\delta_0-\delta_{\mathbf{x}})\big)\big)}{s_\beta\,s_\delta+\lambda(z_1 s_\delta+z_2 s_\beta)}\right)$$
$$= M^*(\mathbf{X})Exp({}^t\mathbf{X}.\omega(\mathbf{z}))$$

with

$$\mathbf{X} = \begin{pmatrix}\beta_{\mathbf{x}}\\ \delta_{\mathbf{x}}\end{pmatrix}$$
$$\mathbf{z} = \begin{pmatrix}z_1\\ z_2\end{pmatrix}$$
$$\omega(\mathbf{x}) = \begin{pmatrix}\frac{\lambda z_1(z_2 s_\beta+z_1 s_\delta)}{s_\beta s_\delta+\lambda(z_2 s_\beta+z_1 s_\delta)}\\ \frac{\lambda z_2(z_2 s_\beta+z_1 s_\delta)}{s_\beta s_\delta+\lambda(z_2 s_\beta+z_1 s_\delta)}\end{pmatrix}$$
$$M^*(\mathbf{X}) = \Big(1+\frac{\lambda z_1}{s_\beta}+\frac{\lambda z_2}{s_\delta}\Big)^{-\frac{n}{2}} Exp\left(\frac{\lambda\,(z_1\,z_2\,m_{\beta\delta}+(z_1\,s_\delta+z_2\,s_\beta)(z_1\beta_0+z_2\delta_0))}{s_\beta\,s_\delta+\lambda(z_1 s_\delta+z_2 s_\beta)}\right) \tag{31}$$

2.2 PDE with mutation

Now that we have computed the MGF of mutational effects, we can rewrite the effect of mutation on the dynamics of the dCGF of life history traits from equation (22) as:

$$\frac{\underset{mut}{\Delta}\,C_t(\mathbf{z})}{\Delta t} = U\left(M^*(\mathbf{X})\frac{\overline{e^{{}^t\mathbf{X}.(\mathbf{z}-\omega(\mathbf{z}))}}}{\overline{e^{{}^t\mathbf{X}.\mathbf{z}}}}-1\right) = U\left(M^*(\mathbf{X})e^{C_t(\mathbf{z}-\omega(\mathbf{z}))-C_t(\mathbf{z})}-1\right) \tag{32}$$

Adding this mutation term to the first PDE (14) yields a PDE describing the dynamics of the dCGF of the joint distribution of life history traits from selection and mutation:

$$\partial_t C_t(\mathbf{z}) = S\,\partial_{z_1} C_t(\mathbf{z}) + \partial_{z_2} C_t(\mathbf{z}) - d + U\left(M^*(\mathbf{X}) e^{C_t(\mathbf{z}-\omega(\mathbf{z}))-C_t(\mathbf{z})} - 1\right) \quad (33)$$

Note that we may use the notations $C_t(z_1, z_2)$ and $C_t(\mathbf{z})$ interchangeably.

2.3 Gaussian form of the dCGF and WSSM approximation

The previous PDE (36) is not directly solvable analytically. In order to do so, we must first find a general expression for the dCGF $C_t(\mathbf{z})$. To this end, we make the assumption that at all times the underlying distribution of phenotypes $\mathbf{x}$ is gaussian and with an equal variance in all directions of the phenotypic space, i.e. it follows a multivariate normal distribution $\mathcal{N}(\overline{\mathbf{x}}(t), I_n V_x(t))$ where $\overline{\mathbf{x}}(t)$ is the mean phenotype and $V_x(t)$ is the phenotypic variance at time t.

In a previous section, we have computed to MGF of the joint distribution of mutational effects on life history traits arising in a phenotype $\mathbf{x}$, in the case of Gaussian mutations with variance λ. This was thus exactly the same problem as the one described in the precedent paragraph with the only differences being: (i) λ is replaced by $V_x(t)$, (ii) we need to add the dependence on the density of infected cells $I(t)$ as we want to compute a dCGF and finally (iii)replacing u_β and u_δ by d_β and d_δ, we now want the dCGF of the joint distribution of $((\beta_{\overline{\mathbf{x}}}(t) + d_\beta, \delta_{\overline{\mathbf{x}}}(t) + d_\delta)|\overline{\mathbf{x}})$ instead of simply $((u_\beta, u_\delta)\,|\mathbf{x})$. Incorporating these differences, we thus get the following expression for the joint dMGF of the distributions of transmission rates and virulence:

$$\begin{aligned} M_t(z_1, z_2) &= \mathbb{E}\left[I(t)\, e^{\beta_{\overline{\mathbf{x}}} z_1 + \delta_{\overline{\mathbf{x}}} z_2 + d_\beta z_1 + d_\delta z_2}\right] \\ &= I(t) e^{\beta_{\overline{\mathbf{x}}}(t) z_1 + \delta_{\overline{\mathbf{x}}}(t) z_2} M^u\big((z_1, z_2)|\overline{\mathbf{x}}\big) \end{aligned} \quad (34)$$

from which follows the dCGF :

$$
\begin{aligned}
C_t(z_1, z_2) &= Log\left(M_t(z_1, z_2)\right) \\
&= Log\big(I(t)\big) - \frac{n}{2} Log\left(1 + \frac{z_1}{s_\beta} + \frac{z_2}{s_\delta}\right) + \\
&\qquad \frac{s_\beta\, s_\delta\, (z_1 \beta_{\overline{\mathbf{x}}}(t) + z_2 \delta_{\overline{\mathbf{x}}}(t)) + V_x(t)\, \big(-z_1\, z_2\, m_{\beta\delta} + (z_2 s_\beta + z_1 s_\delta)\big(z_1\, \beta_0 + z_2\, \delta_0\big)\big)}{s_\beta\, s_\delta + V_x(t)(z_1 s_\delta + z_2 s_\beta)}
\end{aligned}
\tag{35}
$$

Alongside the approximation that the underlying distribution of phenotypes is Gaussian and has the same variance in all phenotypic dimensions, we also assume that there are many mutations of small effects, meaning that we study the dynamics of our model in the WSSM regime (weak selection strong mutation). In this regime, we can linearize the mutation term in PDE (36) using a Taylor expansion in the mutational variance λ, and neglecting the terms in higher order of λ such that $\lambda^2 = \lambda^3 = ... = 0$.

$$
\begin{aligned}
\partial_t C_t(\mathbf{z}) &= S\, \partial_{z_1} C_t(\mathbf{z}) + \partial_{z_2} C_t(\mathbf{z}) - d + \\
&\qquad \frac{\mu^2}{s_\beta\, s_\delta} \left((z_1 s_\delta + z_2 s_\beta) \left(z_1(\beta_0 - \partial_{z_1} C_t(\mathbf{z})) + z_2(\delta_0 - \partial_{z_2} C_t(\mathbf{z})) - \frac{n}{2} \right) - m_{\beta\delta}\, z_1\, z_2 \right)
\end{aligned}
\tag{36}
$$

where $\mu^2 = U\lambda$.

We can also study the dynamics of the underlying phenotypes. Plugging the general expression for the dCGF $C_t(z_1, z_2)$ computed in (35) into this linearized PDE and solving for all values of $\mathbf{z}$ (or (z_1, z_2)) yields the following system of Ordinary Differential Equations describing the evolutionary dynamics:

$$
\begin{aligned}
\dot{\beta}_{\overline{\mathbf{x}}}(t) &= V_x(t)\left(\underbrace{S(t)\frac{2(\beta_0-\beta_{\overline{\mathbf{x}}})}{s_\beta}}_{\text{selection}} + \underbrace{T(t)}_{\text{trade-off}}\right)\\
\dot{\delta}_{\overline{\mathbf{x}}}(t) &= V_x(t)\left(\underbrace{\frac{2(\delta_0-\delta_{\overline{\mathbf{x}}})}{s_\delta}}_{\text{selection}} + \underbrace{S(t)T(t)}_{\text{trade-off}}\right)\\
\dot{V}_x(t) &= \underbrace{-\left(\frac{S(t)}{s_\beta}+\frac{1}{s_\delta}\right)V_x(t)}_{\text{selection}} + \underbrace{\mu^2}_{\text{mutation}}
\end{aligned}
\tag{37}
$$

with

$$
\begin{aligned}
T(t) &= \frac{(\beta_0-\beta_{\overline{\mathbf{x}}})}{s_\delta} + \frac{(\delta_0-\delta_{\overline{\mathbf{x}}})}{s_\beta} - \frac{m_{\beta\delta}}{s_\beta s_\delta}\\
&= \frac{1}{s_\beta\, s_\delta}(\rho\|\mathbf{x}\|\|\mathbf{x}-\mathbf{O}_\delta\|)
\end{aligned}
\tag{38}
$$

Which is minimized when $\|\mathbf{x}\| = \|\mathbf{x}-\mathbf{O}_\delta\| = \frac{\|\mathbf{x}-\mathbf{O}_\delta\|}{2}$ and $\rho = -1$ which is the case when $\mathbf{x}$ is exactly in the middle of the segment between the two optima, that is $\mathbf{x} = \frac{1}{2}\mathbf{O}_\delta$.

With the dynamical system (37), we now have access to the dynamics of the three variables upon which depends the dCGF of the joint distributions of transmission rate and virulence. With these dynamics and recalling the dependence of the mean life history traits on these parameters of equation (6), we can now fully write the dynamics of the mean transmission rate and virulence with mutation:

$$
\begin{pmatrix}\dot{\overline{\beta}}(t)\\ \dot{\overline{\delta}}(t)\end{pmatrix} = \underbrace{\begin{pmatrix}V_\beta(t) & \mathrm{Cov}(\beta,\delta)(t)\\ \mathrm{Cov}(\beta,\delta)(t) & V_\delta(t)\end{pmatrix}}_{\mathbf{G}} \cdot \begin{pmatrix}S(t)\\ 1\end{pmatrix} - \frac{nU\lambda}{2}\begin{pmatrix}1/s_\beta\\ 1/s_\delta\end{pmatrix} \tag{39}
$$

We specifically write the above equation to resemble a classic Price equation (Day and Gandon, 2007; Price, 1970) with a variance covariance matrix G, a selection vector and a mutation vector. Using the properties (11)-(12) of the dCGF $C_t(z_1, z_2)$ we can express the variances and covariance as a function of the phenotypic parameters $\beta_{\overline{\mathbf{x}}}$, $\delta_{\overline{\mathbf{x}}}$ and $V_x(t)$:

$$\begin{aligned} V_\beta &= \frac{2V_x(t)}{s_\beta}(\beta_0 - \beta_{\overline{\mathbf{x}}}(t)) + \frac{nV_x^2(t)}{2s_\beta^2} \\ V_\delta &= \frac{2V_x(t)}{s_\delta}(\delta_0 - \delta_{\overline{\mathbf{x}}}(t)) + \frac{nV_x^2(t)}{2s_\delta^2} \\ \mathrm{Cov}(\beta, \delta) &= V_x(t)T(t) + \frac{nV_x^2(t)}{2s_\beta s_\delta} \end{aligned} \tag{40}$$

With the dynamical expression from system (37), we can compute the dynamics of these variance and covariance and express them as:

$$\begin{aligned} \dot{\mathrm{Cov}}(\beta, \delta)(t) = -\frac{2V_x(t)}{s_\beta\, s\delta}\Big(&\mathrm{Cov}(\beta, \delta)(t)\big(s_\beta + s_\delta S(t)\big) \\ &+ V_x(t)\big((\beta_0 - \beta_{\overline{\mathbf{x}}}) + S(t)\big((\delta_0 - \delta_{\overline{\mathbf{x}}})\big)\big)\Big) \\ &+ U\lambda\left(T(t) + \frac{nV_x(t)}{s_\beta\, s_\delta}\right) \end{aligned} \tag{41}$$

$$\begin{aligned} \dot{V}_\beta(t) = -2V_x^2(t)\Bigg(&S(t)\frac{\big((\beta_0 - \overline{\beta}(t)) + 2(\beta_0 - \beta_{\overline{\mathbf{x}}}(t))\big)}{s_\beta^2} + \underbrace{\frac{T(t)}{s_\beta} + \frac{(\beta_0 - \overline{\beta}(t))}{s_\beta s_\delta}}_{\text{Trade-off term}}\Bigg) \\ &+ 2U\lambda\big(\beta_0 - \overline{\beta}(t)\big) \end{aligned} \tag{42}$$

$$\dot{V}_\delta(t) = -2V_x^2(t)\left(\frac{\big((\delta_0 - \overline{\delta}(t)) + 2(\delta_0 - \delta_{\overline{\mathbf{x}}}(t))\big)}{s_\delta^2} + \underbrace{S(t)\left(\frac{T(t)}{s_\delta} + \frac{(\delta_0 - \overline{\delta}(t))}{s_\beta s_\delta}\right)}_{\text{Trade-off term}}\right)$$
$$2 + U\lambda\big(\delta_0 - \overline{\delta}(t)\big) \tag{43}$$

References

Anderson, R. M. and R. M. May (1982). "Coevolution of hosts and parasites". In: *Parasitology* 85.2, pp. 411–426 (cit. on p. 2).

— (1991). *Infectious diseases of humans: dynamics and control*. Oxford university press (cit. on p. 2).

Day, T. and S. Gandon (2007). "Applying population-genetic models in theoretical evolutionary epidemiology". In: *Ecology Letters* 10.10, pp. 876–888 (cit. on p. 16).

Fenner, F. and I. Marshall (1957). "A comparison of the virulence for European rabbits (Oryctolagus cuniculus) of strains of myxoma virus recovered in the field in Australia, Europe and America". In: *Epidemiology & Infection* 55.2, pp. 149–191 (cit. on p. 2).

Lande, R. (1982). "A quantitative genetic theory of life history evolution". In: *Ecology* 63.3, pp. 607–615 (cit. on p. 2).

Martin, G. and T. Lenormand (2015). "The fitness effect of mutations across environments: Fisher's geometrical model with multiple optima". In: *Evolution* 69.6, pp. 1433–1447 (cit. on pp. 10, 11).

Price, G. R. (1970). "Selection and covariance". In: *Nature* 227, pp. 520–521 (cit. on p. 16).

Roff, D. (1993). *Evolution of life histories: theory and analysis*. Springer Science & Business Media (cit. on p. 2).

Appendix B: The viral escape of CRISPR immunity: impact of mutation rate and host frequency

Martin Guillemet, François Gatchitch, Sylvain Moineau, Sylvain Gandon

This appendix is a first draft for a manuscript.

Introduction

Biological adaptation is led by the emergence of mutations and their subsequent rise in frequency. With a large enough effective population, the speed at which a mutation can emerge and grow in frequency is mainly determined by two parameters. First, the mutation rate determines the probability of a mutation occurring and thus the influx of mutation. Second, the selection coefficient (that is the additional fitness associated with the mutation) influences the probability that a mutation will be initially lost through drift. If this mutation is not lost initially, then the selection coefficient also dictates the speed of invasion of this mutation. However the relative contributions of mutation rate and selection coefficient to the emergence of new dominating mutations can be hard to determine. Particularly in epidemiology, it has become clear that the ability to predict the evolution of pathogens and the successive rise of new variants is crucial.

To address the question of the relative contribution of mutation rate and selection, we used an experimental host pathogen system of CRISPR-resistant bacteria *Streptococcus thermophilus* and its phage 2972. With the CRISPR adaptive immunity system, *S. thermophilus* can acquire resistance to this phage by incorporating fragments of the phage genome in its CRISPR locus, which is used as template to cut phage genetic material upon entry in the cell. These fragments are then called *spacers*. As CRISPR resistance is based on identity between the spacer and the phage genome, this resistance can be escaped by the phage with mutations in the targeted regions: the *protospacers*.

In a previous experiment, Chabas et al. measured the mutation rate in different protospacers of phage 2972 and found significant differences (Chabas, Nicot, et al., 2019). However, in the third chapter of this thesis we found in a coevolution experiment that phage adaptation to CRISPR immunity was driven by the frequency of hosts and found no effect of mutation rate.

Disentangling the contribution of mutation rate and selection coefficient is especially difficult with mutations acting on the ability to escape host immunity. A common issue is that the selection coefficient associated with a given mutation is dependent on time. Indeed for a pathogen, a mutation granting escape to the immunity of a certain host will be more beneficial when this host is frequent in the environment. The change in composition of the host population leads to a change in the selective pressures on the pathogen population. An additional layer of complexity is from the feedback between

the composition of the pathogen population the evolution of the host population: if the escape mutation towards a given host is frequent in the population, then this host will tend to decrease in frequency, thus reducing the selection coefficient associated with this escape mutation. Indeed, there is coevolution and frequency-dependent selection at play, and so the pathogen dynamics cannot be decoupled from the epidemiological dynamics.

To try to circumvent these issues, we designed an experiment where we limit the evolution of the host population to focus on the evolutionary dynamics of the phage population. The experiment is described in Figure 1. We used six strains of *S. thermophilus* divided in two groups which differ in their durability, ie. the escape mutation rate in the corresponding protospacer of phage 2972 according to (Chabas, Nicot, et al., 2019), to test for the effect of protospacer mutation rate. We call the two groups of hosts LM (Low Mutation) and HM (High Mutation), referring to the mutation rate of the corresponding protospacers in the phage: mutation rate in protospacers targeted by HM hosts is higher than mutation rate in protospacers targeted by LM hosts (see Figure S1).

Our evolution experiment took place over five days 1. In a first treatment (A), we only used wild-type bacteria, to observe a baseline level of phage adaptation without host resistance. In treatment (B), we used a 1:1 ratio of LM and HM strains, in which we could expect that the escape mutations associated with the HM strains would invade the phage population the fastest. In another treatment (C), we manipulated the hosts frequencies to go against the expected effect of the mutation rates: the LM strains were more frequent than the HM strains, in a 80:15 ratio. To guarantee that the wild type phage could grow, wild-type susceptible bacteria always made up 5% of the population. Limiting the evolution of the hosts as discussed earlier was done by filtering phages each day, and transferring them to a fresh-made mix of bacteria of constant frequencies. Thus the phages evolved for 5 days in an environment which was reset to the same conditions everyday. We used whole genome sequencing to monitor the frequencies of bacteriophage escape of host immunity.

Results

Escape frequency

We first study the dynamics of the frequency of escape averaged over the six different host strains without distinction between the two groups LM or HM. We observed in Figure 2 that the mean frequency of escape mutations against all resistant hosts was significantly higher in the treatment where LM and HM hosts were initially in the same frequency, than in the treatment where host frequency is initially heterogeneous (Day 1, $T = 3.90$, $df = 12$, $P = 2.11 \times 10^{-3}$; Day 2, $T = 3.86$, $df = 10$ $P = 3.15 \times 10^{-3}$). From day 3, we detected no significant differences between the two treatments. We found that phages emerge quickly, as the mean frequency of escape against the six types of hosts becomes higher than $\frac{1}{6}$. Above this threshold, we know that there is a portion of the phage population that harbors several escape haplotypes and is thus capable of infecting different resistant hosts. There were multi-escaping phages in all replicates in the homogeneous treatment after two days, and in the heterogeneous treatment after 3 days. At the end of our experiment, we observed that no plateau had been reached as the escape haplotype frequencies kept increasing. We do not even notice a clear saturation, as the increase in mean escape mutation frequency seemed close to linear until the end of the experiment. This supports the hypothesis that, with enough time, the phage population will tend towards a population of generalist phages able to infect all six resistant host strains.

We wanted to test the contribution of two parameters on the evolution of the phage populations: escape mutation rate and frequency of the corresponding host (which relates to the selection coefficient). In the homogeneous treatment (B), the six host strains were mixed in equal frequencies to test the effect of mutation rate independently. We show the mean escape mutation frequency per type of hosts in Figure 3.a. We do not detect an effect of mutation rate on escape frequency after the first day as the difference in mean frequency (HM: $f_{\text{escape}} = 0.136$, LM: $f_{\text{escape}} = 0.082$) is not significant ($T = -1.51$,$df = 12$, $P = 0.16$). However we observe that the mean frequency of escape mutation is higher against HM hosts later in the experiment at day 3 ($T = -2.49, df = 12, P = 0.028$) and 4 ($T = -4.56, df = 12, P = 6.5 \times 10^{-1}$). This effect then disappears on the final day. This delayed effect was surprising, but can be linked to the observation from Figure 2 that generalist phages which can infect multiple resistant hosts appear in the first days and keep increasing in frequency after that. Thus mutation is also limiting in the later days, as phages acquire more escape

mutations to escape all different types of hosts. Indeed mutations should be limiting until the appearance of a phage harboring escape mutations against all 6 types of host.

In Figure 3.b we show the results from the heterogeneous treatment, in which the resistant bacterial hosts are initially present in contrasting frequencies. LM hosts initially make up 80% of the host population compared to 15% for the LM hosts. There is then a higher selective pressure selecting for escape mutant to frequent LM hosts rather than rarer HM hosts, which translates to a higher selection coefficient for the former. Therefore we expected that the mean escape mutation frequency would be higher against LM than HM hosts throughout the experiment. We do observe this effect at the end of the first day ($T = 4.75, df = 12, P = 4.7 \times 10^{-4}$) but not later in the experiment. Hence the frequency of hosts only drove phage adaptation early in the experiment, before the frequency of escape mutation against both types of host equalize. However, we found in the homogeneous treatment that mutation rate favored the increase on HM escape mutation frequency in days 3 and 4, so the equal frequencies that we find on days 3 and 4 in the heterogeneous treatment could be due to the effects of mutation rate and host frequency cancelling each other.

Host frequencies

To better understand the dynamics of the frequency of escape mutations, we sequenced the CRISPR locus of the bacterial host population at the end of each day. Indeed to try to limit host evolution, we add a fresh identical mix everyday, yet during the day we expect the frequencies of the hosts to vary. We present the results in Figure 4. Note that in this figure the frequency of wildtype bacteria, or bacteria which acquired news spacers, is not represented. In the homogeneous treatment, we find a significant difference in the frequencies of HM and LM hosts at the end of the first day ($T = 19.5, df = 12, P = 1.87 \times 10^{-10}$). The tendency of LM hosts to have a higher frequency after the first day could be due to a faster increase in the frequency of HM escaping phages, although the difference we observed in Figure 3.a is not significant. After the second day, we detect no difference between the frequencies of HM and LM hosts at the end of each day.

In the second treatment, we find that at the end of the first day the composition of the bacterial population is very similar to the initial mix (Figure 4.a). Strikingly, although the bacterial mix is fresh at the start of each day and with very heterogeneous

frequencies of hosts, we detect no significant differences between the frequencies of LM and HM hosts at the end of the growth phase from day 2, 4 and 5. Even more striking, at the end of day 3 and opposite to the initial mix, there are more HM that LM hosts ($T = -2.75, df = 12, P = 0.017$). Thus the pressure of the adapting phage population seems to homogenize the bacterial population. Towards the end of the experiment, there is balance between LM and HM host strains.

Diversity of escape mutants

With our sequencing data of the phage protospacers, and particularly with the haplotype frequencies which keeps the linkage information between escape mutations on the same protospacer, we can study the diversity of escape haplotypes through time. We define an escape haplotype as a set of mutations which are found in the same protospacer, in the same sequencing reads, and thus initially in the same phage. We use the frequency of escape haplotypes instead of individuals mutations because we find occurrences of several escape mutations in the same read which inflate the escape mutation frequencies, but not escape haplotype frequencies. Note that these haplotypes regroup mutations in the same protospacer, as we do not have linkage information for mutations in different protospacers. We show the dynamics of the effective number of escape haplotypes (Nei, 1973) in Figure 5. This measure is computed with the frequencies of the different haplotypes. With n different haplotypes, this effective number of haplotypes is maximized and equal to n when the frequencies of all haplotypes are equal. We find that this diversity stays relatively constant through time even though we found that the overall frequency of escape haplotypes was increasing quasi linearly (Figure 3 and 2). In the homogeneous treatment (Figure 5.a), we find on day 4 a significantly higher diversity of escape haplotypes in HM rather than LM protospacers ($T = -4.56, df = 12, P = 6.5 \times 10^{-4}$). In the heterogeneous treatment (Figure 5.b), at the end of the first day, the escape haplotype diversity is higher against the initially more frequent LM than the HM hosts ($T = 4.75, df = 12, P = 4.7 \times 10^{-4}$). These observations coincide with the observations from Figure 3 where we find significant differences in frequency between LM and HM hosts, meaning that the difference in haplotype diversity between LM- and HM- escaping phages correlates with the difference in frequency.

This diversity of haplotypes can be observed in more details with the use of Muller plots. We show examples of these plots for one replicate of both the homogeneous

(Figure S2) and the heterogeneous (Figure S3) treatments. With these plots we can precisely see the dynamics of the frequencies of each escape haplotypes against each of the resistant hosts. We observe that the relatively low diversity shown in Figure 5 is also quite static. Indeed against each resistant host, the escape haplotype composition is relatively stable across time.

In Figure S4 we show the presence of each escape haplotypes over the different replicates. With this representation, we can observe whether there are ubiquitous haplotypes that appear in many replicates, or if there are enough mutational targets in a protospacer so that escape haplotypes are different between replicates. This figure shows that many escape haplotypes are found in different replicates. This supports the hypothesis that the number of escape mutation against every host spacer is limited and the same escape mutations are found to drive the viral adaptation to CRISPR immunity.

Discussion

The impact of host CRISPR diversity driving bacteriophages to extinction(Common et al., 2020; Houte et al., 2016; Morley et al., 2017) or limiting their emergence (Chabas, Lion, et al., 2018) has been demonstrated and discussed before. However these earlier studies manipulate diversity through the number of host present, and not their relative frequencies. Manipulating the frequencies of hosts affects the strength of selection. Here we find that in a CRISPR resistant bacteria and phage system, heterogeneity in host frequencies initially reduces the global speed of acquisition of escape mutations in the pathogen. We observe, as expected, that phages acquire escape mutations faster against more frequent hosts, but this effect seems to be transient. Indeed we observe that the initial heterogeneity in the host population vanishes as phages adapt, even with fresh bacteria being added daily. Contrary to expectations, we do not find an effect of escape mutation rate on phage adaptation early in the experiment, when mutations could be thought to be limiting. However we find a delayed effect of mutation rate as later in the experiment we do find a higher frequency of escape haplotypes for which mutation rate is higher. This effect could be explained by multi locus adaptation. We find that phages escaping the resistance from several hosts appear early in the experiment but we do not a reach a point where the whole phage population can infect all hosts. This explains why mutations could still be limiting later in the experiment, and thus why mutation rate can still impact phage adaptation.

Another consequence of multi-locus adaptation is found on the diversity of escape

haplotypes which we observe against a given spacer, within a replicate (Figure S2 and S3). We find that several escape haplotypes against the same spacer can co-exist across time, which we attribute to the possibility that these escape haplotypes are found in phages which infect different sets of other hosts. Hence it is linkage with other escape mutations against other hosts which would help conserve a diversity of escape haplotypes against a given host. This finding highlights the need for long-read sequencing methods to track multi-locus adaptation in this experimental system.

The goal of our experiment was to test the relative importance of mutation rate and host frequency on phage adaptation. To this end we tried to limit host adaptation by resetting the bacterial composition each day. Yet this approach was not perfect as we find that the host frequencies change drastically during each day of the experiment. Besides, we find the emergence of bacteria with additional spacers of resistance (which explains why the sum of the frequencies of HM and LM hosts types in Figure 4 is not equal to 1 at the end of the experiment). Yet this acquisition of new spacers is expected to have a negligible effect on our conclusions. An emergent resistant bacteria would need to grow to a high enough frequency for coevolution to be significant. Moreover, it would be unlikely that the same spacer was acquired several times over in several days in the same replicate (there are more than 600 possible spacers against phage 2972) so the increase in selection coefficient for the corresponding phage mutation would be a one-time event.

Materials and Methods

Bacteria and phages strains

S. thermophilus DGCC 7710 and phage 2972 (Lévesque et al., 2005) were obtained from the Félix d'Hérelle Reference Center for Bacterial Viruses (www.phage.ulaval.ca). Several derivative phage-resistant strains, each with an unique CRISPR spacer were generated previously (Chabas, Nicot, et al., 2019). The ability of phages to mutate and escape the resistance from these strains was assessed. Based on these measures, six strains were chosen, divided in two groups for high and low resistance durability values ie. high and low protospacer escape mutation rate in the phage (see Figure S1).

Experimental procedure

Prior to the experiment, the 7 bacterial strains (including the susceptible wild-type) were grown separately during 6 hours in LM17+$CaCl_2$ (37 g/l of M17 (Oxoid) supplemented with 5 g/l of lactose and 10 mM of sterile $CaCl_2$). The bacterial mixes were then made according to the three treatments described in Figure S1. Everyday, new mixes were made to avoid growth and competition between the bacteria which could lead to dramatic changes in relative frequencies. On the first day, the three bacterial mixes were transferred 1:100 into 10 ml of fresh LM17+$CaCl_2$ and infected with 10^5 wild-type 2972 phages then incubated at 42°C. There were 4 replicates for treatment A, and 7 for both treatment B and C. Every day (after 18 hours of incubation), the cultures were filtered (0.2 µm) to extract phages from the cultures. 100µl of these filtered phages were used to infect the newly made mix each day. Following each transfer, the bacteria and phages from each replicate, as well as the initial bacteria mixes, were kept for sequencing.

Bacteria sequencing

The CRISPR-Cas CR1 locus was amplified through PCR (primers 5'-3': AGTAAGGATTGACAAGGACAGT; CCAATAGCTCCTCGTCATT) and sequenced with Illumina MiSeq. The spacers were extracted from the sequences by searching for the flanking repeats allowing for a maximum of one mismatch. The spacers were then matched with their protospacers on the phage genome using Blast version 2.8.1 (Camacho et al., 2009) and the protospacer database presented in the next section. After these steps, an average sequencing depth of around 126000 was obtained. A minimum identical word size of 10, and a 70% identity threshold was used. The top result of the search, if any, was used to replace the name of the spacer by the middle position of the protospacer in the phage genome. A frequency cutoff of 1% was used to optimize the quality of our dataset.

Phage sequencing

The phage DNA samples were sequenced (Illumina MiSeq) with 150-bp paired-end reads. Trimmomatic (Bolger, Lohse, and Usadel, 2014) was used to clean and trim the sequencing reads, before mapping them on the reference genome (Lévesque et al., 2005) using Bowtie2 (Langmead and Salzberg, 2012). The software FreeBayes (Garrison and Marth, 2012) was then used to detect mutations and their frequencies, filtering with a 0.01 frequency threshold. In order to keep the linkage information between the

escape mutations, we developed another approach using directly the aligned reads. We selected the reads which contained entire protospacers to assess the escape haplotypes against each of the bacterial hosts. We then only kept the reads with escape haplotypes which were seen at least three times. We recovered on average 690 reads spanning each complete protospacer. With this read approach, we keep linkage information between escape mutations in the same protospacer, which allows us to estimate escape haplotype frequencies instead of independent mutation frequencies. For instance, this alleviates the problem of escape mutation frequency becoming higher than one in a population, when some viruses carry multiple escape mutations in the same protospacers.

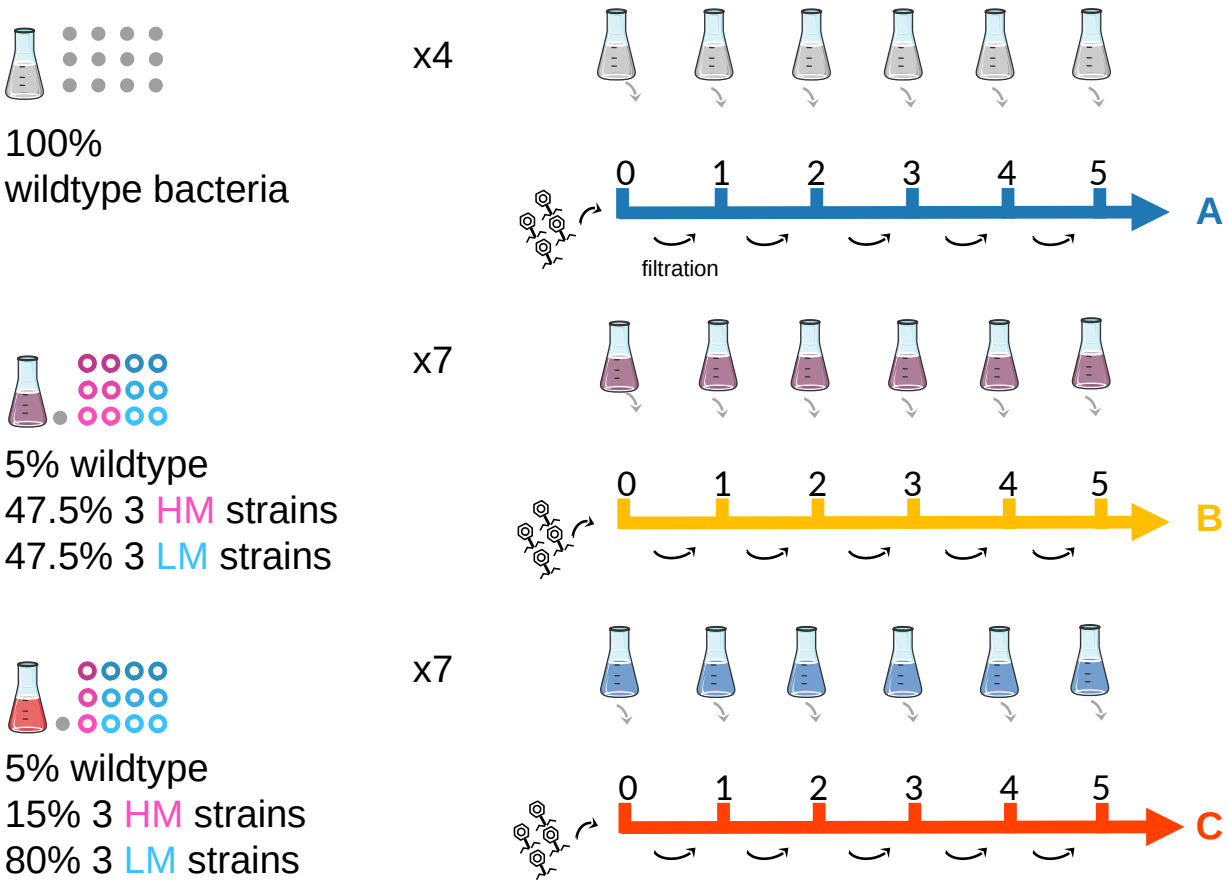


Figure 1: Bacteriophage evolution experiment. On the first day, wild type phages are added to a certain population of bacteria. At the end of each day, the lysate is filtered to obtain only the evolved phages. Everyday, the recovered phages are used to infect a new fresh population of bacteria. In the first treatment A, the bacterial mix is made up of only susceptible cells. In the second treatment B, we add a mix of 5% susceptible bacteria, and in equal frequencies the 6 resistant host strains. Finally in the third treatment C, the bacterial mix is composed of 5% susceptible cells, 15% of the three HM (high mutation) strains and 80% of the three LM (low mutation) strains.

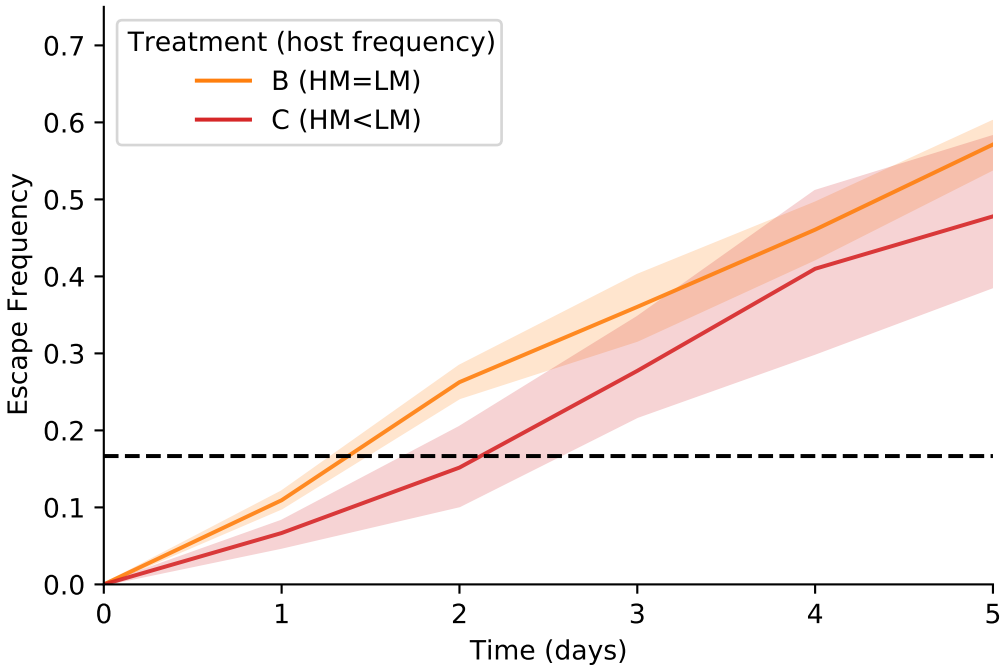


Figure 2: Escape frequency across time averaged over the six protospacers. This value is averaged over the seven replicates of each treatment. The orange line corresponds to the B treatment, where the frequency of HM and LM hosts in the mix is equal, and the red line corresponds to the C treatment where the mix is made up of more LM than HM strains. The dashed horizontal line shows $y = 1/6$ which represents the maximum mean escape mutation frequency if all phages are specialists, that is only harbor escape mutations against one of the six hosts. The shadowed areas denote the bootstrap 95% confidence interval computed over the seven replicates.

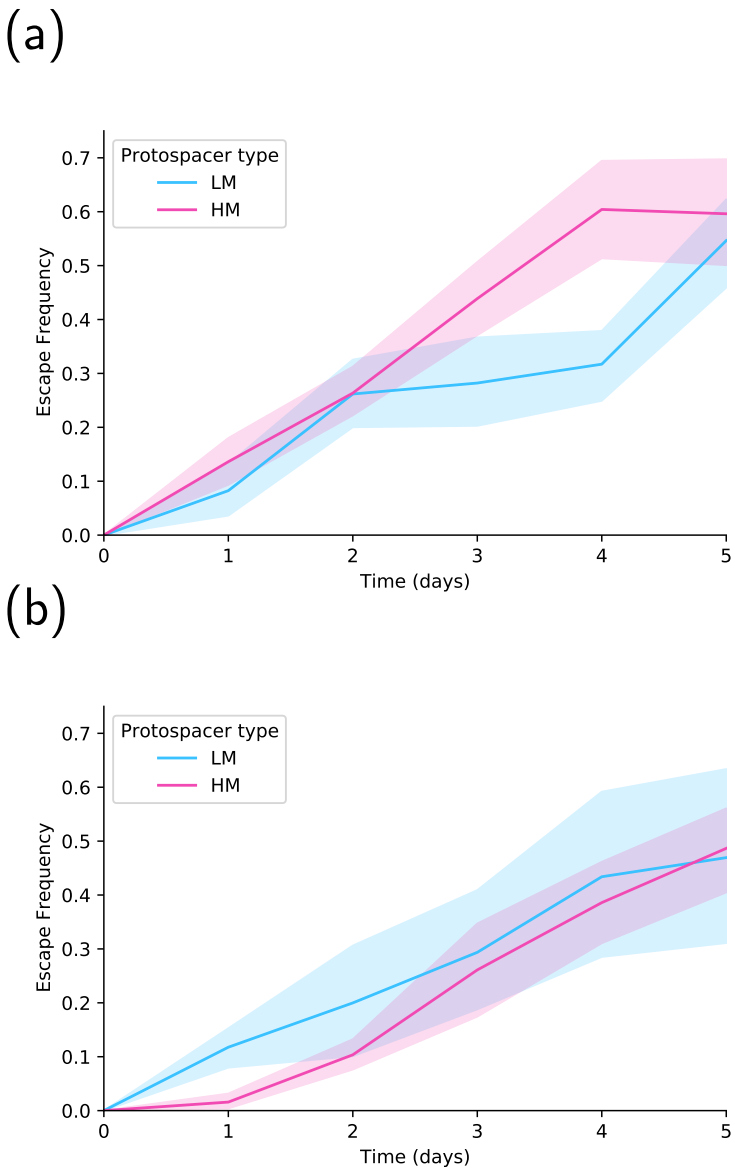


Figure 3: Escape frequencies across time averaged over the three protospacers of each type. Frequency is shown in pink for HM protospacers and in blue for LM protospacers. This value is then averaged over the seven replicates of each treatment. The (a) panel corresponds to the B treatment, where the initial frequency of HM and LM hosts in the mix is equal, the (b) panel corresponds to the C treatment where the mix is made up of more LM than HM strains. The shadowed areas denote the bootstrap 95% confidence interval computed over the seven replicates.

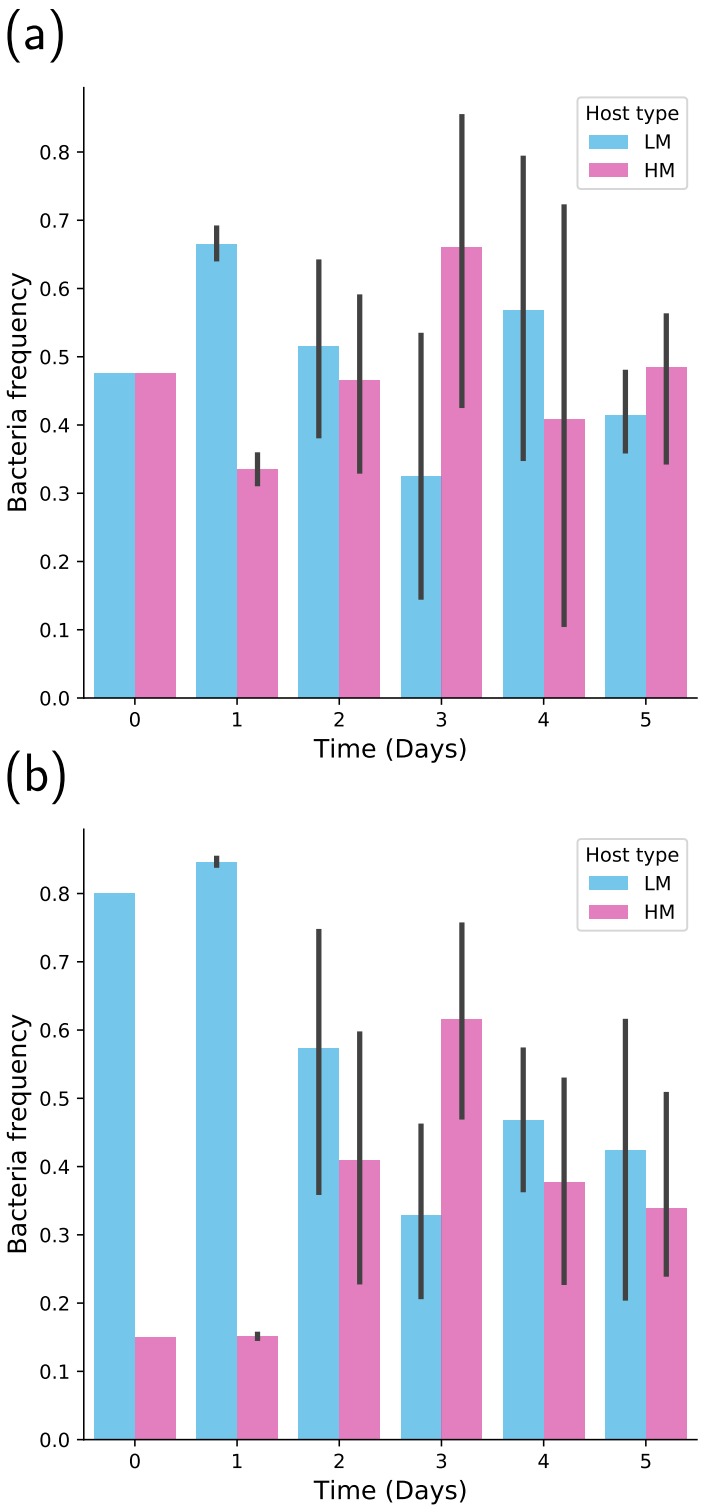


Figure 4: Frequency of the bacterial hosts at the end of each day. The frequency of HM hosts is shown in pink, the frequency of LM hosts is shown in blue. The error bars denote the bootstrap 95% confidence interval computed over the seven replicates. The (a) panel corresponds to the B treatment, where the initial frequency of HM and LM hosts in the mix is equal, the (b) panel corresponds to the C treatment where the mix is made up of more LM than HM strains.

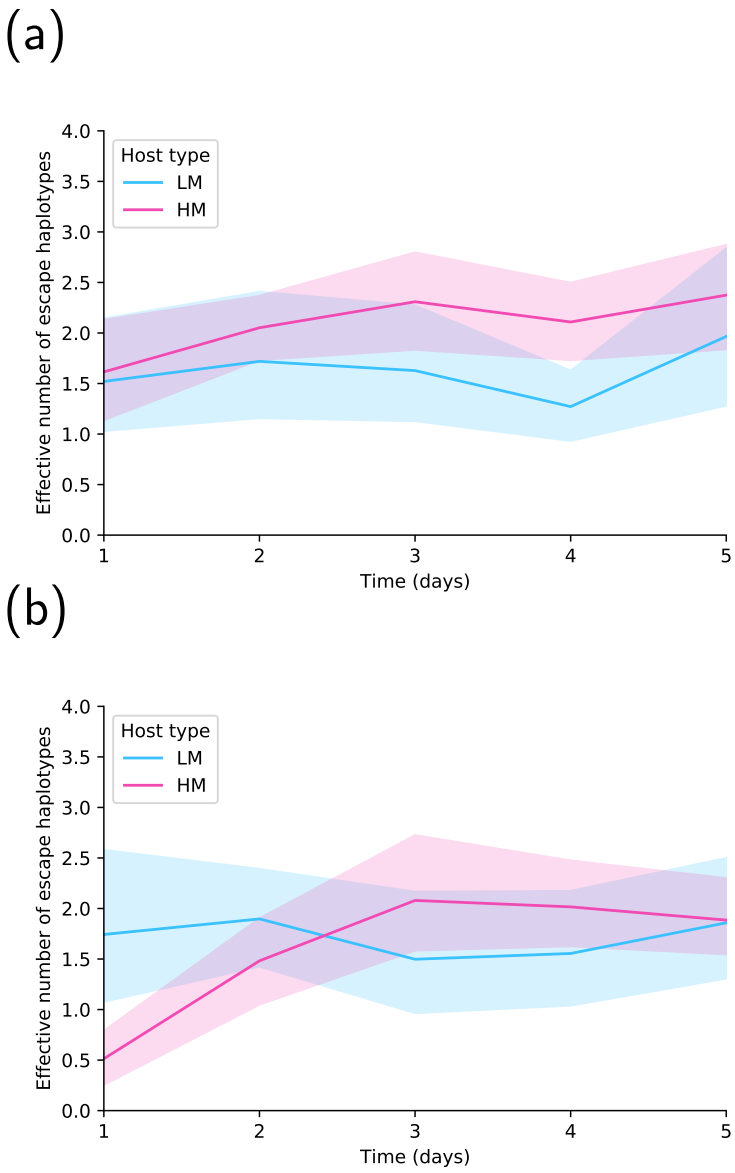


Figure 5: Escape haplotypes diversity. The value shown is the mean effective number of haplotypes against the three different HM (in pink) or LM (in blue) host strains respectively. The effective number of host genotypes is computed using : $1/(\sum_{i=1}^{n} p_i^2)$, where n is the number of escape haplotypes, and p_i is the frequency of escape haplotype i (Nei, 1973). This value is then averaged over the seven replicates of each treatment. The shadowed areas denote the bootstrap 95% confidence interval computed over the seven replicates. The (a) panel corresponds to the B treatment, where the initial frequency of HM and LM hosts in the mix is equal, the (b) panel corresponds to the C treatment where the mix is made up of more LM than HM strains.

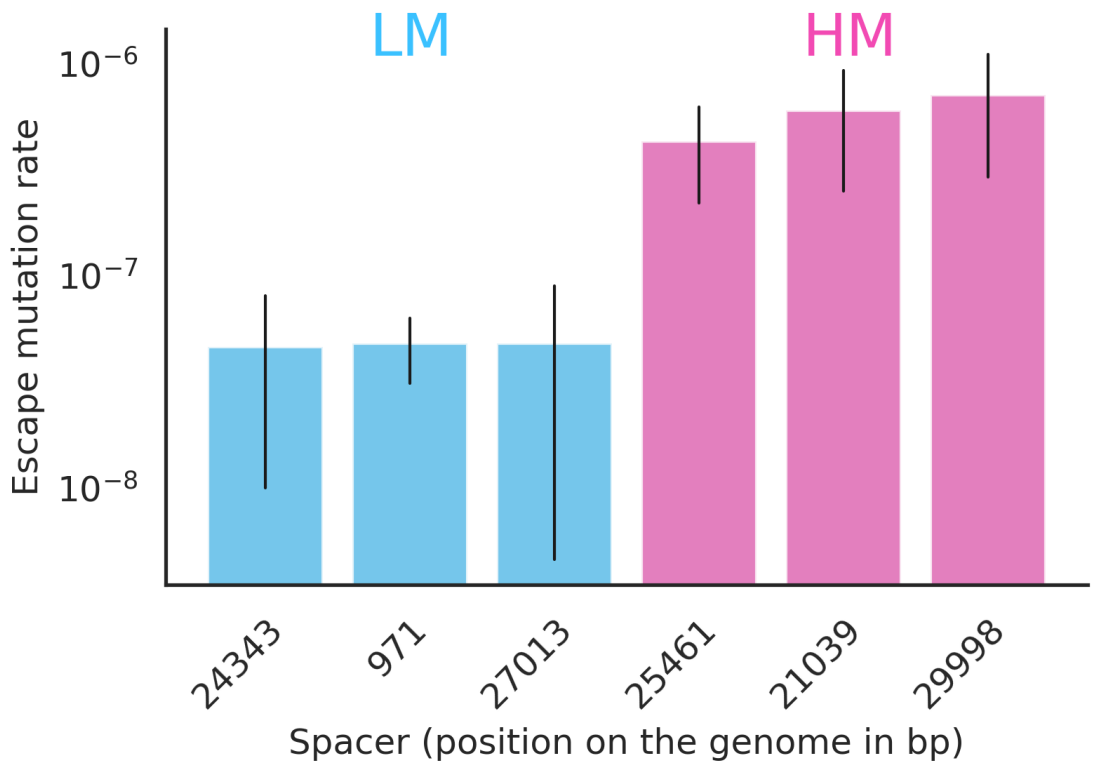


Figure S1: Escape mutation rate in the protospacers targeted by the six different bacterial hosts. Spacers are named according to the position of the corresponding protospacer in the phage genome. The color represent the two groups LM (low mutation) and HM (high mutation). Mutation rate was computed in ref (Chabas, Nicot, et al., 2019).

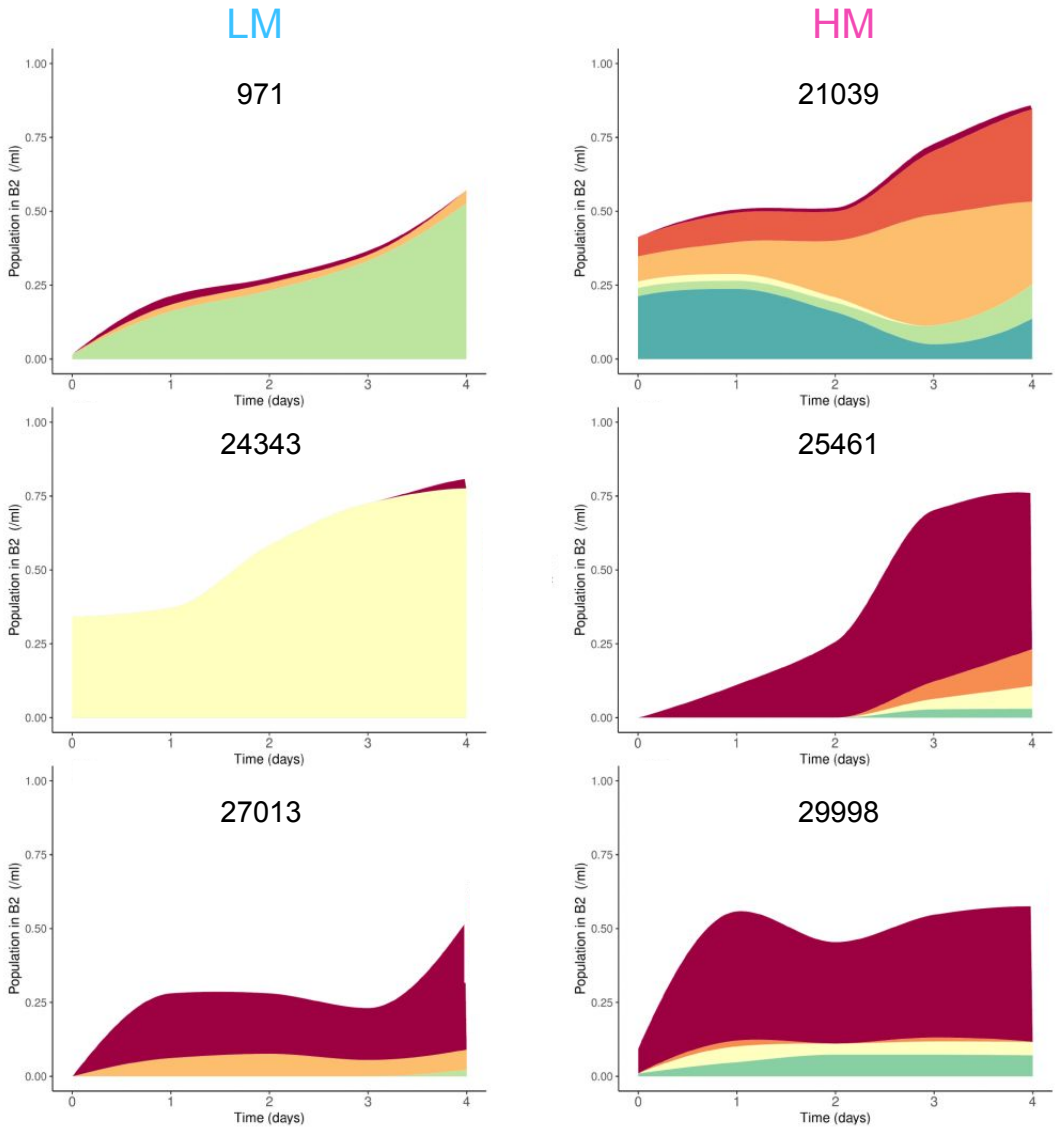


Figure S2: Muller plots of the phage escape haplotypes through time in one replicate of treatment B (where HM and LM hosts are initially in equal frequencies). The frequency of phage escape haplotypes is shown against every resistant host. Each colour corresponds to one distinct escape haplotype. Note that two same colour in different plots do not denote any relationship. The name of the host strain (corresponding to the position of the corresponding protospacer on the phage genome) is written on each plot, with the LM and HM hosts strains being respectively shown in the left and right columns. The lines are smoothed between each day.

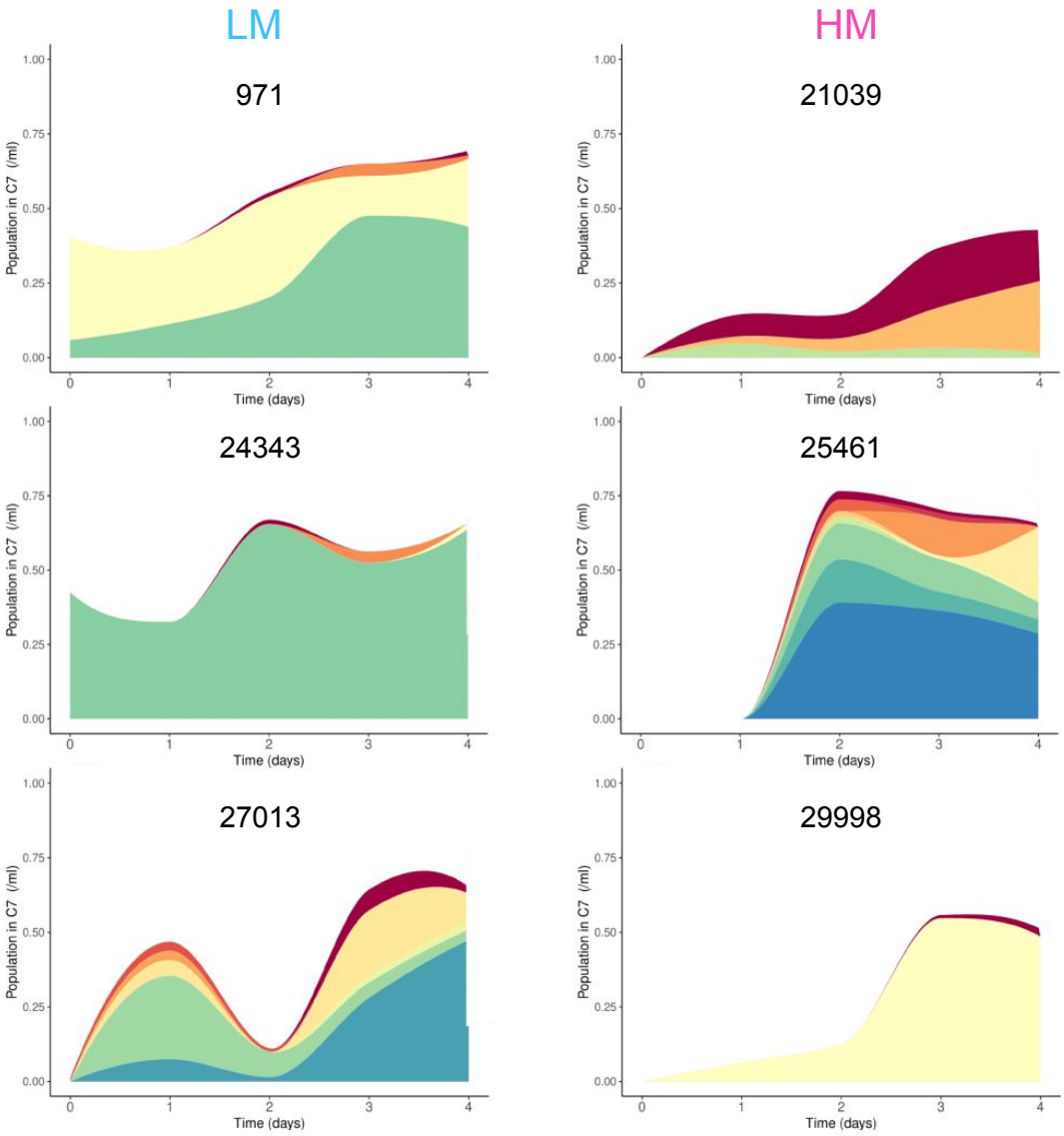


Figure S3: Muller plots of the phage escape haplotypes through time in one replicate of treatment C (where HM hosts are initially less frequent than LM hosts). The frequency of phage escape haplotypes is shown against every resistant host. Each colour corresponds to one distinct escape haplotype. Note that two same colour in different plots do not denote any relationship. The name of the host strain (corresponding to the position of the corresponding protospacer on the phage genome) is written on each plot, with the LM and HM hosts strains being respectively shown in the left and right columns. The lines are smoothed between each day.

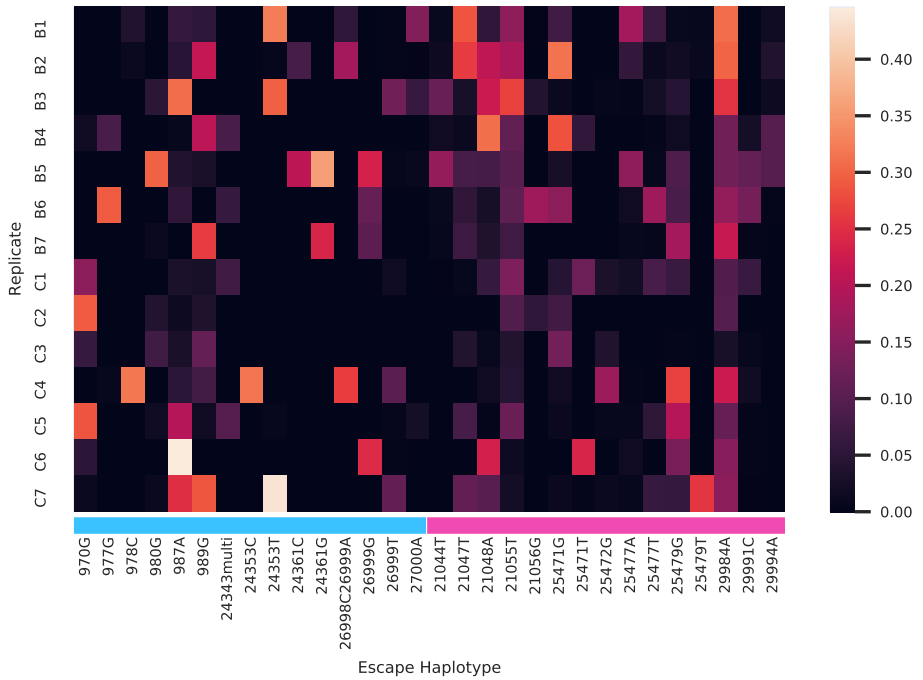


Figure S4: Heatmap of the different escape haplotypes in different replicates. We show for each replicate of the B and C treatments the frequencies averaged over time of every escape haplotype. The haplotypes are filtered to keep only those for which the sum over all replicates of these averages frequencies are over 0.25. The haplotype names denote the position of the mutation followed by the changed nucleotide. This is repeated in case of a haplotype with two mutations. Haplotype '24343multi' is a haplotype with more than 5 mutations. The blue line marks the haplotypes in LM protospacers, and the pink line marks haplotypes in HM protospacers.

References

Bolger, A. M., M. Lohse, and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15, pp. 2114–2120 (cit. on p. 9).

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10.1, pp. 1–9 (cit. on p. 9).

Chabas, H., S. Lion, A. Nicot, S. Meaden, S. van Houte, S. Moineau, L. M. Wahl, E. R. Westra, and S. Gandon (2018). "Evolutionary emergence of infectious diseases in heterogeneous host populations". In: *PLoS Biology* 16.9, e2006738 (cit. on p. 7).

Chabas, H., A. Nicot, S. Meaden, E. R. Westra, D. M. Tremblay, L. Pradier, S. Lion, S. Moineau, and S. Gandon (2019). "Variability in the durability of CRISPR-Cas immunity". In: *Philosophical Transactions of the Royal Society B* 374.1772, p. 20180097 (cit. on pp. 2, 3, 8, 16).

Common, J., D. Walker-Sünderhauf, S. van Houte, and E. R. Westra (2020). "Diversity in CRISPR-based immunity protects susceptible genotypes by restricting phage spread and evolution". In: *Journal of Evolutionary Biology* 33.8, pp. 1097–1108 (cit. on p. 7).

Garrison, E. and G. Marth (2012). *Haplotype-based variant detection from short-read sequencing* (cit. on p. 9).

Houte, S. van, A. K. Ekroth, J. M. Broniewski, H. Chabas, B. Ashby, J. Bondy-Denomy, S. Gandon, M. Boots, S. Paterson, A. Buckling, et al. (2016). "The diversity-generating benefits of a prokaryotic adaptive immune system". In: *Nature* 532.7599, pp. 385–388 (cit. on p. 7).

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4, pp. 357–359 (cit. on p. 9).

Lévesque, C., M. Duplessis, J. Labonté, S. Labrie, C. Fremaux, D. Tremblay, and S. Moineau (2005). "Genomic organization and molecular analysis of virulent bacteriophage 2972 infecting an exopolysaccharide-producing *Streptococcus thermophilus* strain". In: *Applied and Environmental Microbiology* 71.7, pp. 4057–4068 (cit. on pp. 8, 9).

Morley, D., J. M. Broniewski, E. R. Westra, A. Buckling, and S. van Houte (2017). "Host diversity limits the evolution of parasite local adaptation". In: *Molecular ecology* 26.7, pp. 1756–1763 (cit. on p. 7).

Nei, M. (1973). "Analysis of gene diversity in subdivided populations". In: *Proceedings of the National Academy of Sciences* 70.12, pp. 3321–3323 (cit. on pp. 6, 15).

Appendix C: An introduction to evolutionary epidemiology theory: Evolution of virulence and transmission

During this PhD, I taught approximately 200 hours at the University of Montpellier in the Biology and Ecology Department, to bachelor and master students. The classes I taught were mostly focused on biostatistics, but also linear algebra and evolutionary genetics.

I also co-developed a 3-hour practical course on evolutionary epidemiology for the Winter school "Quantitative viral dynamics across scales" held in Paris in 2022, organized by Joshua Weitz. This work was done in collaboration with Sylvain Gandon and PhD student Wakinyan Benhamou. The "instructor" version of this class, which includes the expected answers, is presented in this appendix as it squarely fits the subject of this thesis.

Introduction

The aim of the course is to provide an introduction to the analysis of the joint epidemiological and evolutionary dynamics of infectious diseases (*i.e.*, evolutionary epidemiology theory). Throughout the course we have combined an analytic approach with a numerical exploration of the models. The plan is to present/discuss briefly the analytic part and ask the participants to work mainly on the numerical part. The goal is to show how a little bit of analysis can help a lot to interpret numerical simulations.

There will be two main parts:

1. **Epidemiology**

   1.1. **Analytical approach**

   - Introduction of the SIR model.
   - Derivation of the epidemic condition $R_0 > 1$.
   - Derivation of the disease-free equilibrium.
   - Derivation of the endemic equilibrium.

   1.2. **Simulation approach**

   - Presentation of the simulation of the disease-free equilibrium.
   - Simulation of the epidemic until the endemic equilibrium. Validation of the analytical results **(Q1)**.

2. **Evolution**

   2.1 **Dynamics of an epidemic with two pathogens**

   - Modification of the SIR model to account for a polymorphic pathogen population - the wild type and the mutant - **(Q2)**.
   - Simulation of an epidemic with two pathogens **(Q3)**.
   - Analytical derivation from the analysis of the model.
   - Computation of the selection coefficient (s(t)) and the density of susceptible hosts (S(t)) as functions of time **(Q4, Q5)**.

   2.2 **Adaptive dynamics (AD) approach - evolutionary invasion analysis**

   - Condition of invasion when the resident strain is at the endemic equilibrium.
   - Numerical solution for the Evolutionary Stable Strategy (ESS) with a Pairwise Invasibility Plot (PIP) and comparison with the analytical solution **(Q6)**
   - Geometric construction for the ESS

   2.3 **Adaptive dynamics (long term) vs. evolutionary epidemiology (transient epidemic)**

   - We want to show and discuss scenarios where a mutant may transiently outcompete the ESS strategy.
     Test an ESS in a population at endemic equilibrium **(Q7)**. Find a situation where an ESS may transiently be outcompeted; discuss the results **(Q8)**.

1 Epidemiology

1.1 Analytical approach

Let's assume that the dynamics of a host population is governed by the balance between an influx λ of new individuals (birth and immigration) and a natural death rate δ. This host can be infected by a pathogen characterised by three main life-history traits: the horizontal transmission rate β, the mortality rate induced by the infection α (also called the virulence) and the recovery rate γ. The dynamics of this system - a version of the famous **S**usceptible-**I**nfectious-**R**ecovered (SIR) model - can be described by the following set of ordinary differential equations (ODE) where the dot refers to differentiation with respect to time:

$$\begin{aligned} \dot{S}(t) &= \lambda - \beta I(t)S(t) - \delta S(t) \\ \dot{I}(t) &= \beta I(t)S(t) - (\delta + \alpha + \gamma)I(t) \\ \dot{R}(t) &= \gamma I(t) - \delta R(t) \end{aligned} \tag{1}$$

Before analysing the epidemiological dynamics of the pathogen, we need to characterise the host population prior to the introduction of the pathogen. The above system reduces to:

$$\dot{S}(t) = \lambda - \delta S(t)$$

The disease-free equilibrium (sometimes noted DFE) is:

$$S_0 = \frac{\lambda}{\delta}$$

If a pathogen is introduced at the DFE its dynamics will be governed by:

$$\dot{I}(t) = \Big(\beta S_0 - (\delta + \alpha + \gamma)\Big) I(t)$$

The pathogen will grow if and only if $r_0 = \beta S_0 - (\delta + \alpha + \gamma) > 0$, where r_0 is the instantaneous growth rate of the pathogen.

This condition is equivalent to $R_0 = \frac{\beta S_0}{\delta+\alpha+\gamma} > 1$, where R_0 is the basic reproduction number of the pathogen (this is not a rate).

When the above condition is satisfied, the introduction of a small quantity of pathogen will lead to an epidemic that will eventually reach an endemic equilibrium:

$$\begin{aligned} S_e &= \frac{\delta + \alpha + \gamma}{\beta} \\ I_e &= \frac{\lambda\beta - \delta(\delta + \alpha + \gamma)}{\beta(\delta + \alpha + \gamma)} \\ R_e &= \frac{\gamma}{\delta} I_e \end{aligned}$$

1.2 Simulation approach

```
# Cleaning objects from the workplace
rm(list=ls())

# Packages (may first require installations: install.packages("name of the package"))
library(tidyverse)
library(ggplot2)
library(cowplot)
library(deSolve)
library(scales)
library(lattice)
library(knitr)
```

```
ODE_SIR <- function(t, y, parms){

  # t, the current time
  # y, the current state of the system (/!\ to the order of the state variables)
  # parms, the parameters of the model

  # State variables
  S <- y[1]
  I <- y[2]
  R <- y[3]

  # Parameters
  lambda <- parms["lambda"]
  delta <- parms["delta"]
  beta <- parms["beta"]
  alpha <- parms["alpha"]
  gamma <- parms["gamma"]

  # Temporal derivatives
  dS <- lambda - delta*S - beta*I*S
  dI <- (beta*S - (delta + alpha + gamma))*I
  dR <- gamma*I - delta*R

  result <- c(dS, dI, dR)

  # Return
  list(result)
}
```

```
# Time points

t0 <- 0 # initial time
tf <- 10 # final time
times <- seq(from=t0, to=tf, by=0.1)

# Parameters

lambda = 1
delta = 1
```

```r
beta = 5
gamma = 0.1
alpha = 0.1

parms = c("lambda"=lambda, "delta"=delta, "beta"=beta, "alpha"=alpha, "gamma"=gamma)
```

1.2.1 Disease-free population

```r
# Initialization of each compartment (at time t = t0)

init_disease_free <- c("S" = 0.1, # S(t0), all the population is susceptible (S) to the disease
                       "I" = 0,   # I(t0), disease-free population
                       "R" = 0)   # R(t0), no recovered (R) individuals
```

Numerical integration

```r
simul_disease_free <- lsoda(y = init_disease_free, times = times, func = ODE_SIR, parms = parms)

head(simul_disease_free, n = 2) # 2 first rows of the table
```

```
##      time         S I R
## [1,]  0.0 0.1000000 0 0
## [2,]  0.1 0.1856454 0 0
```

```r
tail(simul_disease_free, n = 2) # 2 last rows of the table
```

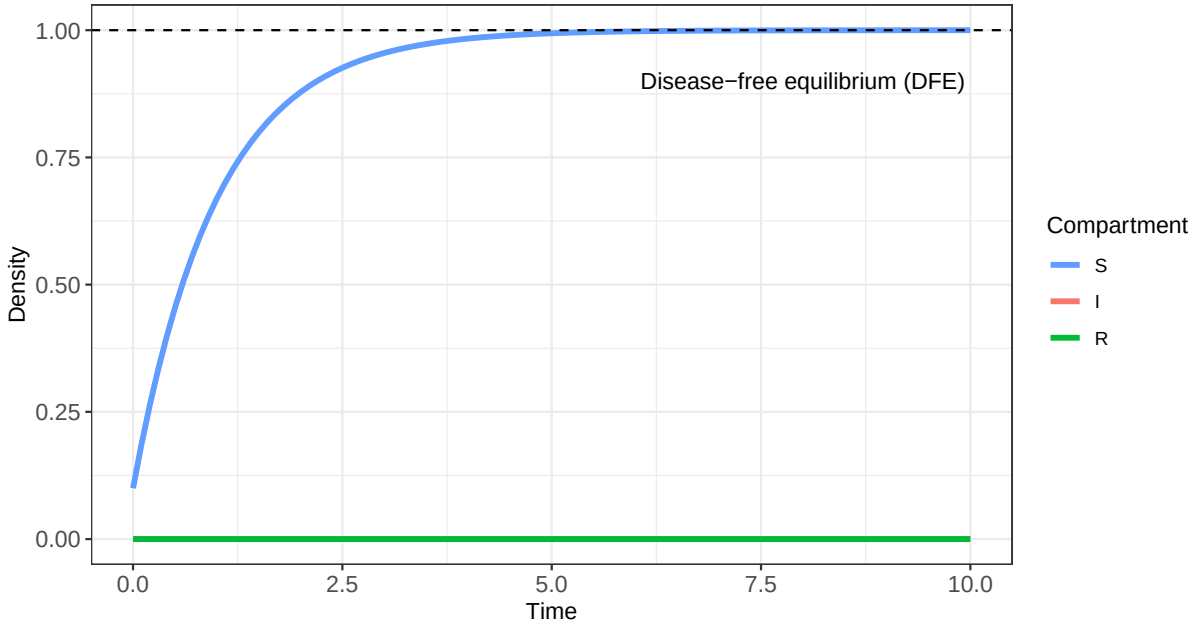
```
##        time         S I R
## [100,]  9.9 0.9999548 0 0
## [101,] 10.0 0.9999591 0 0
```

Formatting of simulated data & graphical visualization

```r
plot_simul <- function(simul, title = element_blank(), parms = NULL){

  data <- data.frame("Time" = simul[,1] %>% rep(3),
                     "Compartment" = c("S", "I", "R") %>% rep(each = dim(simul)[1]),
                     "Density" = simul[,-1] %>% c)
  data$Compartment <- factor(data$Compartment, levels = c("S", "I", "R"))

  # Other possibility (more advanced in R):
  #
  # data <- simul %>% as.data.frame %>% tidyr::gather(Compartment, Density, -time) %>%
  #   dplyr::mutate(Compartment = factor(Compartment, labels = c("S", "I", "R"))) %>%
  #   dplyr::rename(Time = time)

  caption <- ifelse(is.null(parms), yes = "",
              no = paste("\n Parameters:", paste(names(parms), parms, sep = " = ", collapse = " ; ")))

  return(ggplot(data, aes(x = Time, y = Density, color = Compartment)) +
```

```
            geom_line(cex = 1.3) +
            labs(title = title, caption = caption) +
            theme_bw() +
            scale_color_manual(values = c("#619CFF", "#F8766D", "#00BA38")) +
            theme(axis.text.x = element_text(size=11),
                  axis.text.y = element_text(size=11),
                  plot.caption = element_text(face = 'bold')))
}
plot_simul(simul_disease_free,
            title = "Fig. 1. Simulation of the SIR model (1) for a disease-free population\n",
            parms = parms) +
  geom_hline(yintercept = lambda/delta, lty = 'dashed') + # disease-free equilibrium for S
  annotate(geom="text", x=8, y=0.9*(lambda/delta), label="Disease-free equilibrium (DFE)")
```

Fig. 1. Simulation of the SIR model (1) for a disease−free population



1.2.2 Introduction of a low initial density of infected/infectious individuals in a population at the disease-free equilibrium

Q1. Use the code given above, adding a low initial density of infected individuals to find the endemic equilibrium - *i.e.* the values of S_e, I_e and R_e. Compare your results to the expected analytical values.

```
# Initialization of each compartment (at time t = t0)

I_t0 <- 0.001 # I(t0), (low) initial density of I

init_disease <- c("S" = (lambda/delta)-I_t0,
                  # S(t0), almost all the population is susceptible (S) at the DFE
```

```
                  "I" = I_t0, # I(t0), low initial density of infected (I) individuals
                  "R" = 0)    # R(t0), no recovered (R) individuals
```

Numerical integration

```
simul_disease <- lsoda(y = init_disease, times = times, func = ODE_SIR, parms = parms)

head(simul_disease, n = 2) # 2 first rows of the table
```

```
##      time         S           I            R
## [1,]  0.0 0.9990000 0.001000000 0.000000e+00
## [2,]  0.1 0.9985145 0.001462207 1.162723e-05
```

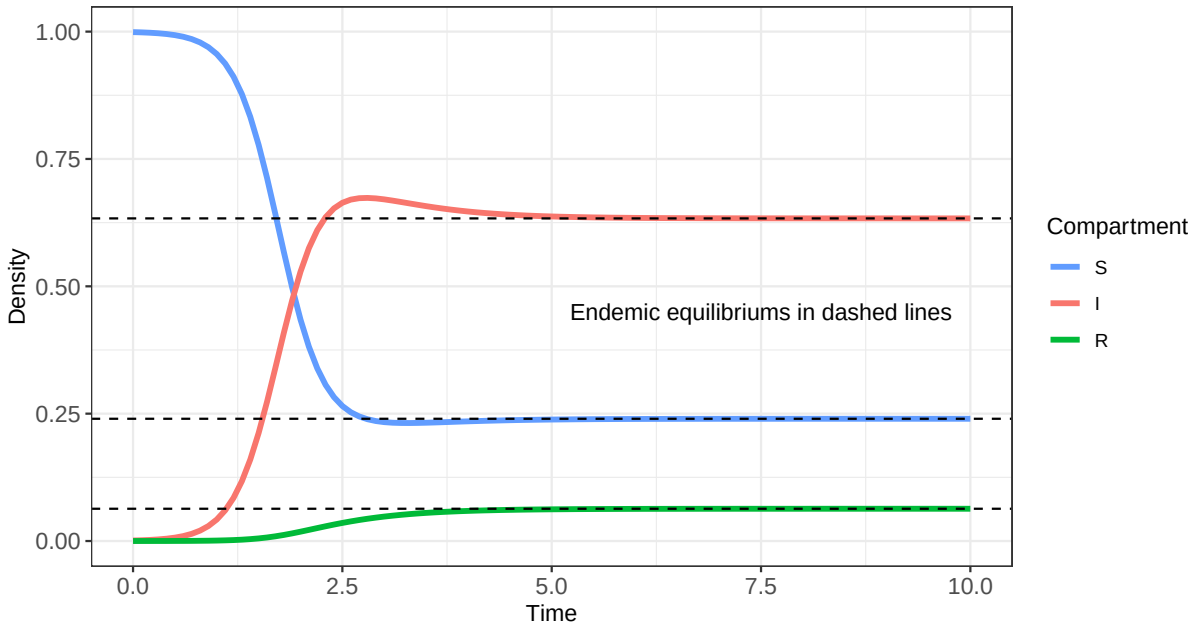
```
tail(simul_disease, n = 2) # 2 last rows of the table
```

```
##        time         S         I          R
## [100,]  9.9 0.2399978 0.6333386 0.06333182
## [101,] 10.0 0.2399980 0.6333379 0.06333201
```

Formatting of simulated data & graphical visualization

```
plot_simul(simul_disease, title = "Fig. 2. Simulation of the SIR model (1)\n", parms = parms) +
  geom_hline(yintercept = c((delta+alpha+gamma)/beta,              # endemic equilibrium for S,
                            lambda/(delta+alpha+gamma) - delta/beta,                      # I,
                            (gamma/delta)*(lambda/(delta+alpha+gamma) - delta/beta)), # and R
             lty = 'dashed') +
  annotate(geom="text", x=7.5, y=0.45, label="Endemic equilibriums in dashed lines")
```

Fig. 2. Simulation of the SIR model (1)



Parameters: lambda = 1 ; delta = 1 ; beta = 5 ; alpha = 0.1 ; gamma = 0.1

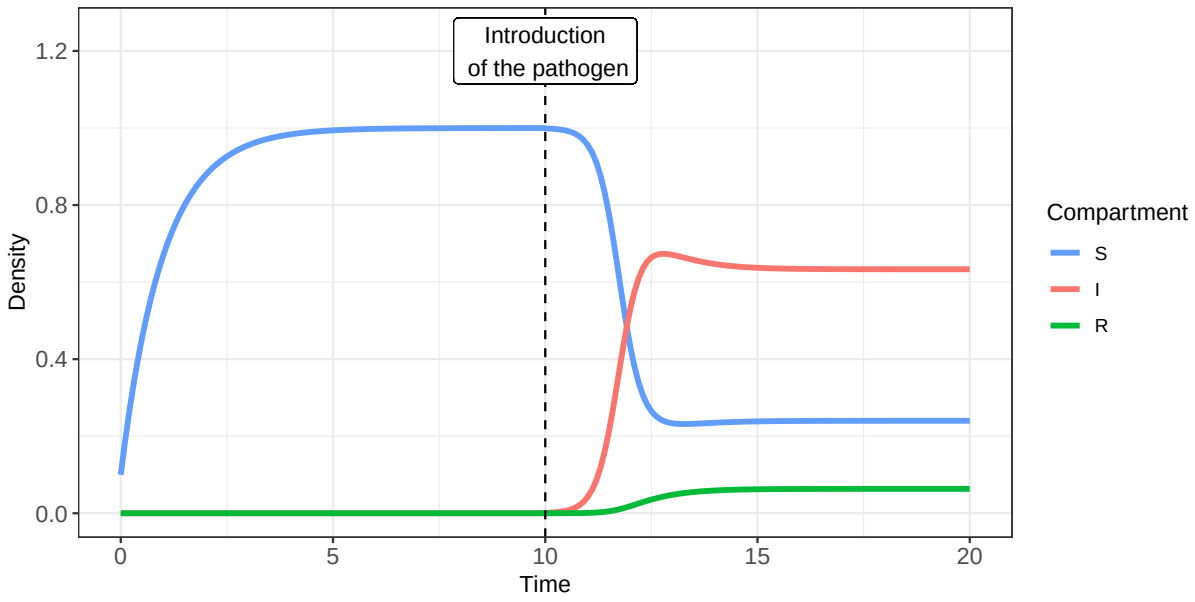
As expected from the analysis of the model, the density of infected hosts increases because $R_0 = 4.16 > 1$. After a transient phase, the dynamical variables $S(t)$, $I(t)$ and $R(t)$ converge toward the equilibrium values derived above (*i.e.*, S_e, I_e and R_e).

1.2.3 Overview

To sum up this section, **Fig. 3** shows the establishment of the disease-free equilibrium, then the introduction of a small density of infected individuals, eventually leading to the endemic equilibrium.

```
n_row_df <- dim(simul_disease_free)[1]

simul_disease[,1] <- simul_disease_free[n_row_df,1] + simul_disease[,1]

plot_simul(simul = rbind(simul_disease_free[-n_row_df,], simul_disease),
           title = "Fig. 3. Overview of the simulations of the SIR model (1)
           before and after the introduction of the pathogen\n", parms = parms) +

  geom_vline(xintercept = simul_disease[1,1], lty = 'dashed') +
  geom_label(aes(x = simul_disease[1,1], y = 1.2*(lambda/delta),
                 label = "Introduction\n of the pathogen"), fill = "white", col = 'black') +
  ylim(c(0, 1.25*(lambda/delta)))
```

Fig. 3. Overview of the simulations of the SIR model (1)
before and after the introduction of the pathogen




```
rm(list=ls()) # Cleaning objects from the workplace
```

2 Evolution

2.1 Dynamics of an epidemic with two pathogens

2.1.1 Analytical approach

Let's assume that a new variant appears by mutation. Will this mutant invade and replace the previously dominant form of the pathogen?

To answer this question we need to account for the circulation of this new variant which requires a new system of ODE:

Q2. Write the system of ODE describing the epidemiological dynamics of two pathogenic strains, respectively with parameters (β, α, γ) and $(\beta_m, \alpha_m, \gamma_m)$

$$\begin{aligned}
\dot{S}(t) &= \lambda - \beta I(t)S(t) - \beta_m I_m(t)S(t) - \delta S(t) \\
\dot{I}(t) &= \underbrace{(\beta S(t) - (\delta + \alpha + \gamma))}_{r(t)} I(t) \\
\dot{I_m}(t) &= \underbrace{(\beta_m S(t) - (\delta + \alpha_m + \gamma_m))}_{r_m(t)} I_m(t) \\
\dot{R}(t) &= \gamma I(t) + \gamma_m I_m(t) - \delta R(t)
\end{aligned} \tag{2}$$

Adding one strain requires an additional equation but do not forget to modify the other equations as the presence of the mutant is also affecting the dynamics of $S(t)$ and $R(t)$.

2.1.2 Simulation approach

Q3. Using a modified version of the earlier code, simulate the epidemiological dynamics dictated by this new system of ODE. Describe the dynamics of the two infected compartments. Did you expect this behaviour?

```
ODE_SIR.2 <- function(t, y, parms){

  # t, the current time
  # y, the current state of the system (/!\ to the order of the state variables)
  # parms, the parameters of the model

  # State variables
  S <- y[1]
  I <- y[2]
  I_m <- y[3]
  R <- y[4]

  # Parameters
  lambda <- parms["lambda"]
  delta <- parms["delta"]
  beta <- parms["beta"]
  alpha <- parms["alpha"]
  gamma <- parms["gamma"]
  beta_m <- parms["beta_m"]
```

```
  alpha_m <- parms["alpha_m"]
  gamma_m <- parms["gamma_m"]

  # Temporal derivatives
  dS <- lambda - delta*S - (beta*I + beta_m*I_m)*S
  dI <- (beta*S - (delta + alpha + gamma))*I
  dI_m <- (beta_m*S - (delta + alpha_m + gamma_m))*I_m
  dR <- gamma*I + gamma_m*I_m - delta*R

  result <- c(dS, dI, dI_m, dR)

  # Return
  list(result)
}
```

```
# Time points
t0 <- 0 # initial time
tf <- 15 # final time
times <- seq(from=t0, to=tf, by=0.01)

# Initialization of each compartment (at time t = t0)
I_t0 <- 0.001 # I(t0), initial density of I
I_m_t0 <- 0.001 # I_m(t0), initial density of I_m
I_T_t0 <- I_t0 + I_m_t0

init <- c("S" = 1-I_T_t0,    # S(t0)
          "I" = I_t0,        # I(t0), individuals initially infected by the WT strain (ancestral)
          "I_m" = I_m_t0,    # I_m(t0), individuals initially infected by the variant
          "R" = 0)           # R(t0)

# Parameters
lambda = 1
delta = 1
beta = 10.5
gamma = 0.1
alpha = 1.1
beta_m = 12
gamma_m = 0.1
alpha_m = 1.5

parms = c("lambda"=lambda, "delta"=delta, "beta"=beta, "alpha"=alpha, "gamma"=gamma,
          "beta_m"=beta_m, "alpha_m"=alpha_m, "gamma_m"=gamma_m)
```

Numerical integration

```
simul <- lsoda(y = init, times = times, func = ODE_SIR.2, parms = parms)

head(simul, n = 2) # 2 first rows of the table
```

```
##      time         S           I         I_m            R
## [1,] 0.00 0.9980000 0.001000000 0.001000000 0.000000e+00
## [2,] 0.01 0.9977861 0.001086352 0.001098356 2.081938e-06
```

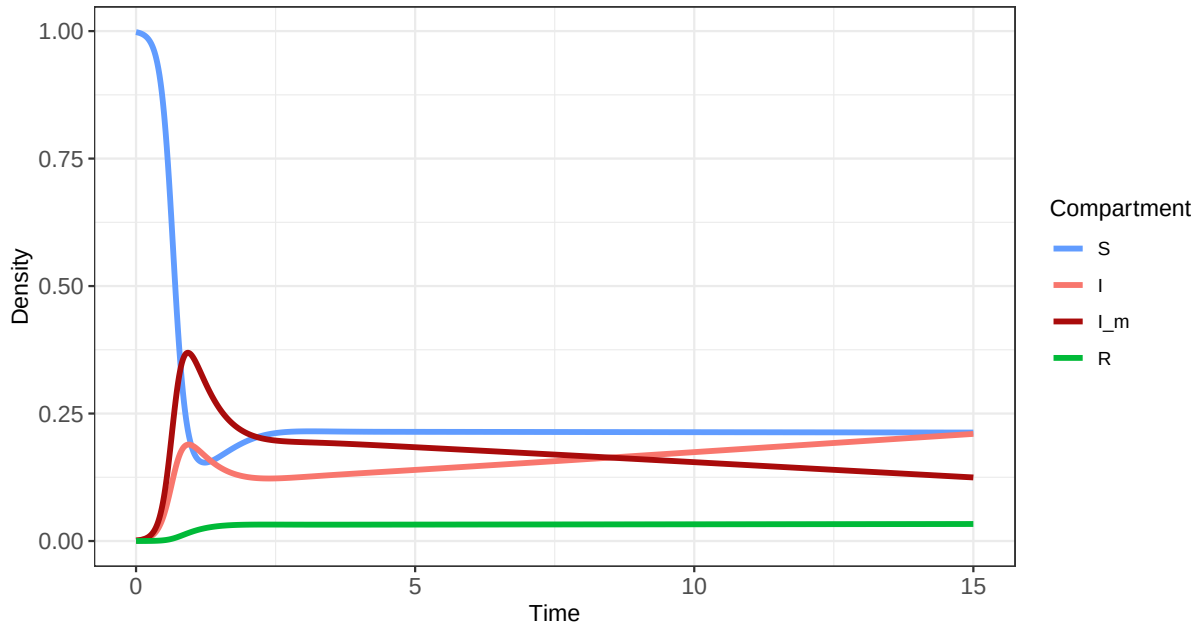
```r
tail(simul, n = 2) # 2 last rows of the table
```

```
##          time         S         I       I_m          R
## [1500,] 14.99 0.2127208 0.2099941 0.1247275 0.03335710
## [1501,] 15.00 0.2127194 0.2100646 0.1246684 0.03335825
```

Formatting of simulated data & graphical visualization

```r
plot_simul.2 <- function(simul, title = element_blank(), parms = NULL){

  compartments <- colnames(simul)[2:5]
  data <- data.frame("Time" = simul[,1] %>% rep(4),
                     "Compartment" = compartments %>% rep(each = dim(simul)[1]),
                     "Density" = simul[,2:5] %>% c)
  data$Compartment <- factor(data$Compartment, levels = compartments)
  caption <- ifelse(is.null(parms), yes = "",
              no = paste("\n Parameters:", paste(names(parms), parms, sep = " = ", collapse = " ; ")))

  return(ggplot(data, aes(x = Time, y = Density, color = Compartment)) +
           geom_line(cex = 1.3) +
           labs(title = title, caption = caption) +
           theme_bw() +
           scale_color_manual(values = c("#619CFF", "#F8766D", "#A90B0B", "#00BA38")) +
           theme(axis.text.x = element_text(size=11),
                 axis.text.y = element_text(size=11),
                 plot.caption = element_text(hjust = 0, face = 'bold')))
}
plot_simul.2(simul, title = "Fig. 4. Simulation of the SIR model (2)\n", parms = parms)
```

Fig. 4. Simulation of the SIR model (2)



Parameters: lambda = 1 ; delta = 1 ; beta = 10.5 ; alpha = 1.1 ; gamma = 0.1 ; beta_m = 12 ; alpha_m = 1.5 ; gamma_m = 0.1

In the simulation example presented in **Fig. 4**, both strains are introduced at very low densities with a 1:1 ratio and the variant (or mutant strain m) differs from the ancestral strain by a higher transmission rate and a higher virulence. In this case, the mutant strain grows much faster at the beginning of the epidemic but is then gradually replaced by the ancestral strain which dominates from $t = 8.5$. However, note that a change in $I(t)$ or in the parameter values - *e.g.* the traits of the mutant, the initial densities - can have a dramatic impact on the dynamics.

2.1.3 Population genetics approach - derivation of the selection coefficient

At this stage and to understand these dynamics, it is useful to rewrite the above system of 4 equations (2) in the following way:

$$\begin{aligned} \dot{S}(t) &= \lambda - \overline{\beta}(t) I_T(t) S(t) - \delta S(t) \\ \dot{I}_T(t) &= \overline{\beta}(t) I_T(t) S(t) - (\delta + \overline{\alpha}(t) + \overline{\gamma}(t)) I_T(t) \\ \dot{R}(t) &= \overline{\gamma}(t) I_T(t) - \delta R(t) \end{aligned} \tag{3a}$$

$$\text{where } I_T(t) = I(t) + I_m(t) \quad \text{and} \quad \begin{aligned} \overline{\beta}(t) &= (1 - p_m(t))\beta + p_m(t)\beta_m \\ \overline{\alpha}(t) &= (1 - p_m(t))\alpha + p_m(t)\alpha_m \\ \overline{\gamma}(t) &= (1 - p_m(t))\gamma + p_m(t)\gamma_m \end{aligned} \quad \text{with } p_m(t) = \frac{I_m(t)}{I_T(t)}$$

$$\dot{p}_m(t) = \underbrace{p_m(t)(1 - p_m(t))}_{\text{genetic variance}} \underbrace{(r_m(t) - r(t))}_{\text{selection coefficient}} \tag{3b}$$

Note again that (2) and (3) are equivalent but the second formulation decoupled epidemiological dynamics (3a) and evolutionary dynamics (3b). In particular, it is insightful to examine the selection coefficient $s(t) = r_m(t) - r(t)$ (*Day & Gandon*). To understand the effect of each life-history trait, it is important to write the selection coefficient as:

$$s(t) = (\beta_m - \beta) S(t) + (\alpha + \gamma) - (\alpha_m + \gamma_m) \tag{4}$$

Strains favoured by selection:

- Larger transmission rate
- Lower virulence rate
- Lower recovery rate

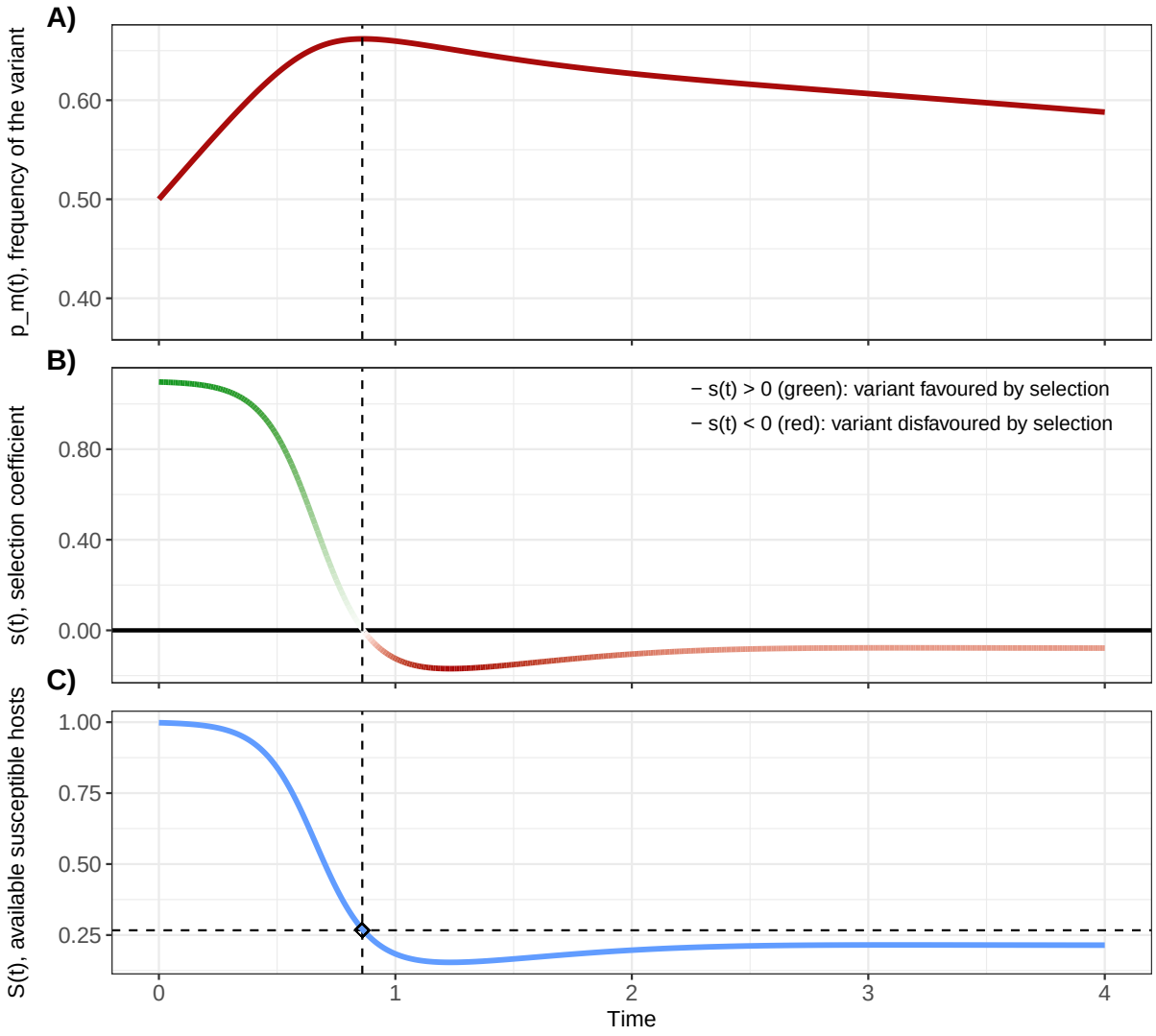
Note that the first term acts on the production of new infections (*i.e.* birth rate of the infection) while the last two points act on the duration of infection (*i.e.* lower death rate of the infection).

Q4. To understand the dynamics of the two pathogenic strains, plot the frequency $p_m(t)$ as well as the selection coefficient $s(t)$ each as a function of time.

```r
simul <- simul %>% as.data.frame

simul$p_m <- simul$I_m / (simul$I_m+simul$I) # compute p_m(t)

simul$selection_coef <- (beta_m-beta)*simul$S+(alpha+gamma)-(alpha_m+gamma_m) # compute s(t)

s_threshold_index <- simul$selection_coef %>% abs %>% which.min
# Index of the value of s(t) closest to 0 in our simulation

S_threshold <- ((alpha_m+gamma_m)-(alpha+gamma))/(beta_m-beta)
# Analytical value of S(t) such that the selection coefficient of the variant is: s(t) = 0

plot_grid(

  ggdraw() + draw_label(
    "Fig. 5. Temporal dynamics of the frequency (A) and of the selection coefficient (B) of the variant
           and of the density of available hosts (C) based on a simulation of the SIR model (2)-(3)\n",
    x = 0.025, hjust = 0, size = 13),

  ggplot(simul %>% as.data.frame, aes(x = time, y = p_m)) +
    geom_line(cex = 1.3, col = "#A90B0B") +
    geom_vline(xintercept = simul[s_threshold_index, 1], lty = 'dashed') +
    labs(x = "Time", y =  "p_m(t), frequency of the variant\n") +
    scale_y_continuous(labels = scales::label_number(accuracy = 0.01)) +
    xlim(c(0,4)) +
    theme_bw() +
    theme(axis.text.x = element_blank(), axis.title.x = element_blank(),
          axis.text.y = element_text(size=11)),

  ggplot(data = simul, aes(x = time, y = selection_coef, color = selection_coef)) +
    geom_hline(yintercept = 0, cex = 1) +
    geom_line(cex = 1.3) +
    geom_vline(xintercept = simul[s_threshold_index, 1], lty = 'dashed') +
    labs(y = "s(t), selection coefficient\n") +
    scale_color_gradientn(colors = c("#AB0707", "white", "#169822"),
                          values = rescale(c(min(simul$selection_coef), 0,
                                             max(simul$selection_coef)))) +
    scale_y_continuous(labels = scales::label_number(accuracy = 0.01)) +
    annotate(geom="text", label = "- s(t) > 0 (green): variant favoured by selection",
             x = 2.25, y = 1.067, size = 3.5, hjust = 0) +
    annotate(geom="text", label = "- s(t) < 0 (red): variant disfavoured by selection",
             x = 2.25, y = 0.917, size = 3.5, hjust = 0) +
    xlim(c(0,4)) +
    theme_bw() +
    theme(axis.text.x = element_blank(), axis.title.x = element_blank(),
          axis.text.y = element_text(size=11), legend.position = 'none'),

  ggplot(data = simul, aes(x = time, y = S)) +
    geom_hline(yintercept = S_threshold, lty = 'dashed') +
    geom_vline(xintercept = simul[s_threshold_index, 1], lty = 'dashed') +
    geom_line(cex = 1.3, col = "#619CFF") +
    geom_point(x = simul[s_threshold_index, 1], y = S_threshold, pch = 5, size = 2) +
    scale_y_continuous(labels = scales::label_number(accuracy = 0.01)) +
```

```
  labs(x = "Time", y = "S(t), available susceptible hosts\n") +
  xlim(c(0,4)) +
  theme_bw() +
  theme(axis.text.x = element_text(size=11), axis.text.y = element_text(size=11),
        plot.caption = element_text(hjust = 0, face = 'bold')),

ggdraw() + draw_label(paste("Parameters:",
                            paste(names(parms), parms, sep = ' = ', collapse = ' ; ')),
                      size = 9, fontface = 'bold'),

labels = c("", "A)", "B)", "C)", ""), label_x = 0.03, label_y = c(0, 1.05, 1.05, 1.12, 0),
ncol = 1, rel_heights = c(0.3, 1, 1, 1, 0.1))
```

Fig. 5. Temporal dynamics of the frequency (A) and of the selection coefficient (B) of the variant and of the density of available hosts (C) based on a simulation of the SIR model (2)−(3)



Parameters: lambda = 1 ; delta = 1 ; beta = 10.5 ; alpha = 1.1 ; gamma = 0.1 ; beta_m = 12 ; alpha_m = 1.5 ; gamma_m = 0.1

Fig. 5-A and **B** show that the frequency of the variant increases at the beginning of the epidemic and then gradually decreases (in this example, the maximum is reached around $t = 0.86$). When the frequency of the variant increases, its selection coefficient is positive (the variant has a selective advantage). When the variant no longer increases in frequency, its selection coefficient is zero. Eventually, when the variant is progressively replaced by the other strain - *i.e.* the variant decreases in frequency -, its selection coefficient becomes negative (and its value reflects the speed of this decay).

We added here the temporal dynamics of the S compartment (*cf.* **Fig. 5-C**). Note how the dynamics of $S(t)$ mirrors the dynamics of the selection coefficient. A particular value of $S(t)$ is associated with the time point when $s(t) = 0$ (*i.e.*, when the variant reaches its maximum frequency).

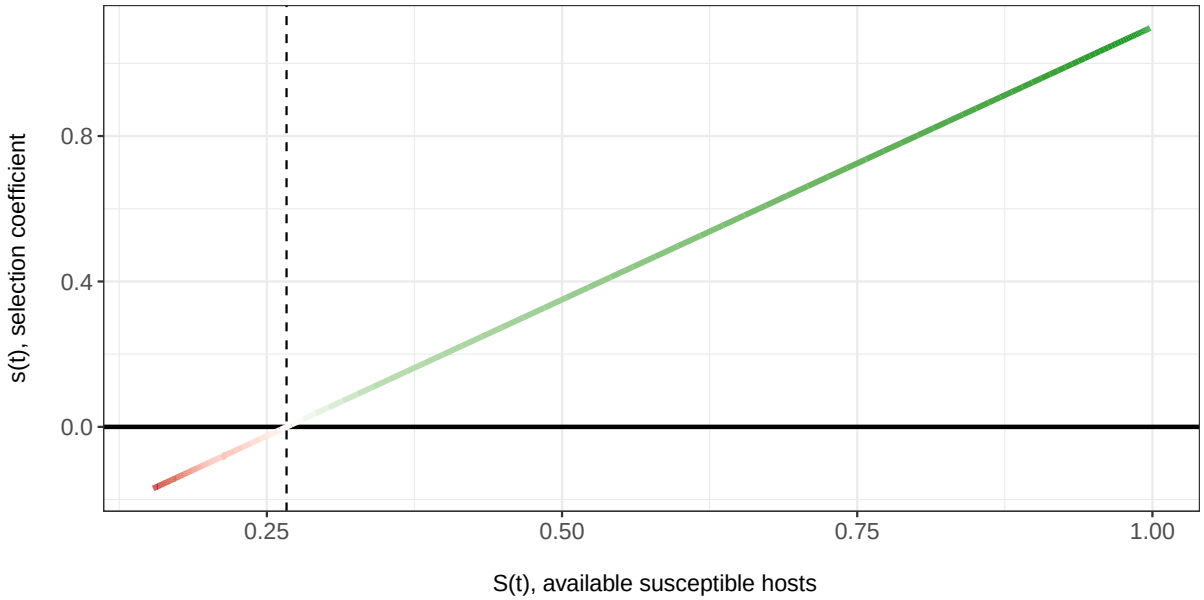
Q5. Find the threshold value of $S(t)$ for which the more selected strain changes, both analytically and graphically (with a plot $s(t) = f(S(t))$).

The coefficient of selection of the mutant changes when $s(t) = 0$. Thus, we can define $\hat{S}(t)$ the threshold value of $S(t)$ where $s(t) = 0$ and using equation (4):

$$s(t) = 0 \iff (\beta_m - \beta)\hat{S}(t) + (\alpha + \gamma) - (\alpha_m + \gamma_m) = 0 \iff \hat{S}(t) = \frac{(\alpha_m + \gamma_m) - (\alpha + \gamma)}{\beta_m - \beta}$$

```
ggplot(data = simul, aes(x = S, y = selection_coef, col = selection_coef)) +
  geom_hline(yintercept = 0, cex = 1) +
  geom_line(cex = 1.3) +
  labs(x = "\n S(t), available susceptible hosts", y = "s(t), selection coefficient\n",
       title = "Fig. 6. Selection coefficient of the variant against the density of available hosts",
       subtitle= "Based on a simulation of the SIR model (2)-(3)\n",
       caption =  paste("\n Parameters:", paste(names(parms), parms,
                                                sep = " = ", collapse = " ; "))) +
  geom_vline(xintercept = S_threshold, lty = 'dashed') +
  scale_color_gradientn(colors = c("#C31515", "white", "#169822"),
                        values = rescale(c(min(simul$selection_coef), -0.1, 0.1,
                                           max(simul$selection_coef)))) +
  theme_bw() +
  theme(axis.text.x = element_text(size=11),
        axis.text.y = element_text(size=11),
        legend.position = 'none',
        plot.caption = element_text(hjust = 0, face = 'bold'))
```

Fig. 6. Selection coefficient of the variant against the density of available hosts
Based on a simulation of the SIR model (2)−(3)



Parameters: lambda = 1 ; delta = 1 ; beta = 10.5 ; alpha = 1.1 ; gamma = 0.1 ; beta_m = 12 ; alpha_m = 1.5 ; gamma_m = 0.1

The selection coefficient of the variant $s(t)$ is a linear function of $S(t)$ as shown in **Fig. 6** which is consistent with (4). Here, the threshold density $\hat{S}(t)$ is about 0.27. Below this threshold, the selection coefficient is negative (*i.e.* the variant is selected against), above, the selection coefficient is positive (*i.e.* the variant is selected for). This is because this variant is more transmissible but more virulent than the other strain. As shown in (4), this transmission advantage depends on the number of available hosts ($S(t)$). The selective advantage of this kind of variant changes with the availability of susceptible hosts $S(t)$. When there are no longer enough susceptible hosts - *i.e.* below the calculated threshold density $\hat{S}(t)$ -, the virulence burden is no longer compensated by the transmission advantage and the frequency of the variant drops.

```
rm(list = setdiff(ls(), lsf.str())) # Cleaning objects from the workplace except for the functions
```

2.2 Adaptive dynamics (AD) approach - evolutionary invasion analysis

2.2.1 Analytical approach

Another classical approach to model the evolution of life-history is to focus on a situation where the mutant is introduced when the epidemiological system is at the endemic equilibrium. This assumption makes sense when the mutation rate is assumed to be very small. In this case, the epidemiology reaches the endemic equilibrium before a new variant is introduced by mutation. In this case $r = 0$ and $r_m = \beta_m S_e - (\delta + \alpha_m + \gamma_m)$. In other words, the mutant can invade if and only if: $r_m > 0$ which yields:

$$\frac{\beta_m}{\delta + \alpha_m + \gamma_m} > \frac{\beta}{\delta + \alpha + \gamma} \tag{5}$$

This condition is particularly useful when we want to assume some covariation among different life-history traits (*e.g.*, trade-off between transmission and virulence: impossible to increase transmission without higher exploitation of the host). In this case, one can write the transmission rate as an increasing function of virulence: $\beta(\alpha)$. Here we propose to use the trade-off function: $\beta(\alpha) = 10\sqrt{\alpha}$

The condition (5) means that adaptation is maximizing: $R(\alpha) = \frac{\beta(\alpha)}{\delta+\alpha+\gamma}$

The strategy α^* that maximizes this ratio must verify:

$$\frac{dR(\alpha)}{d\alpha} = 0 \tag{6}$$

$$\text{and} \qquad \frac{d^2R(\alpha)}{d\alpha^2} < 0$$

After some rearrangements (6) yields the following condition:

$$\frac{d\beta(\alpha)}{d\alpha} = \frac{\beta(\alpha)}{\delta + \alpha + \gamma} \tag{7}$$

For the special case where $\beta(\alpha) = 10\sqrt{\alpha}$, one can find that:

$$\alpha^* = \delta + \gamma$$

2.2.2 Numerical approach

Q6. Using the condition (5) and the trade-off function $\beta(\alpha) = 10\sqrt{\alpha}$, find if possible the parameters β^* and α^* of a strain which cannot be invaded by any other strain. This strain is said to be at an Evolutionary Stable Strategy (ESS). Compare your numerical approximation of α^* with with the analytical solution. For the sake of simplicity, use the following function for the trade-off:

```
Trade_off <- function(alpha, k=10, c=1/2){ # Concave relationship between transmission and virulence
  return(k*alpha^c) # = beta(alpha)
}
```

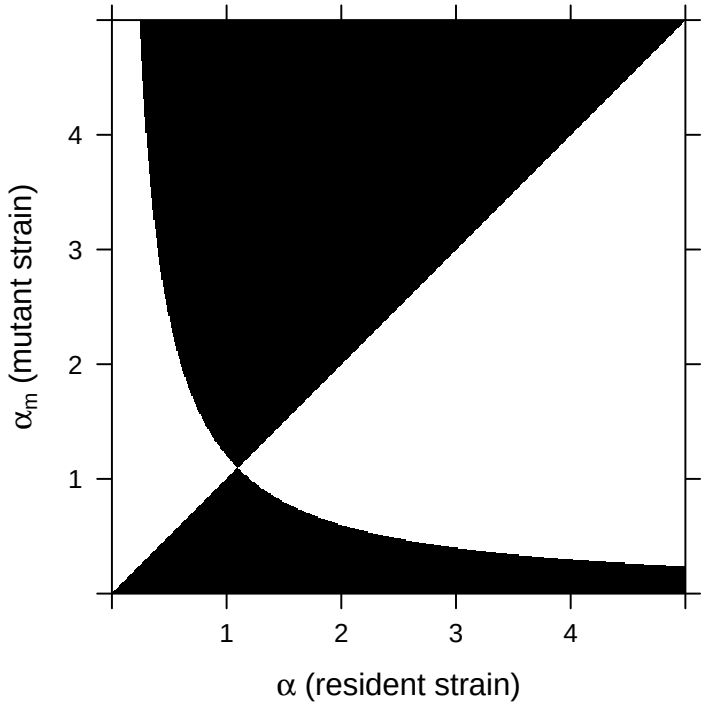
```
# Parameters

k <- 10
c <- 0.5

lambda <- 1
```

```r
delta <- 1
gamma <- gamma_m <- 0.1

alpha_vec <- seq(from=0, to=5, length.out = 500)
n_alpha <- length(alpha_vec)
```

Pairwise comparisons

```r
PIP <- matrix(ncol = n_alpha, nrow = n_alpha) # matrix for Pairwise Invasibility Plot
diag(PIP) <- 0 # A variant cannot invade the resident strain with the same strategy

for(i in 1:(n_alpha-1)){

  # Resident strain
  alpha <- alpha_vec[i] # Virulence
  beta <- Trade_off(alpha, k, c) # Transmission rate using the trade-off function

  for(j in (i+1):n_alpha){

    # Variant / Mutant strain
    alpha_m <- alpha_vec[j] # Virulence
    beta_m <- Trade_off(alpha_m, k, c) # Transmission rate using the trade-off function

    # Eventually, can the mutant invade the resident strain: r_m > r ?
    invasion <- ifelse(beta_m/(delta+alpha_m+gamma_m) > beta/(delta+alpha+gamma), # cf. equation (5)
                       yes = 1, no = 0)
    PIP[i,j] <- invasion
    PIP[j,i] <- 1-invasion
  }
}
lim <- c(alpha_vec[1], alpha_vec[n_alpha])

levelplot(PIP, row.values = alpha_vec, column.values = alpha_vec, xlim = lim, ylim = lim,
          colorkey = FALSE, col.regions = c('black', 'white'),
          xlab = expression(paste(alpha, " (resident strain)")),
          ylab = expression(paste(alpha[m], " (mutant strain)")),
          main = list(label = "Fig. 7. Pairwise Invasibility Plot based on the the SIR model (2)-(3)",
                      cex = 1, font = 'plain'))
```

Fig. 7. Pairwise Invasibility Plot based on the the SIR model (2)−(3)




```
# Looking for the ESS (Evolutionary Stable Strategy)
ESS_index <- which(apply(PIP, 1, sum) == 0) # Only row with only '0'

if(length(ESS_index) == 0){
  print("No Evolutionary Stable Strategy (ESS)")

}else{
  alpha_approx_ESS <- alpha_vec[ESS_index]

  tab <- c(alpha_approx_ESS, (alpha_vec[n_alpha]-alpha_vec[1])/(2*(n_alpha-1))) %>% round(3) %>%
                                    paste(collapse=" +/- ") %>% c(delta+gamma) %>% as.data.frame
  colnames(tab) <- "$\\alpha^{*} (time^{-1})$"
  rownames(tab) <- c("Numerical approximation", "Analytical solution")
  kable(tab)
}
```

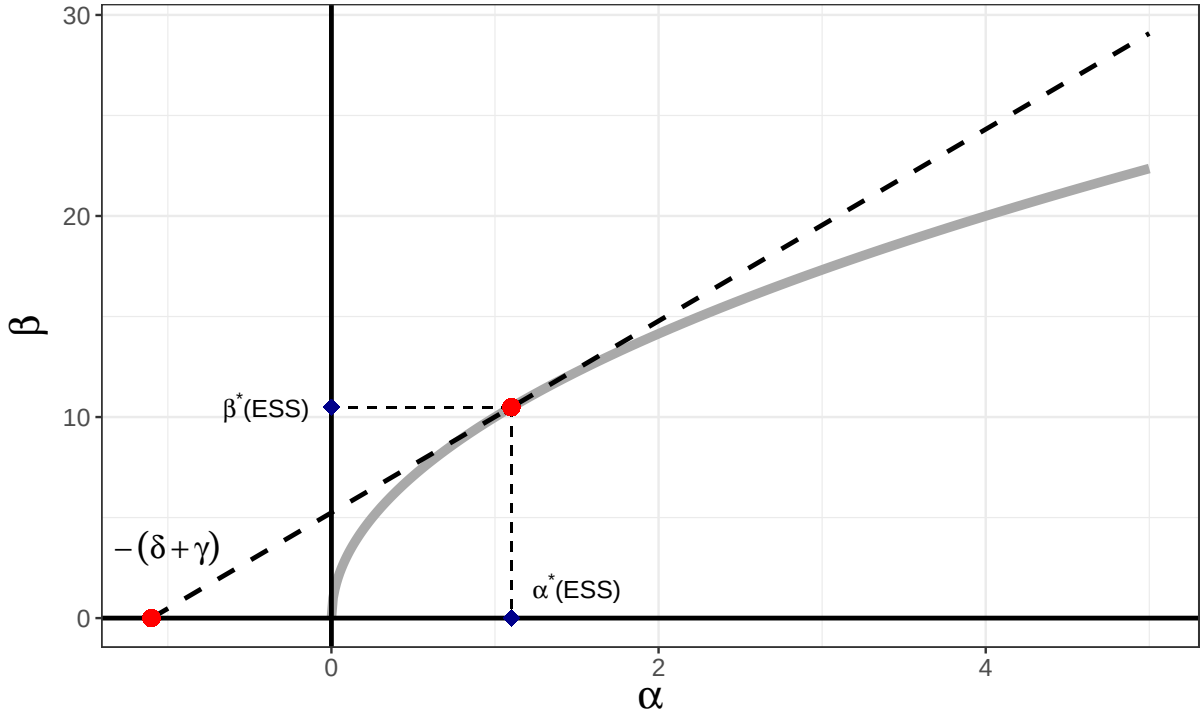
	$\alpha^*(time^{-1})$
Numerical approximation	1.102 +/- 0.005
Analytical solution	1.1

The virulence ESS α^* may be found graphically on a Pairwise Invasibility Plot (PIP) where the diagonal is intersected by the other boundary of the regions associated with an invasion of the resident strain (in white). In this example, the PIP allows us to obtain a good approximation for virulence: $\alpha^* \approx 1.1$.

2.2.3 Geometric construction

Let's simply note here that equation (7) yields a very useful geometric representation that one can use to study the effect of various parameters on the evolutionary stable virulence strategy.

Fig. 8. Geometric construction to find the Evolutionary Stable Strategy (ESS)



2.3 Adaptive dynamics (long term) vs. evolutionary epidemiology (transient epidemic)

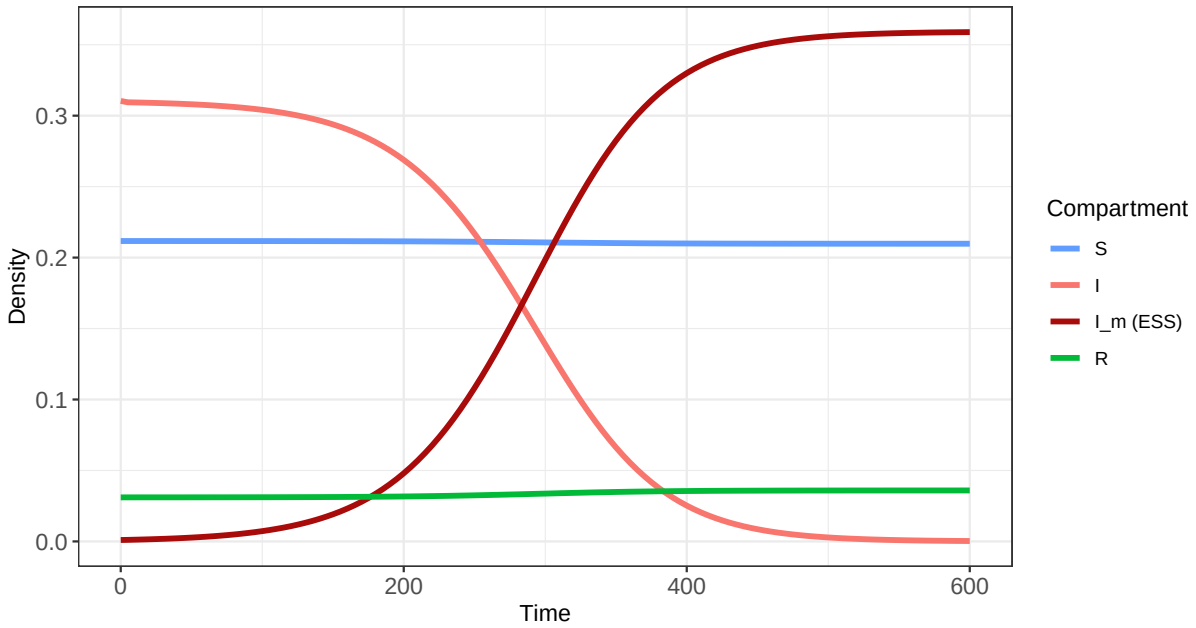
2.3.1 The ESS wins in the long term...

Q7. Starting from the endemic equilibrium of any pathogen with a strategy different from the ESS (Evolutionary Stable Strategy), check with some simulations that it is always invaded by the ESS pathogen (both strains following the same trade-off function).

```
# Time points
t0 <- 0 # initial time
tf <- 600 # final time
times <- seq(from=t0, to=tf, by=5)

# Parameters

k <- 10
c <- 0.5
```

```
lambda = 1
delta = 1
gamma = gamma_m = 0.1

alpha <- 1.44 # different from the ESS
beta <- Trade_off(alpha, k, c)

alpha_m <- alpha_ESS # ESS
beta_m <- Trade_off(alpha_m, k, c)

parms <- c("lambda"=lambda, "delta"=delta, "beta"=beta, "alpha"=alpha, "gamma"=gamma,
           "beta_m"=beta_m, "alpha_m"=alpha_m, "gamma_m"=gamma_m)

# Initialization at endemic equilibrium

S_e <- (delta+alpha+gamma)/beta
I_e <- lambda/(delta+alpha+gamma) - delta/beta
R_e <- (gamma/delta)*I_e

I_m_t0 <- 0.001 # I_m(t0), (very low) initial density of individuals infected by the variant

init_endemic <- c("S" = S_e, "I" = I_e, "I_m (ESS)" = I_m_t0, "R" = R_e)

# Simulation (long term)
simul_long_term <- lsoda(y = init_endemic, times = times, func = ODE_SIR.2, parms = parms)

plot_simul.2(simul_long_term, title = "Fig. 9. Simulation of the SIR model (2)-(3) in the long term\n",
             parms = round(parms, 2))
```

Fig. 9. Simulation of the SIR model (2)−(3) in the long term



Parameters: lambda = 1 ; delta = 1 ; beta = 12 ; alpha = 1.44 ; gamma = 0.1 ; beta_m = 10.49 ; alpha_m = 1.1 ; gamma_m = 0.1

We explore a scenario where the ancestral strain has reached the endemic equilibrium, the strain with the ESS strategy is introduced at very low density. The latter gradually replaces the previously dominant strain (it becomes dominant around $t = 285$). In this case, the replacement is quite slow. We can verify that the ESS always invades when we use other ancestral strains. The speed of the invasion varies with the ancestral strains.

2.3.2 ... but the ESS can be outcompeted by other virulence strategies during transient epidemics

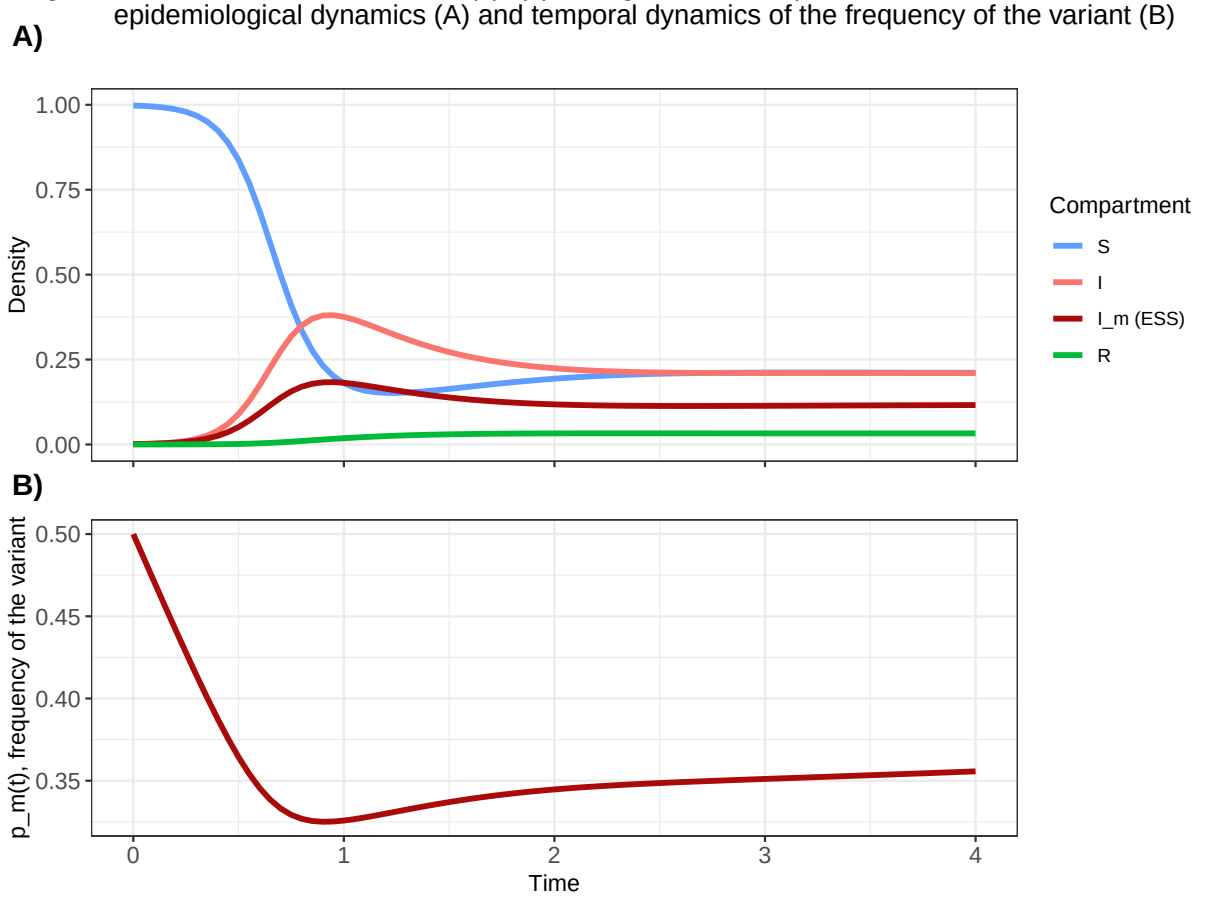
Q8. Starting from the disease-free equilibrium, introduce two pathogen (one at the ESS) in small but equal densities, both following the same trade-off function. Is the ESS pathogen always more selected than the other pathogen? For the second pathogen, try with $\beta < \beta_m$ and $\beta > \beta_m$. What do you notice? Suggest an explanation.

```
# Time points
t0 <- 0 # initial time
tf <- 4 # final time
times <- seq(from=t0, to=tf, by=0.05)

# Initialization of each compartment
I_t0 <- I_m_t0 <- 0.001 # Initial density of infected individuals (resident and mutant strains)
I_T_t0 <- I_t0 + I_m_t0 # Total density of infected individuals

init_transient <- c("S" = 1-I_T_t0, "I" = I_t0, "I_m (ESS)" = I_m_t0, "R" = 0)

# Simulation (transient epidemic)
simul_transient <- lsoda(y = init_transient, times = times, func = ODE_SIR.2, parms = parms)
```

```
# Plot
Fig_transient <- plot_simul.2(simul_transient)

simul_transient <- as.data.frame(simul_transient)
simul_transient$p_m <- simul_transient[,4] / (simul_transient[,4]+simul_transient[,3])

plot_grid(

  ggdraw() + draw_label(
    "Fig. 10. Simulation of the SIR model (2)-(3) during a transient epidemic:
            epidemiological dynamics (A) and temporal dynamics of the frequency of the variant (B)\n",
    x = 0.025, hjust = 0, size = 13),

  Fig_transient + theme(axis.text.x = element_blank(), axis.title.x = element_blank(),
                        legend.position = 'none'),

  ggplot(simul_transient, aes(x = time, y = p_m)) +
    geom_line(cex = 1.3, col = "#A90B0B") +
    labs(caption = paste("\n Parameters:",
                         paste(names(parms), round(parms, 2), sep = ' = ', collapse = ' ; ')),
         x = "Time", y = "p_m(t), frequency of the variant") +
    theme_bw() +
    theme(axis.text.x = element_text(size=11),
          axis.text.y = element_text(size=11),
          plot.caption = element_text(hjust = 0, face = 'bold')),
```

```
ggplot() + theme_void(),

get_legend(Fig_transient),

ncol = 2, rel_heights = c(0.2, 1, 1), rel_widths = c(0.85, 0.15), byrow = FALSE,
labels = c("", "", "A)", "", "B)", ""), label_y = c(0, 0, 1.1, 0, 1.1, 0))
```

Fig. 10. Simulation of the SIR model (2)−(3) during a transient epidemic:
epidemiological dynamics (A) and temporal dynamics of the frequency of the variant (B)



Parameters: lambda = 1 ; delta = 1 ; beta = 12 ; alpha = 1.44 ; gamma = 0.1 ; beta_m = 10.49 ; alpha_m = 1.1 ; gamma_m = 0.1

In this simulation example (same parameters as above (**Q7**)), starting with small densities for both strains (ratio 1:1), the one at the ESS (here, the variant) is always dominated by the other strain (here, the resident strain) (*cf.* **Fig. 10-A**). At the beginning of the epidemic, the frequency of the variant drops from 0.5 to 0.33. This shows that, even if a strain has the best strategy in the long term (ESS), it may be transiently outcompeted (when the host population is not at the endemic equilibrium) by another strain. If we continue the simulation, however, the ESS strain will eventually invade. We already see for example in **Fig. 10-B** that (albeit weakly) the frequency of the variant rises from $t = 0.9$.

As in **Q4-5**, the strain favoured in the short term is the most transmissible (and the most virulent according to our trade-off function) because the available host density $S(t)$ is not limiting (beginning of the epidemic), while in the longer term (when $S(t)$ is much lower) the transmission advantage no longer compensates for the burden of a higher virulence (more details in **§2.1.3. Population genetics approach - derivation of the selection coefficient**).

Theoretical results from the adaptive dynamics approach assume that evolutionary processes are much slower than epidemiological dynamics and, therefore, that the system has always reached an equilibrium when a new variant emerges. Evolutionary epidemiology does not rely on this assumption and allows us to understand what factors affect the change in frequency of the mutant strain (*e.g.* the availability of susceptible hosts $S(t)$) as discussed above in **Q4-5**.

3 References

Day T. & Gandon S. (2006) Insights from Price's equation into evolutionary epidemiology. In: *Disease evolution: models, concepts and data analyses. (Feng, Z. Dieckmann U.; Levin, S., eds.) American Mathematical Society*, p. 23-43.

Day T. & Gandon S. (2007) Applying population-genetic models in theoretical evolutionary epidemiology. In: *Ecology Letters* (**10**): 876-888.

Dynamique de l'adaptation virale: approches théoriques et expérimentales

Introduction

Dans le cadre de cette thèse, nous avons mené plusieurs projets sous le thème général de la dynamique de l'adaptation virale. Les virus peuvent muter et s'adapter en très peu de temps, ce qui a un impact direct, par exemple, sur la manière dont les maladies virales comme le VIH sont traitées ou dont les stratégies de vaccination sont déployées. La pandémie de SARS-CoV-2 nous a brutalement rappelé la menace que représentent les agents pathogènes émergents ou réémergents. Dans ce contexte, il est crucial de comprendre les mécanismes d'adaptation virale afin de concevoir des interventions prophylactiques, thérapeutiques ou non pharmaceutiques efficaces qui limiteraient les conséquences indésirables de l'évolution virale.

Dans cette thèse, nous utilisons à la fois des approches théoriques et expérimentales. Par la théorie, nous essayons de démêler les effets des différents mécanismes évolutifs et de fournir des prédictions qualitatives et quantitatives pour le résultat de l'adaptation virale. Nous utilisons également une approche expérimentale pour valider certaines de ces prédictions et pour essayer de découvrir de nouveaux processus biologiques.

Dans cette thèse nous avons étudié plusieurs facettes de l'adaptation virale, liées

à différents scénarios épidemiologiques. Dans une première partie, nous avons étudié grâce à des modèles analytiques l'évolution de traits d'histoire de vie des pathogènes comme le taux de transmission, et la possibilité d'utiliser la mutagénèse létale comme approche thérapeutique. Ensuite, nous avons considéré l'adaptation virale quand la population d'hôtes contient des hôtes résistants à l'infection. Nous avons modélisé la probabilité d'émergence évolutive: que des mutations d'échappement à la résistance se produisent après l'introduction de virus et mènent à une épidémie. Nous confirmons ces résultats expérimentalement. Enfin, nous avons étudié la dynamique des fréquences de ces mutations d'échappement dans le cas où une épidémie se produit, et plus particulièrement dans des scénarios de coévolution avec les hôtes.

Chapitre 1: Evolution des traits d'histoire de vie

Evolution du taux de transmission et mutagénèse létale

Dans le premier chapitre, nous avons construit un modèle qui associe la dynamique épidémiologique à l'intérieur de l'hôte à la dynamique évolutive de l'agent pathogène. Nous avons adapté le modèle géométrique de Fisher (Martin and Gandon, 2010; Tenaillon, 2014) pour relier les phénotypes de l'agent pathogène au taux de transmission. Nous sommes en mesure de dériver la dynamique évolutive de la distribution des phénotypes et du taux de transmission moyen dans un régime de sélection faible et de mutation forte (WSSM), dont nous montrons qu'il peut être assoupli pour permettre des effets mutationnels plus importants. Nous utilisons ce cadre de modélisation pour étudier la possibilité d'une mutagénèse létale : conduire des populations virales à l'extinction par une augmentation du taux de mutation (J. J. Bull, Sanjuan, and Wilke, 2007; Lynch, Bürger, et al., 1993). Cela augmente l'afflux de mutations létales, que nous interprétons comme un terme de mortalité supplémentaire, et de mutations non létales qui ont en moyenne un effet négatif sur le taux de transmission et provoquent donc une charge mutationnelle : un coût sur la fitness dûe à la variance du taux de transmission dans la population pathogène. Dans ce travail, nous montrons comment les mutations peuvent conduire à l'extinction des pathogènes, mais aussi à des pathogènes mieux adaptés. En effet, contrairement à de nombreuses études sur la mutagenèse létale, nous considérons également les mutations bénéfiques qui augmentent le taux de transmission. Nous intégrons également la rétroaction démographique de la dynamique épidémiologique. Nous montrons que, tout comme la dynamique de la densité des cellules infectées, la vitesse d'adaptation est aussi conditionnée par la densité des cellules sensibles.

Cette rétroaction pourrait avoir un effet important dans les scénarios de mutagénèse létale. Lorsque la population de cellules infectées diminue, la population de cellules sensibles augmente, ce qui pourrait aider la population pathogène à échapper à l'extinction.

Evolution jointe du taux de transmission et de la virulence

Dans le premier chapitre, nous avons examiné l'évolution d'un seul caractère quantitatif, le taux de transmission, qui est habituellement lié à l'évolution d'un autre caractère la virulence. Le concept de stratégie évolutivement stable a été utilisé pour décrire l'état d'équilibre du taux de transmission et de la virulence. Pour la période transitoire, la dynamique des traits moyens peut être calculée sur la base de la matrice **G** de variance-covariance de ces traits avec une équation de Price (Day and Gandon, 2006; Day, Parsons, et al., 2020). Cependant, cette solution est incomplète car la dynamique de l'évolution des variances et des covariances n'est pas explicitement modélisée.

Pour combler cette lacune, nous avons développé une approche par équations aux dérivées partielles basée sur des travaux antérieurs (Martin and Roques, 2016). Nous avons étendu cette approche pour modéliser simultanément les distributions du taux de transmission et de la virulence, en incorporant des optima distincts pour chacun de ces traits d'histoire de vie des pathogènes. Cela a permis d'introduire un compromis entre les deux traits, car les phénotypes ne pouvaient pas être simultanément proches des deux optima. Grâce à nos dérivations, nous avons obtenu une équation de Price et des expressions pour la dynamique de la variance et de la covariance. Ceci nous a permis de comprendre la dynamique d'adaptation conjointe du taux de transmission et de la virulence.

En outre, notre cadre de modélisation intègre les dynamiques évolutives et épidémiologiques, en reconnaissant le rôle crucial de la densité des hôtes sensibles dans l'orientation de l'adaptation vers l'optimisation du taux de transmission ou de la virulence. La pression sélective sur le taux de transmission est influencée par la densité d'hôtes, tandis que la pression sélective sur la virulence reste indépendante de ce facteur. La prise en compte de l'impact de la dynamique épidémiologique est donc particulièrement importante pour les agents pathogènes à évolution rapide, tels que les virus, pour lesquels les changements évolutifs et épidémiologiques se produisent à des échelles de temps similaires.

Chapitre 2: Evolution de l'échappement à la résistance des hôtes

Le système expérimentale *Streptococcus thermophilus* et phage 2972

Ce système modèle offre des possibilités uniques d'explorer les processus microévolutifs induits par la coévolution antagoniste entre les bactéries et leurs pathogènes viraux. Contre le phage lytique 2972, *S. thermophilus* utilise comme seule système de défense le système CRISPR-Cas. Ce dernier permet à la bactérie d'incorporer des séquences d'ADN d'environ 30 paires de bases (nommées *spacers*) à partir du génome du phage. Pendant l'infection, ces spacers sont utilisés comme guide par le complexe Cas pour reconnaître et cliver le génome du phage à la séquence correspondante, le *protospacer*, stoppant ainsi la réplication du virus. Les phages peuvent échapper à l'immunité CRISPR par des mutations dans les protospacers qui empêchent la reconnaissance par le complexe Cas. Le séquençage du locus CRISPR des populations de bactéries et du génome entier des bactériophages permettent de caractériser pleinement la spécificité du réseau d'infection, sans aucun test phénotypique.

Quelle stratégie deploiement de la résistance chez les hôtes pour limiter l'émergence d'un pathogène

Dans notre deuxième chapitre, nous nous sommes concentrés sur l'étude de la probabilité d'émergence de l'agent pathogène en fonction de différentes stratégies de déploiement de la résistance dans la population hôte. Alors que des études antérieures ont exploré l'impact de la diversité des hôtes résistants sur la limitation de la propagation des pathogènes, nous nous sommes concentrés sur la profondeur de la résistance, en particulier sur le nombre de mutations d'échappement nécessaires à la réussite de l'infection. Nous avons étudié trois stratégies de déploiement de la résistance : Stratégie de mélange : la moitié des hôtes résistants sont porteurs de la résistance A, tandis que l'autre moitié était porteuse de la résistance B, ce qui nécessite pour le phage une mutation d'échappement différente pour échapper à la résistance de chacun des hôtes. Stratégie de pyramidage : tous les hôtes sont doublement résistants (AB), ce qui signifie que les agents pathogènes doivent acquérir

deux mutations d'échappement distinctes pour réussir l'infection. Stratégie de combinaison : la moitié des hôtes résistants présentent une résistance simple (A ou B) et l'autre moitié une résistance double (AB). Nos résultats ont révélé que des inoculums plus importants augmentaient la probabilité d'émergence de deux manières. Premièrement, ils augmentent le nombre de mutants préexistants dans la population. Deuxièmement, ils provoquent une épidémie initiale plus importante parmi les hôtes sensibles, ce qui permet une plus grande réplication et génère potentiellement plus de mutants.

D'un point de vue analytique, nous avons prédit que la stratégie de pyramide était la plus efficace pour empêcher l'émergence du pathogène, car l'acquisition simultanée des deux mutations d'échappement représentait un défi important. En revanche, la stratégie de mélange était plus propice à l'émergence de pathogènes en raison des avantages immédiats en termes de fitness que procurent les mutations d'échappement uniques, qui peuvent se propager rapidement dans la population. La stratégie de combinaison a démontré une efficacité intermédiaire, car l'une des deux mutations d'échappement (contre des hôtes monorésistants) offrait des avantages considérables en termes d'aptitude et servait de tremplin pour l'acquisition de l'autre mutation d'échappement afin d'infecter des hôtes doublement résistants.

Nous avons testé ces prédictions avec le système expérimental de *Streptococcus thermophilus* résistant à CRISPR et son phage virulent 2972. Nous avons confirmé la prédiction sur l'effet de la taille de l'inoculum et de la hiérarchie des traitements selon laquelle la probabilité d'émergence était plus élevée dans le traitement de mélange, intermédiaire pour l'un des traitements de combinaison et la plus faible dans le traitement de pyramide. Cependant, nous avons constaté que l'autre traitements de combinaison n'était pas significativement différent du traitement de mélange en termes de probabilité d'émergence évolutive. Cela pourrait être dû à une différence de taux de mutation entre les deux protospacers concernés.

Dynamique de la fréquence des mutations d'échappement

Dans l'annexe B, nous décrivons une expérience d'évolution avec le même système de bactéries et de phages. Dans ce travail, nous avons voulu suivre la dynamique évolutive des mutations d'échappement après l'émergence initiale. Nous avons testé l'effet sur la dynamique évolutive du coefficient de sélection associé aux mutations d'échappement, que nous avons manipulé par des différences dans les fréquences ini-

tiales des différents hôtes. Nous avons constaté qu'une fréquence plus élevée d'hôtes était associée à une fréquence accrue des mutations d'échappement correspondantes au début de l'expérience, mais que cet effet disparaissait les jours suivants. Nous avons également testé l'effet du taux de mutation d'échappement en comparant la fréquence des mutations d'échappement dans deux groupes de protospacers qui diffèrent par leur taux de mutation. Contrairement à nos attentes, nous n'avons pas trouvé d'effet significatif du taux de mutation au début de l'expérience, mais nous avons trouvé un effet dans les jours suivants. En effet, nous nous attendions à ce que le taux de mutation soit limitant au début de l'expérience, mais qu'une fois les mutants apparus dans la population, il n'ait plus d'impact. Nous avons observé que les phages porteurs de multiples mutations d'échappement apparaissaient au début de l'expérience, mais qu'à la fin de l'expérience, la fréquence des mutations d'échappement continuait d'augmenter et que nous n'avions pas atteint un point où tous les phages pouvaient infecter tous les hôtes. Nous pouvons donc comprendre que le taux de mutation peut jouer un rôle plus tard dans l'expérience, car il peut accélérer l'acquisition de mutations d'échappement supplémentaires dans les phages qui infectent déjà d'autres hôtes résistants.

Il est intéressant de noter que nous obtenons des résultats contrastés par rapport au chapitre 2, où nous prédisions que dans le traitement Mélange, la probabilité d'émergence est minimisée lorsque $f_A = 1/2$, c'est-à-dire lorsque les deux hôtes résistants sont présents à des fréquences égales. Dans l'annexe B, nous étudions la dynamique des fréquences d'échappement après l'émergence. Dans ce cas, nous constatons que la fréquence globale des mutations d'échappement est plus élevée lorsque les hôtes résistants sont en fréquences égales (traitement B). Ainsi, l'hétérogénéité de la fréquence des hôtes résistants pourrait favoriser l'émergence du pathogène, mais limiter la propagation ultérieure des mutants d'échappement une fois l'émergence réalisée.

La somme des différentes fréquences des mutations d'échappement étant rapidement supérieure à un, nous savons que les phages qui échappent à la résistance de plusieurs hôtes émergent. Cependant, en utilisant une technologie de séquençage à lecture courte, nous ne récupérons pas les informations de liaison entre ces mutations d'échappement. Nous ne connaissons donc pas la composition de la population de phages en termes de génotypes d'échappement et nous ne pouvons pas suivre avec précision la dynamique de ces phages généralistes. Pour résoudre ce problème, nous aimerions utiliser une technologie de séquençage à lecture longue pour suivre la

dynamique évolutive dans le même système de bactéries et de phages résistants à CRISPR.

Chapitre 3: Coévolution des virus avec leurs hôtes

Compétition entre les hôtes et conséquences sur l'évolution de la population virale

Dans le troisième chapitre nous avons étudié une expérience similaire, mais cette fois-ci en permettant la coévolution des bactériophages et des hôtes bactériens. Nous avons surveillé à la fois l'évolution des mutations d'échappement des phages et le locus CRISPR des bactéries. Dans tous les traitements, nous avons utilisé un mélange de 16 souches bactériennes résistantes qui différaient par leur locus CRISPR, et une souche de type sauvage totalement sensible. Nous avons observé dans un traitement sans phages que la diversité bactérienne était rapidement perdue en raison des différences de fitness intrinsèques entre les différentes souches. En particulier, nous avons trouvé deux souches qui étaient plus compétitives et qui ont complètement envahi la population au cours des 4 jours de l'expérience dans toutes les répétitions. Avec une telle dynamique répétée de l'hôte en l'absence de phages, nous avons utilisé notre expérience pour étudier les effets réciproques de la compétition de l'hôte sur l'adaptation du pathogène et vice versa. Nous avons constaté qu'à travers la sélection négative dépendante de la fréquence, les phages ont limité la perte de diversité de l'hôte. Cependant, cette NFDS n'a pas conduit à un scénario de "Kill-the-winner" (Thingstad, 2000; Weinbauer, 2004). Ce que nous avons constaté, c'est que les souches les plus adaptées identifiées dans le contrôle ont d'abord dépassé les autres, puis la fréquence de la mutation d'échappement correspondante dans la population de phages a augmenté. Cependant, avant de s'éteindre, ces souches initialement plus compétitives ont acquis à plusieurs reprises de nouveaux espaceurs de résistance et ont vu leur fréquence augmenter à nouveau. Ce que nous avons observé concernant la diversité des hôtes est donc dû au fait que les phages génèrent de la diversité au niveau du locus CRISPR des souches déjà dominantes, plutôt qu'à une conservation de la diversité initiale.

Notre système nous permet également de suivre la dynamique de l'aptitude moyenne de la population de phages, en utilisant à la fois la fréquence des génotypes de résistance de l'hôte et la fréquence des mutations d'échappement. Nous calcu-

lons également l'aptitude de la population de phages lorsqu'elle est confrontée à des hôtes contemporains, ou à des hôtes de périodes passées et futures, imitant ainsi les expériences de décalage temporel. Nous constatons que les phages sont les plus aptes face à des hôtes du passé proche, mais que cette aptitude diminue rapidement face à des hôtes du futur. Cela met en évidence l'adaptation des phages pour échapper aux espaceurs présents dans la population hôte, ainsi que l'adaptation des hôtes qui acquièrent rapidement de nouveaux espaceurs pour résister aux phages. Nous trouvons également des preuves de l'adaptation locale des bactériophages en comparant l'aptitude des populations de phages contre des hôtes du même réplicat (sympatrie) ou contre des hôtes d'autres réplicats (allopatrie).

Conclusion

Dans cette thèse, nous avons étudié différents aspects de la dynamique de l'adaptation virale, avec une variété d'approches, à la fois théoriques et expérimentales. Si les différents chapitres semblent très distincts, comme le modèle théorique sur l'évolution du taux de transmission et l'expérience de coévolution avec des bactéries résistantes à CRISPR, dans ces projets, l'évolution des virus est guidée par les mêmes forces que nous pouvons résumer par l'équation suivante :

$$\Delta r = \Delta r_{ns} + \Delta r_m + \Delta r_{ec} \tag{28}$$

La dynamique de la fitness malthusienne, ou taux de croissance, est déterminée par la sélection naturelle (Δr_{ns}), par l'effet direct de la mutation (Δr_m) et par les changements dans l'environnement (Δr_{ec}).

Nous avons vu comment la sélection naturelle peut conduire l'évolution des caractères quantitatifs vers un optimum, ou sélectionner des caractères qualitatifs tels que la résistance à l'échappement. La sélection naturelle opère sur la variance dans la population, qui est générée par les mutations. Ces mutations peuvent permettre d'échapper à la résistance de l'hôte ou d'augmenter le taux de transmission. Mais elles peuvent également conduire à l'extinction en raison de leur caractère en moyenne délétère. Enfin, nous avons étudié la manière dont les changements environnementaux biotiques peuvent influer sur l'évolution virale : par le biais de la densité des cellules sensibles, des changements dans la fréquence des différents hôtes et même de l'apparition de nouveaux hôtes résistants dans des scénarios coévolutifs.

Nous avons mis en évidence, à l'aide d'approches théoriques et expérimentales, l'interaction entre l'épidémiologie et la dynamique évolutive, en soulignant comment ces processus peuvent se produire aux mêmes échelles de temps, ce dont nous avons également été témoins lors de la pandémie de SARS-CoV-2. Cette thèse montre que tous ces processus doivent être pris en compte conjointement pour mieux comprendre l'évolution virale et éventuellement concevoir de meilleures approches thérapeutiques ou politiques de gestion épidémiologiques.

Dynamique évolutive des traits d'histoire de vie et applications à la gestion des épidémies

Dans cette thèse, nous avons modélisé la dynamique évolutive du taux de transmission (et de la virulence) dans un cadre qui permet une rétroaction entre l'évolution et la dynamique épidémiologique. Cette rétroaction a été largement ignorée dans le cadre plus classique de la maximisation de R_0. Dans le cadre de la dynamique adaptative, les rétroactions écologiques sont prises en compte, mais on a supposé que l'échelle de temps des changements évolutifs était beaucoup plus élevée que celle des changements écologiques, et ces derniers ont donc été considérés comme immédiats. En outre, dans ce cadre, les mutations sont considérées comme des événements rares et ne sont donc pas modélisées explicitement, et la sélection est alimentée par une variance permanente.

La pandémie de SARS-CoV-2 nous a montré que cette séparation des échelles de temps n'est pas toujours justifiée. Nous avons observé des changements évolutifs avec l'apparition et l'augmentation subséquente de la fréquence de plusieurs nouveaux variants au cours des premières phases de la pandémie, donc avant qu'un (quasi-)équilibre épidémiologique n'ait pu être atteint puisque, par exemple, la proportion de personnes immunisées était encore relativement faible. Cela souligne le fait que les changements épidémiologiques et évolutifs doivent être étudiés ensemble, au moins pour les épidémies de l'ampleur de la pandémie de SARS-CoV-2. Un autre aspect qui fait défaut dans la compréhension de l'évolution de ce virus est l'absence de prédiction de l'évolution du taux de transmission ou de la virulence. En l'absence d'un modèle mutationnel et d'un paysage phénotypique spécifiques, il s'est avéré difficile de prédire si les variantes initiales seraient associées à un taux de transmission et/ou à une virulence plus ou moins élevés. Notre approche utilisant le modèle de Fisher nous permet de présenter de telles prédictions, ce qui pourrait expliquer pourquoi initialement un virus mal adapté pourrait évoluer pour à la fois augmenter le taux de transmission et réduire la virulence, avant qu'un compromis

ne soit atteint qui limite l'optimisation de ces deux traits simultanément.

CRISPR: un modèle expérimentale pour l'épidemiologie

Dans cette thèse, nous avons utilisé le système expérimental de *Streptococcus thermophilus* et son phage virulent 2972. Avec ce système, nous avons exploré la dynamique coévolutive de la résistance et de l'échappement à CRISPR. Ce système est intéressant en soi car cette bactérie est massivement utilisée dans l'industrie de la fermentation laitière (Samson and Moineau, 2013). Avec la possibilité de plus en plus envisagée de la thérapie par les phages - qui consiste à traiter les bactéries pathogènes (multirésistantes aux médicaments) avec un cocktail sélectionné de phages - il devient de plus en plus important d'étudier CRISPR et, plus généralement, la coévolution bactéries-bactériophages.

Nous pensons également que ce système expérimental est un modèle approprié pour l'étude de l'évolution de l'échappement des pathogènes à la résistance des hôtes en général. Les bactéries peuvent être résistantes avec un ou potentiellement plusieurs espaceurs CRISPR et nous connaissons exactement le déterminisme génétique des mutations d'échappement dans les phages. Nous avons montré comment cela pouvait être utilisé pour explorer l'efficacité de certaines stratégies de déploiement de la résistance dans la population hôte pour limiter l'émergence du pathogène. Nous pensons que ces conclusions pourraient s'appliquer à une variété de systèmes pour lesquels il est difficile d'obtenir des données expérimentales, comme les systèmes agronomiques ou encore une population humaine vaccinée. Nous montrons également qu'après l'émergence d'un pathogène, nous pouvons suivre la dynamique des pathogènes échappés grâce au séquençage tout en contrôlant, dans une certaine mesure, la composition de la population hôte. Une telle approche pourrait être utilisée pour étudier les stratégies de déploiement dynamique de la résistance, comme le déploiement progressif de plusieurs vaccins potentiels, à l'image du scénario de la pandémie de SARS-CoV-2.

Abstract

Key words: Evolutionary epidemiology, Host-pathogen interactions, Virus, Pathogen adaptation, CRISPR, Bacteriophages

Most living organisms on the tree of life can be infected by viruses. The ubiquity of viruses is driven by different factors including high mutation rates, high population sizes and low generation times, which allow for quick adaptation to very different host species. The dynamics of adaptation - the rate of change of the mean fitness of the viral population - results from the interplay between multiple evolutionary forces that may promote or hamper viral adaptation. During this PhD we developed a combination of theoretical and experimental approaches to disentangle the influence of some of these factors on viral adaptation.

First, we explored the dynamics of viral adaptation to a homogeneous host population. We used Fisher's Geometric Model of adaptation and studied the joint evolutionary and epidemiological dynamics of a viral population spreading in a host population. This modeled allowed us to explore the lethal mutagenesis hypothesis: is it possible to treat viral infections with mutagenic drugs to increase the mutation load of the viral population beyond a threshold that may result in the extinction of the within-host population? We show which parameters affect the critical mutation rate leading to viral extinction and we show how epidemiology and evolution can affect the transient within-host dynamics of the viral population when a single virus life-history trait (transmission rate) is under selection. We extend this modeling framework to study the joint evolution of transmission and virulence during the adaptation of an emerging pathogen.

Second, we studied viral adaptation in heterogeneous host populations when the virus spreads among a diversified population of resistance host. We studied the evolutionary emergence of viruses: can viruses avoid extinction by the acquisition of escape mutations allowing them to infect some of the resistant hosts in the population? We developed a birth-death process model to predict the probability of evolutionary emergence as a function of the composition of the host population. In particular, we show how the proportion of multiple resistant hosts can reduce the risk of pathogen evolutionary emergence. We put some of these predictions to the test using bacteriophages spreading in bacterial populations. We manipulate the diversity of CRISPR immunity in *Streptococcus thermophilus* bacteria and we confirm the key influence of multiple resistance on the risk of viral adaptation.

Third, we also studied viral adaptation in time-varying environments where the host population is allowed to coevolve with the virus. In this experimental project we monitored the adaptation of bacteriophages as they coevolved with the CRISPR immunity of *S. thermophilus bacteria*. We track reciprocal adaptive changes in which bacteria acquire new layers of resistance (new spacers in the CRISPR array) and phages acquire new escape mutations in the corresponding protospacers. This experiment allows us to monitor the dynamics of viral adaptation across time and space. Interestingly, we find a significant asymmetries in competitive abilities among different bacterial strain in the absence of phage predation. This asymmetric competition has dramatic consequences on the maintenance of diversity of host resistance and on the coevolutionary dynamics with the virus. This thesis demonstrates the possibility to use experimental evolution with microbial microcosms to explore the validity of some theoretical predictions on the dynamics of viral adaptation. This experimental validation is particularly important if one wants to use evolutionary models to make public-health recommendations.

Résumé

Mots clés: Epidémiologie évolutive, Interactions hôte-pathogènes, Virus, Adaptation des pathogènes, CRISPR, Bactériophages

La plupart des organismes vivants peuvent être infectés par des virus. Cette omniprésence est due à différents facteurs, notamment des taux de mutation élevés, des populations de grande taille et des temps de génération courts, qui permettent une adaptation rapide à des espèces hôtes très différentes. La dynamique de l'adaptation des populations virales résulte de l'interaction entre de multiples forces évolutives. Au cours de cette thèse, nous avons développé une combinaison d'approches théoriques et expérimentales pour démêler l'influence de certains de ces facteurs sur l'adaptation virale.

Tout d'abord, nous avons exploré la dynamique de l'adaptation virale face à une population hôte homogène. Nous avons utilisé le modèle géométrique d'adaptation de Fisher et étudié les dynamiques évolutive et épidémiologique d'une population virale en modèle intra-hôte. Ce modèle permet d'explorer l'hypothèse de la mutagenèse létale: est-il possible de traiter les infections virales avec des médicaments mutagènes pour augmenter la charge de mutation au-delà d'un seuil qui peut entraîner l'extinction de la population? Nous montrons quels paramètres affectent le taux de mutation critique conduisant à l'extinction virale et nous montrons comment l'épidémiologie et l'évolution peuvent affecter la dynamique transitoire de la population virale à l'intérieur de l'hôte lorsqu'un seul trait du cycle de vie du virus (taux de transmission) est soumis à la sélection. Nous étendons ce cadre de modélisation à l'étude de l'évolution conjointe de la transmission et de la virulence au cours de l'adaptation d'un pathogène émergent.

Deuxièmement, nous avons étudié l'adaptation virale dans des populations d'hôtes hétérogènes lorsque le virus se propage parmi une population diversifiée d'hôtes résistants. Nous avons étudié l'émergence évolutive des virus : les virus peuvent-ils éviter l'extinction par l'acquisition de mutations d'échappement leur permettant d'infecter certains des hôtes résistants de la population? Nous avons développé un modèle de naissance/mort pour prédire la probabilité d'émergence évolutive en fonction de la composition de la population d'hôtes. En particulier, nous montrons comment la proportion d'hôtes multi-résistants peut réduire le risque d'émergence évolutive de l'agent pathogène. Nous mettons certaines de ces prédictions à l'épreuve en utilisant des bactériophages se propageant dans des populations bactériennes. Nous manipulons la diversité de l'immunité CRISPR dans les bactéries *Streptococcus thermophilus* et nous confirmons l'influence clé de la résistance multiple sur le risque d'adaptation virale.

Troisièmement, nous avons également étudié l'adaptation virale dans des environnements variables dans le temps où la population hôte est autorisée à coévoluer avec le virus. Dans ce projet expérimental, nous avons suivi l'adaptation des bactériophages au fur et à mesure qu'ils évoluaient avec l'immunité CRISPR des bactéries *S. thermophilus*. Nous suivons les changements adaptatifs réciproques dans lesquels les bactéries acquièrent de nouvelles couches de résistance (nouveaux spacers dans le locus CRISPR) et les phages acquièrent de nouvelles mutations d'échappement dans les protospacers correspondants. Cette expérience nous permet de suivre la dynamique de l'adaptation virale dans le temps et l'espace. Nous avons noté des asymétries significatives dans les capacités de compétition entre les différentes souches bactériennes. Cette compétition asymétrique a des conséquences dramatiques sur le maintien de la diversité de la résistance de l'hôte et sur la dynamique coévolutive avec le virus. Cette thèse démontre la possibilité d'utiliser l'évolution expérimentale en microcosmes microbiens pour explorer la validité de certaines prédictions théoriques sur la dynamique de l'adaptation virale. Cette validation expérimentale est particulièrement importante si l'on veut utiliser des modèles évolutifs pour faire des recommandations de santé publique.