Evolutionary epidemiology of infectious diseases: Theoretical and statistical approaches

Wakinyan Benhamou

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

En Biostatistique

École doctorale I2S – Information, Structures et Systèmes

Unité de recherche :
Centre d'Écologie Fonctionnelle et Évolutive (CEFE) – UMR 5175

Evolutionary epidemiology of infectious diseases: Theoretical and statistical approaches

Présentée par Wakinyan BENHAMOU
Le 26 septembre 2024

Sous la direction de Rémi CHOQUET
et Sylvain GANDON

Devant le jury composé de

C. Jessica E. METCALF, Professor, Princeton University	**Rapportrice**
Samuel ALIZON, Directeur de recherche, CNRS, CIRB	**Rapporteur**
Benoîte DE SAPORTA, Professeure, Université de Montpellier, IMAG	**Présidente du jury**
Lulla OPATOWSKI, Professeure, UVSQ, Institut Pasteur	**Examinatrice**
Rémi CHOQUET, Ingénieur de recherche, CNRS, CEFE	**Co-directeur de thèse**
Sylvain GANDON, Directeur de recherche, CNRS, CEFE	**Co-directeur de thèse**

UNIVERSITÉ DE MONTPELLIER

This thesis compiles my research work as a PhD student at the *Centre d'Écologie Fonctionnelle et Évolutive* (CEFE) in Montpellier, France, from September 2021 to August 2024 (three years) under the supervision of Rémi Choquet and Sylvain Gandon. The first version of the present manuscript was submitted on July 18, 2024. My PhD scholarship comes from the French Ministry of Higher Education and Research, and was granted by the École Normale Supérieure Paris-Saclay. In this thesis, I have conducted several projects in evolutionary epidemiology of infectious diseases, at the interface between theoretical models and empirical/experimental data.

Remerciements

Je tiens tout d'abord à remercier **Jessica Metcalf** et **Samuel Alizon** qui ont accepté d'être rapportrice/rapporteur pour cette thèse. Je remercie également **Benoîte de Saporta** et **Lulla Opatowski** pour avoir accepté de faire partie de mon jury et d'évaluer mon travail.

Un grand merci évidemment à **Rémi** et **Sylvain**, mes directeurs de thèse. Lorsque je vous avais contacté la première fois, j'étais intéressé pour travailler sur des sujets à l'interface entre modélisation mathématique et analyses de données, mais je ne connaissais pas grand-chose ni en épidémiologie (évolutive), ni en inférence statistique... Merci d'avoir accepté de m'accueillir dans l'équipe au CEFE, d'abord en stage de M2, puis pour poursuivre en thèse. Merci beaucoup à tous les deux pour les échanges scientifiques, votre écoute, vos relectures, vos précieux conseils, vos encouragements, votre soutien et pour vous être rendu disponibles chaque fois que j'en avais manifesté le besoin, même lorsque vous étiez déjà débordés.

Je remercie également **Mircea Sofonea**, **Luis-Miguel Chevin**, **Frédéric Mortier** et **Joseph Salmon** qui ont participé au suivi de ma thèse. Merci pour vos retours, vos suggestions et vos encouragements.

Je remercie **Jessica Metcalf** (une fois encore !) et **Bryan Grenfell** pour m'avoir invité à Princeton en avril dernier et m'avoir proposé de les rejoindre en post-doc pour la suite.

Merci **Seb**, pour ton aide sur le premier projet de ma thèse et, plus généralement, pour ta disponibilité chaque fois que j'avais une question à te poser.

Merci à **François Blanquart**, **Marc Choisy** et **Thomas Berngruber** pour votre collaboration sur le deuxième projet de ma thèse.

Évidemment, un grand merci à **Armelle**, **Julia**, **Mathilde**, **Tristan** et **Augustin** – la 8^e merveille du monde après le phare de Rodez – qui font vivre le bureau 204 et ont réussi à me supporter. À toutes ces discussions plus ou moins sérieuses, dans le bureau, à Montmaur ou au bar. Merci **Augustin** pour toujours s'assurer que je ne manque pas de café, de viennoiseries ou de bière, pour me rappeler de ne pas me disperser lorsque je dois avancer le Chapitre 4 (pas merci en revanche pour avoir essayé de me remplacer par un bananier). Merci aussi, **Armelle** et **Augustin**, pour vos relectures de ma thèse.

Pour les autres membres de l'équipe E^3, merci aussi beaucoup à **Jérémy** et **Thierry**.

Un grand merci aux anciens membres de l'équipe E^3. Pour les thésards : merci à **Martin**, **Valentin** et **Alicia**. En particulier, un grand merci à **Martin**, pour tous ces bons moments, pour la bonne ambiance qui a régné, pour toutes ces discussions sur la thèse, les cours à la fac, les stats, la cuisine ou la politique et aussi pour avoir relu mon manuscrit dans la dernière ligne droite. Merci aussi beaucoup pour avoir considérablement accéléré ma thèse en me montrant comment me servir du cluster de l'équipe. Merci aussi à **Amélie** et **Marina**, de passage en post-doc dans l'équipe. Merci à **Erwan (R1)**, de passage en stage. Je ne connaîtrais sans doute pas les maîtres-dauphins, les cristaux d'aquasirius ou les minéraux-dragons de la constellation des pléiades sans toi. Merci pour tes conseils vélo (est-ce qu'il y a un jour où tu ne parles pas vélo ?) et bon courage pour ton Master à Rennes.

Évidemment, un grand merci à toutes ces personnes exceptionnelles rencontrées au CEFE et que je n'aurais pas encore cité. Pour le couloir de l'aile B – 2^e étage : un immense merci à **Laurine**, **Lisa**, **Jérémy** (même si parfois il y a eu quiproquo), **Pablo**, **Téo** (je te parle et je sais que tu m'écoutes), **Nicolas** (même si tu seras docteur avant moi), **Erwan**, **Nico**, **Coralie (Coco)**, **Constance**, **Sofia**, **Mathis**, **Yseult**, **Lana**, **Noa**, **Flora**, **Louis**, ... et tous ceux que j'ai peut-être oubliés (vraiment désolé !). Merci énormément – déjà pour m'avoir supporté – pour avoir égayé l'ambiance du couloir (je cherche encore une cartouche de nerf au passage), pour tous ces moments hautement qualitatifs passés ensemble, au CEFE et ailleurs, les repas du midi, les discussions plus ou moins sérieuses, les parties de tarot (mention spéciale à la Nico, la Téo *sensu stricto*, la Téo *sensu* *lato*), d'échecs avec **Erwan** et **le J**, les soirées, les apéros, les week-end, les randonnées, les sorties à la plage, ... Tellement ravi d'avoir fait ma première transhumance chez toi **Laurine**, en Ardèche (et j'en profite pour saluer toute la **famille Mathieu**). J'ai passé une super troisème année grâce à vous, merci pour tout, du fond du cœur.

Je remercie évidemment la team BEV. Un grand merci à **Maurine**, **Laureline**, **Bastien**, **Jeanne (Jeannotte)** – je te rajoute ici **Clément** –, **Sacha** et **Marie-Émilie**, sans oublier **Letizia**, **Dimitri** et **Julia**. Ça a toujours été un immense plaisir d'être en votre compagnie. Merci pour toutes ces discussions plus moins sérieuses, et aussi pour les séances photos (c'est quoi le thème vendredi prochain ?).

Un grand merci à **Rémi** (ravi d'avoir terraformé Mars en buvant des coktails), **Léo**, **Victoria**, **Samson**, **Lilian** (Allez, on y est presque !), **Fanny (FanFan)**, **Jeanne**, **Étienne**, **Tristan**, **Nico (châtaigne)**, **Louis**, **Maëlis**, **Killian**, **Célian**, **Soumaya**, **Christie**, **Mellina**, **Lise**, **Nati**, **Tati**, **Catharina**, **Guillaume** (ravi d'avoir recroisé ta route après l'agreg), **Arnaud**, **Alycia**, **Victor**, **Ryan**, **Carole**, **Yohan**, **Val**, **Simon**, **Tanguy**, **Jan**, **Noa**, **Tom**, ... et tous ceux que j'ai peut-être oubliés (vraiment désolé encore !).

Merci beaucoup aussi à la team foot et à la team volley. Pour ceux que je n'aurais pas encore cité : merci à **Marilou**, **Léo**, **Aryan**, **Tomasín**, **Silvere**, **Teddy**, ...

Merci aux comités d'animation qui ont rythmé et enflammé la vie du labo en organisant les apéros et les apéros cross-over, dans les bars ou à Montmaur, les week-end d'intégr... de cohésion, les journées des non-permanents, les barbecues/pique-niques du TE, etc...

Un grand merci à **Lucas** et à **Florine**, hâte de refaire griller des tas de trucs sur une plancha.

À l'ISEM également, merci beaucoup **Ariane**, c'est toujours un plaisir de discuter avec toi ; merci aussi à **Adrien** (pour parler de la thèse, des confs, de boissons alcoolisées ou de plongée sous-marine) ; et merci aussi beaucoup à **Mathieu**.

Merci également à **Guillaume**, de l'IMAG. Hâte de rediscuter de la conjecture de Poincaré ou autre problème mathématique avec toi.

Un grand merci évidemment à **Romane**, **Chloé** et **Rébecca**. **Romane** et **Chloé**, voilà déjà plus d'une décennie qu'on se connaît, et **Rébecca**, ce fut un plaisir de te rencontrer. Merci **Romane** – et j'en profite pour faire coucou à **Elwë** et à **Dorian** –, pour toutes ces soirées à jouer ou à discuter autour d'un verre (de vin), au Spiritus ou ailleurs. Merci **Chloé** et **Rébecca**, notamment pour nous avoir accueilli chez vous en Bourgogne, terre de cassis (et de Puligny-Montrachet).

Un grand merci aux amis de prépa, **Cécile**, **Lou**, **Hugo** et **Valentin**, que j'ai toujours eu autant de plaisir à revoir durant ces trois années.

Merci à **Anaïs**, pour ces moments à Montpellier, à Paris ou pour cette virée en Normandie.

Je remercie également toutes les personnes de l'administration et de la cantine qui assurent chaque jour les très bonnes conditions de l'environnement de travail au CEFE.

Merci aussi à tous les collègues des équipes pédagogiques et tous les étudiants avec qui j'ai eu la chance d'interragir à la fac. J'ai pris un grand plaisir à donner toutes ces heures de cours.

Je remercie également les profs que j'ai eus jusqu'ici, et sans qui je ne serais probablement pas là.

De manière générale, je remercie toutes celles et ceux qui ont fait un bout de chemin avec moi. Merci à toutes les personnes qui ont été liées, de près ou de loin, à moi ou à ma thèse, et que j'aurais pu oublier.

Enfin, je remercie bien sûr toute ma **famille**.

Notamment, un grand merci à mes **parents** [1], qui m'ont toujours soutenu et qui m'ont toujours aidé lorsque j'en avais besoin (comme mes premiers problèmes de vélo et que je n'y connaissais rien). Merci pour tous ces bons moments, dans le Jura, à Montpellier ou en région stéphanoise. Et mention spéciale à toutes ces chouettes bouteilles de vin qu'on a bu: Meursault, Puligny-Montrachet, Chassagne-Montrachet, Chablis,

Chambolle-Musigny, Gevrey-Chambertin, Clos de Vougeot, Pomerol, Saint-Émilion, Margaux, Pauillac, Sauternes (mention spéciale au d'Yquem 97), Condrieu, Côte-Rôtie, Saint-Joseph, Vouvray, vin jaune, vin de paille et autres champagnes.

Résumé

L'épidémiologie évolutive des maladies infectieuses vise à comprendre les interactions entre les processus épidémiologiques et évolutifs. Cette approche est particulièrement pertinente pour étudier la dynamique transitoire des maladies émergentes, lorsque les échelles de temps épidémiologiques et évolutives se chevauchent. Il est par exemple crucial pour comprendre la pandémie de COVID-19 de tenir compte de l'évolution du virus et de la succession des variants préoccupants. Cette adaptation peut affecter des traits phénotypiques clés et diminuer notre capacité à contrôler les épidémies. L'avènement des méthodes de séquençage haut-débit permet aujourd'hui de collecter des données génétiques et de suivre à la fois spatialement et temporellement la distribution des différentes souches. Dans cette thèse, je combine épidémiologie évolutive théorique et inférence statistique en utilisant des données épidémiologiques et génétiques pour estimer les phénotypes des agents pathogènes, dans le domaine de la santé publique et de la microbiologie expérimentale. Mon travail repose sur l'analyse de modèles déterministes basés sur des systèmes dynamiques d'équations différentielles ordinaires.

Tout d'abord, je m'intéresse au variant Alpha du SARS-CoV-2 en Angleterre et j'étudie les caractéristiques phénotypiques à l'origine de son avantage sélectif. Pour cela, je développe une approche en deux étapes basée sur des modèles épidémiologiques SEIR (*Susceptible-Exposed-Infectious-Recovered*). Dans une première étape, avant l'émergence du variant, j'estime l'impact des interventions non pharmaceutiques sur la propagation du virus. Dans une deuxième étape, après l'émergence du variant, j'exploite la dynamique lente-rapide des processus éco-évolutifs pour estimer les différences phénotypiques entre les deux lignées virales en compétition. Je montre que l'avantage sélectif du variant Alpha est davantage dû à un taux de transmission accru qu'à une période de contagiosité plus longue. Deuxièmement, je m'intéresse à une expérience d'évolution expérimentale qui suit la dynamique épidémiologique et évolutive du bactériophage tempéré λ au cours de sa propagation dans une population bactérienne d'*Escherichia coli*. Je développe ici une nouvelle approche d'inférence pour estimer les phénotypes viraux à différents stades de l'épidémie - y compris des traits phénotypiques très difficiles à estimer par ailleurs. Je modélise des processus cachés tels que la lyse et la lysogénie et j'ajuste ce nouveau modèle à un jeu de données incomplètes. Troisièmement, j'analyse comment la migration entre des populations d'hôtes peut impacter l'épidémiologie et l'évolution transitoire d'un agent pathogène. Pour cela, je simule la dynamique évolutive transitoire de la compétition entre deux souches dans un modèle SIRS. Je montre comment la migration peut biaiser la quantification de la force de la sélection et fausser les interprétations de l'avantage sélectif réel des variants.

Ces trois projets me permettent de développer de nouveaux outils pour exploiter des jeux de données donnant une description incomplète (processus cachés et données manquantes) de la dynamique d'agents pathogènes se propageant et évoluant dans un environnement hétérogène. Je montre notamment que prendre en compte la structure de l'habitat du pathogène dans différents compartiments peut être essentielle pour estimer les paramètres des modèles, d'où l'importance de la disponibilité de données stratifiées. Ce travail souligne comment l'analyse théorique et statistique des dynamiques épidémiologiques et évolutives des maladies infectieuses peut éclairer notre compréhension de l'évolution phénotypique et de l'adaptation des agents pathogènes.

Mots clés: Épidémiologie évolutive – Modélisation des maladies infectieuses – Équations différentielles ordinaires – Inférence statistique – SARS-CoV-2 – Bactériophage

Abstract

Evolutionary epidemiology theory of infectious diseases aims to understand the interplay between epidemiological and evolutionary processes. This approach is particularly useful to study the transient dynamics of emerging pathogens, when epidemiological and evolutionary timescales overlap. For instance, it is essential for understanding the COVID-19 pandemic to account for the evolution of the virus and the succession of variants of concern. This adaptation can affect key phenotypic traits and erode our ability to mitigate epidemics. The advent of sequencing methods makes it now possible to collect genetic data and track the distribution of different strains across space and time. In this thesis, I combine evolutionary epidemiology theory and statistical inference using both epidemiological and genetic data to estimate pathogen phenotypes in public health and experimental microbiology. Throughout, my work relies on the analysis of deterministic models based on dynamical systems of ordinary differential equations.

First, I focus on the rise of the SARS-CoV-2 Alpha variant in England and I explore which phenotypic traits drive its selective advantage. For this purpose, I develop a two-step approach based on SEIR (*Susceptible-Exposed-Infectious-Recovered*) epidemiological models. In the first step, before the emergence of the variant, I estimate the impact of the intensity of non-pharmaceutical interventions on the spread of the virus. In a second step, after the emergence of the variant, I exploit the slow-fast dynamics of eco-evolutionary processes to infer the phenotypic differences between the two competing lineages. I show that the selective advantage of the Alpha variant is likely driven by a higher transmission rate than by a longer infectious period. Second, I focus on an evolution experiment that tracks the epidemiological and evolutionary dynamics of the temperate bacteriophage λ throughout its spread in a bacterial population of *Escherichia coli*. I develop a new inference approach to estimate the viral phenotypes at different stages of the epidemic – including phenotypic traits very difficult to estimate otherwise. I model hidden processes such as lysis and lysogeny and fit this new model to an incomplete dataset. Third, I analyse how migration between host populations can affect the transient epidemiology and evolution of the pathogen. To do so, I track the transient evolutionary dynamics of the competition between two strains in an SIRS model. I show how migration can bias the quantification of the strength of selection and lead to misinterpretations about the real selective advantage of variants.

These three projects allow me to develop new tools to exploit datasets that give access to an incomplete description (hidden processes and missing data) of the dynamics of a pathogen spreading and evolving in a heterogeneous host environment. Notably, I show that accounting for the pathogen structure among different compartments can be key to estimate model parameters, highlighting the importance of the availability of stratified data. This work emphasizes how the theoretical and statistical analysis of the joint epidemiological and evolutionary dynamics of infectious diseases can provide insights on the phenotypic evolution driving pathogen adaptation.

Keywords: Evolutionary epidemiology – Infectious diseases modelling – Ordinary differential equations – Statistical inference – SARS-CoV-2 – Bacteriophage

Contents

Chapter one

General introduction | 1

1.1 Preamble

Late 2019, less than two years before the start of this thesis, the SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) emerged in the market of Wuhan, Hubei Province, China [2]. The respiratory disease transmitted by the virus, namely COVID-19 (Coronavirus Disease 2019), spread rapidly and worldwide. By March 2020, COVID-19 reached at least 114 countries and was categorized as a pandemic by the World Health Organization [3]. Understanding and predicting the epidemiological dynamics of COVID-19 was therefore a major challenge for public health; epidemiologists urgently needed to understand its transmission dynamics, assess the potential burden of the pandemic, and design strategies to mitigate its spread. Mathematical modelling has been a critical tool in this effort.

[2]: Lu et al. (2020), 'Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle'

[3]: WHO (2020), 'WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020'

A model is a simplified representation of the reality: many details of the underlying processes driving the dynamics of complex systems have to be approximated (i.e., factors considered to be less important are neglected). While theoretical approaches have often been underrated compared to empirical approaches [4], mathematical models are really useful and powerful tools that allow to formalize, analyse, understand and provide qualitative or quantitative predictions about biological processes and support decision-making. Models are typically either made to improve our understanding of complex systems or to make predictions. The study of infectious disease epidemiology has always been closely intertwined with mathematical modelling. In the second half of the 18th century, Bernoulli developed an epidemiological model to analyse data of smallpox morbidity and mortality and investigated the benefit of pathogen inoculation [5]. Late 1800 early 1900, Ronald Ross (Nobel Prize 1902, Physiology or Medicine) proposed the first mathematical models of malaria transmission and made pioneering contributions to quantitative theory in epidemiology, which he coined *a priori* pathometry. In particular, he demonstrated that malaria was transmitted through bites of infected *Anopheles* mosquitoes (vector-borne disease) and proposed intervention strategies to control the disease [6–8]. In 1927, Kermack and McKendrick published an article that has popularized the use of deterministic compartmental models to simulate epidemiological dynamics [9]. These compartmental models have since been used extensively, and in particular for the COVID-19 pandemic. Moreover, mathematical models can be fitted to data (e.g., reported number of positive cases or deaths) to estimate key parameters, such as the mean number of secondary infections – estimated for instance at around 2.9 (95% confidence interval: 2.81 to 3.01) at the beginning of the epidemic of COVID-19 in France [10]. Mathematical models have also been used to forecast the future dynamics of COVID-19 in different scenarios – for example with different control strategies –, notably to anticipate saturation in health care demand in hospitals (e.g., [11, 12]).

[4]: Goldstein (2018), 'Are theoretical results 'Results'?'

[5]: Bernoulli (1760), 'An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it'

[6]: Ross (1899), 'Inaugural lecture on the possibility of extirpating malaria from certain localities by a new method'
[7]: Ross (1905), 'The logical basis of the sanitary policy of mosquito reduction'
[8]: Ross (1911), *The prevention of malaria*

[9]: Kermack et al. (1927), 'A contribution to the mathematical theory of epidemics'

[10]: Salje et al. (2020), 'Estimating the burden of SARS-CoV-2 in France'

[11]: Ferguson et al. (2020), 'Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand'
[12]: Paireau et al. (2022), 'An ensemble model based on early predictors to forecast COVID-19 health care demand in France'

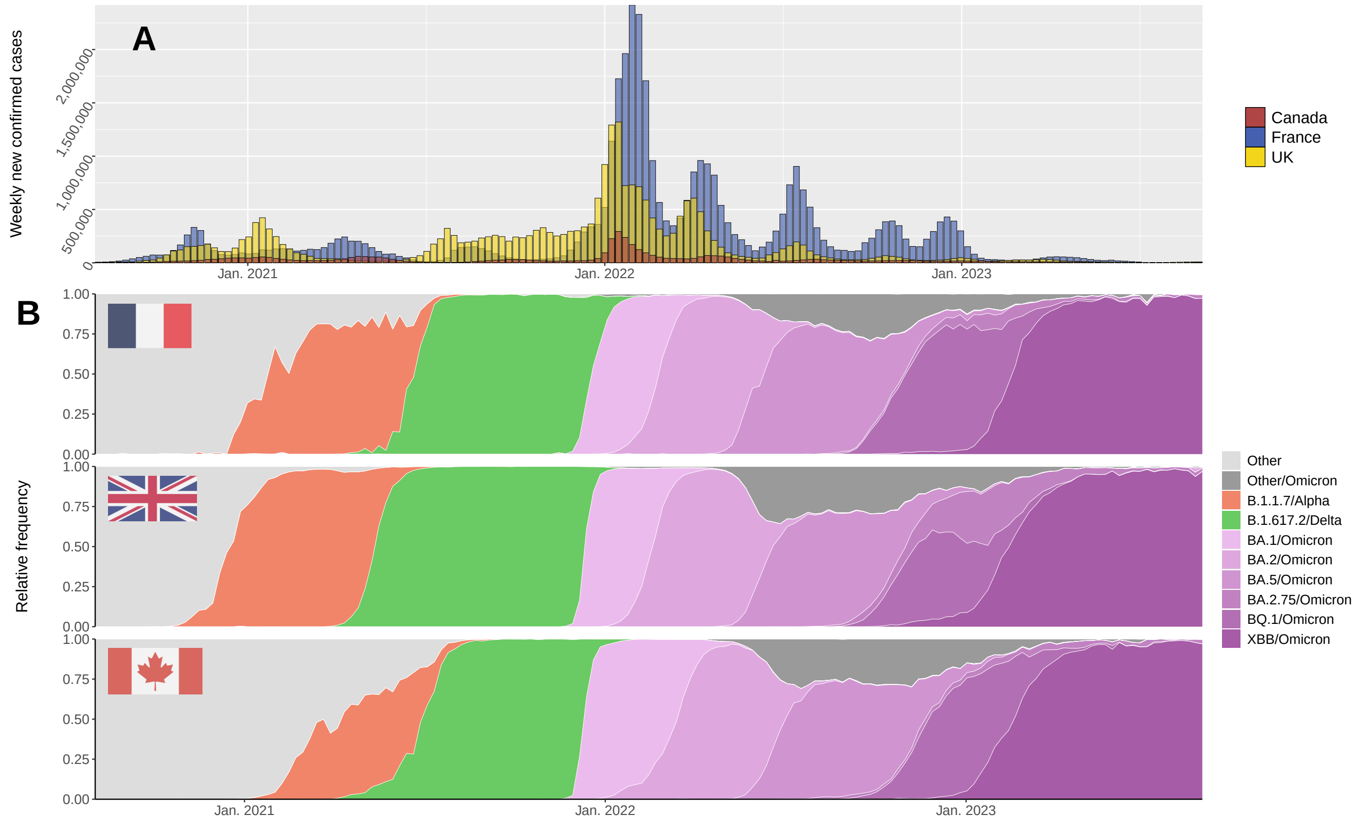


Figure 1.1: Epidemiology and evolution of SARS-CoV-2 across France, the UK and Canada. I use publicly available data from 2020-08-01 to 2023-09-01. (A) Epidemiology: weekly new confirmed COVID-19 cases (WHO data, downloaded from *Our World in Data*); (B) Evolution: relative frequencies of several variants of concern (metadata source: 4,020,732 sequences available on GISAID).

At the onset of the pandemic, the evolutionary potential of SARS-CoV-2 was thought to be very limited [13, 14]. Evolutionary dynamics are generally assumed to occur on timescales that are much slower than epidemiological dynamics. Yet, the D614G mutation (spike protein substitution, likely to increase transmission) emerged during the early phase of the pandemic (around May 2020) and became the dominant strain of the COVID-19 pandemic [15–17]. The virus was mutating away from the ancestral strain but those mutations were first considered to be either detrimental or neutral for viral fitness. The rise of some mutations, like D614G, could result from the influence of demographic stochasticity [18]. Later in 2020, the SARS-CoV-2 Alpha variant (Pango lineage B.1.1.7) emerged in England [19, 20]. The independent increase in frequency of this variant in different countries (i.e., demographic stochasticity cannot explain this parallel evolution, see **Figure 1.1**) changed dramatically the way evolution was considered in the COVID-19 pandemic. Viral adaptation (i.e., when evolution is driven by natural selection) was suddenly becoming an important factor that needed to be monitored and many countries started to sequence the virus after the rise of the Alpha variant. Alpha was categorized as variant of concern (VOC) – that is a variant with a selective advantage – and was the first of a succession of VOCs that successively emerged and replaced the previous lineage – e.g., Delta (Pango lineage B.1.617.2), or Omicron (first Pango lineage B.1.1.529) – (**Figure 1.1**, using sequences data from GISAID [21]). The COVID-19 pandemic thus illustrates how crucial it is to characterize pathogen phenotypes, to track

[13]: Grubaugh et al. (2020), 'We shouldn't worry when a virus mutates during disease outbreaks'
[14]: Rausch et al. (2020), 'Low genetic diversity may be an Achilles heel of SARS-CoV-2'
[15]: Korber et al. (2020), 'Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus'
[16]: Plante et al. (2021), 'Spike mutation D614G alters SARS-CoV-2 fitness'
[17]: Volz et al. (2021), 'Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity'
[18]: Grubaugh et al. (2020), 'Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear'
[19]: Public Health England (2020), *Investigation of novel SARS-COV-2 variant 202012/01: technical briefing 5*
[20]: Volz et al. (2021), 'Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England'
[21]: Khare et al. (2021), 'GISAID's role in pandemic response'

pathogen phenotypic evolution and to understand the underpinnings of such evolution. Yet, while statistical approaches are really common in epidemiology, coupling evolution and epidemiology has generally been limited to theoretical approaches.

In this thesis, I combine a theoretical approach based on dynamical mechanistic models and a statistical approach to estimate model parameters, such as key phenotypic traits of pathogens, in public health and experimental microbiology. **Figure 1.2** is a schematic of the interactions between models and data. Data and observations are used qualitatively to tailor a dynamical model to the biology of a particular host-pathogen system, tracking the interplay between epidemiology and evolution. Through theoretical analyses and numerical simulations, models provide useful insights to understand the evolutionary epidemiology of infectious diseases, and even provide theoretical predictions. The confrontation between demographic and genetic data and the outcomes of models can qualitatively confirm theoretical predictions. More quantitatively, statistical models built from dynamical models can be fitted to time series data to estimate model parameters and to make model comparisons. Combining effectively theoretical and statistical approaches is an iterative process, back and forth between models and data. In this chapter, I begin with dynamical models and introduce the field of the evolutionary epidemiology of infectious diseases. Next, I focus on data and statistical inference. In particular, I present some concepts and methods in statistical inference that I used in this thesis to estimate model parameters. As a guiding thread throughout these two parts, I frequently rely on a deterministic epidemic model to illustrate the different concepts – model notations are summarized in **Table 1.1**. Lastly, I present in more details the objectives of my research projects.

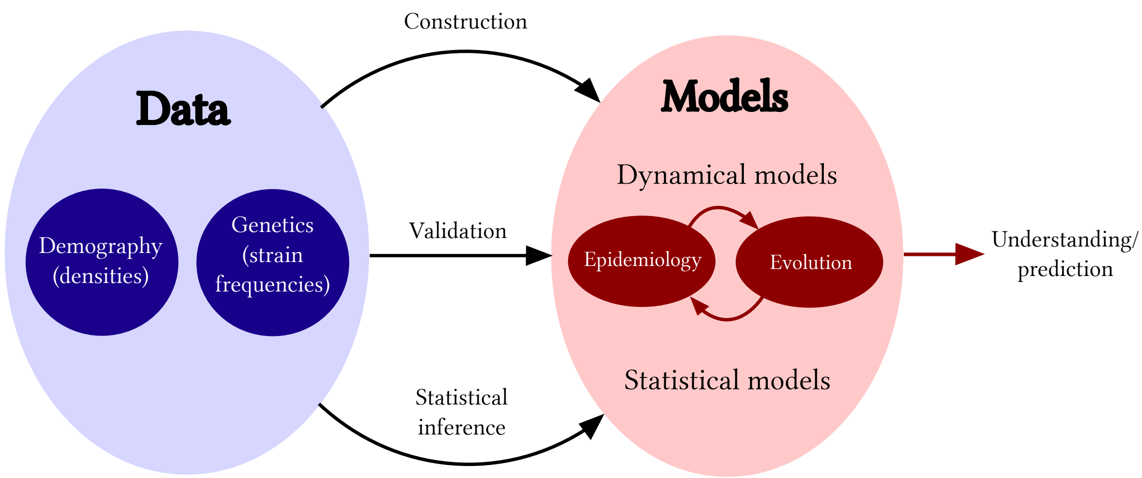


Figure 1.2: Schematic of the relationship between models and data in evolutionary epidemiology. Mathematical models in evolutionary epidemiology are simplified and formalised representations of the coupled epidemiological and evolutionary processes; epidemiological and evolutionary feedback shape both epidemiological and evolutionary dynamics. Through theoretical analyses and numerical simulations, dynamical/mechanistic models are useful to understand the evolutionary epidemiology of infectious diseases and yield theoretical predictions on both the demography and the evolution of the system. Such predictions can be (in)validated by experimental/empirical data. Data guide modelling choices for a particular host-pathogen system. Fitting statistical models to the data enables to estimate parameter values. Here, I use a combination of demographic data (densities) and genetic data (strain frequencies of the pathogen).

Table 1.1: Notations. The subscripts w and m refer to the wildtype strain and the mutant strain (or variant), respectively. Overlines refer to mean values of life-history traits across all genotypes.

Term	Definition
N	Host population
S	Susceptible hosts
$I,\ I_w,\ I_m$	Infected (and infectious) hosts
R	Recovered (and immune) hosts
q	Frequency of the variant
$\beta, \overline{\beta}, \beta_w, \beta_m$	*Per capita* transmission rates; $\overline{\beta} = (1-q)\beta_w + q\,\beta_m$
$\gamma, \overline{\gamma}, \gamma_w, \gamma_m$	*Per capita* recovery rates; $\overline{\gamma} = (1-q)\gamma_w + q\,\gamma_m$
$\alpha, \overline{\alpha}, \alpha_w, \alpha_m$	Virulence (*per capita* pathogen-induced mortality rates); $\overline{\alpha} = (1-q)\alpha_w + q\,\alpha_m$
$\Delta\beta,\ \Delta\gamma,\ \Delta\alpha$	Phenotypic differences between the variant and the wildtype; $\Delta\beta = \beta_m - \beta_w$, $\Delta\gamma = \gamma_m - \gamma_w$, $\Delta\alpha = \alpha_m - \alpha_w$
λ	Recruitment of susceptible hosts (births, net migration)
δ	*Per capita* natural mortality rate
ζ	*Per capita* rate of immunity waning
$r,\ \overline{r},\ r_w,\ r_m$	Growth rates of the epidemic (absolute pathogen fitness)
$\mathscr{R}_0$	Basic reproduction number of the pathogen
$\mathscr{R}$	Effective reproduction number of the pathogen
G	Generation interval
$\overline{G}$	Mean generation interval
$\mathscr{S}$	Selection gradient (relative fitness) of the variant (rate at which it grows or declines in frequency on the logit scale)
ρ	Reporting rate of infected hosts
θ	Parameter(s) of interest to estimate
$\hat{\theta}$	Estimator/estimation of the parameter(s) of interest
Θ	Space of the parameter(s) of interest
σ	Nuisance parameter(s)
$\hat{\sigma}$	Estimator/estimation of the nuisance parameter(s)
$y_{1:n}$	Data (sequence of n observations)
$\mathscr{L}$	Likelihood

1.2 On the evolutionary epidemiology of infectious diseases

Evolutionary epidemiology theory of infectious diseases aims to understand and disentangle the interplay between epidemiological and evolutionary processes that are acting upon host-pathogen systems. In that respect, evolutionary epidemiology theory focuses on the joint temporal dynamics of epidemiological densities and of phenotypic traits' distributions; it investigates not only their long-term dynamics but also their transient (short-term) dynamics, especially when epidemiological and evolutionary timescales overlap (quantitative genetics approach [22]). Crucially, the phenotypic evolution of pathogens is shaped by their environment which, at the population level and from the point of view of the pathogen, is the availability and the quality of susceptible hosts. It is therefore important to take epidemiological feedback into account and I thus start by presenting some general concepts in epidemiology before introducing evolution in a second step. Next, I consider a polymorphic pathogen population where two strains with distinct strategies (phenotypes) co-circulate and I show how the classical framework of adaptive dynamics can be used to identify the long-term outcome of the competition. Under the assumption that mutations are rare, this framework relies on decoupling the epidemiological and evolutionary dynamics : a variant only emerges when the previous strains has reached an endemic equilibrium [23, 24]. This is not always the case however, and I later show how an evolutionary epidemiology framework enables to track transient dynamics.

[22]: Lion et al. (2023), 'Extending eco-evolutionary theory with oligomorphic dynamics'

[23]: Geritz et al. (1998), 'Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree'
[24]: Dieckmann (2002), 'Adaptive dynamics of pathogen-host interactions'

1.2.1 Epidemiology: example with a monomorphic SIRS model

Epidemiological dynamics

Modelling the epidemiological dynamics of a given infectious disease is most frequently tackled through the framework of deterministic compartmental models, as popularized by [9]. This class of epidemic models assumes an homogeneous-mixing host population (mean-field approach), stratified between different compartments (or state variables) depending on their epidemiological status. For instance, in the famous SIR or SIRS model, compartments are: S (susceptible), I (infected/infectious) and R (recovered, and immune). With such between-host (in contrast to within-host) models, one track the density of hosts over time in each compartment but not explicitly that of pathogens [25]. Furthermore, unlike individual-based models, dynamical models do not rely on an explicit tracking of each individual in the system but only of the density of individuals in each compartment.

[9]: Kermack et al. (1927), 'A contribution to the mathematical theory of epidemics'

[25]: Anderson et al. (1991), *Infectious diseases of humans: dynamics and control*

In the following, I use a version of an SIRS model in continuous time – a system of ordinary differential equations (ODEs) – for illustration (**Figure 1.3**). In contrast with the SIR model – the flagship model of mathematical epidemiology –, in which recovered hosts acquire lifelong immunity against reinfection, an SIRS model allows for the recovered hosts to be immune only temporarily (waning of immunity) and to return

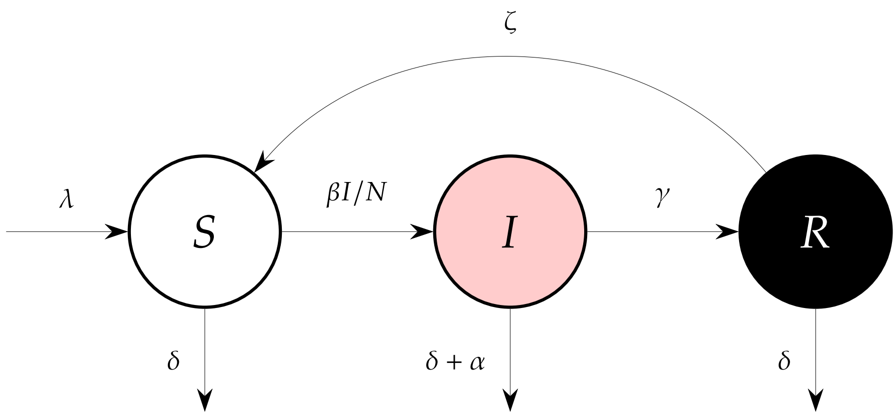


Figure 1.3: Flow chart of an SIRS model. The host population (with total density N) is divided between susceptible (S), infected and infectious (I) and recovered (R) hosts. Each epidemiological status, or compartment, is represented by circles and arrows indicate transitions between compartments. Parameters governing each transition are reported with the corresponding arrow. New susceptible hosts are recruited with a constant influx λ and natural death occurs in all compartments at a *per capita* rate δ; β is the the effective, direct and horizontal *per capita* transmission rate; infected hosts recover at a *per capita* rate γ and become fully immune or die of the disease at a *per capita* rate α (virulence); immunity wanes at a *per capita* rate ζ.

at some point to the S compartment. Hence, the SIR model is a special case of the more general SIRS model. The second motivation to use an SIRS model here is that immunity waning is an important feature of many infectious diseases, such as COVID-19 [26–28] which has been a strong focus of this thesis. Throughout, the density of a compartment X at the current time t is denoted $X(t)$, for example the density of the S compartment at a given time is $S(t)$. $N(t) = S(t) + I(t) + R(t)$ is the total size of the host population. In terms of demography, new susceptible hosts are recruited with a constant influx λ (births, migration) and natural death occurs in all compartments at a *per capita* rate δ. Transmission is assumed to be horizontal — i.e., no parent-progeny relationship – and to happen through direct contacts between individuals. Let β be the effective *per capita* transmission rate, a product between the host contact rate and the probability of transmission per contact with an infectious individual. The force of infection – i.e., the *per capita* infection rate of susceptible hosts – is here given by $\beta I(t)/N(t)$, where the division by $N(t)$ indicates that transmission is frequency-dependant rather than density-dependant [29]. Infected hosts either recover at a *per capita* rate γ or die at a *per capita* rate α. I will then refer to α as the virulence, defined in theoretical biology as the additional (pathogen-induced) mortality rate. This clarification is important as the term 'virulence' have different meanings across fields – e.g,. degree of ability to infect, pathogen-induced damages, notably in plant pathology. Recovered hosts are fully immune to the disease but may become susceptible again at a *per capita* rate ζ governing immunity waning. These epidemiological trajectories are modeled using the following system of non-linear ODEs where the dot refers to differentiation with respect to time:

[26]: Levin et al. (2021), 'Waning immune humoral response to BNT162b2 Covid-19 vaccine over 6 months'
[27]: UKHSA (2022), *COVID-19 vaccine surveillance report – Week 16*
[28]: Carabelli et al. (2023), 'SARS-CoV-2 variant biology: immune escape, transmission and fitness'

[29]: McCallum et al. (2001), 'How should pathogen transmission be modelled?'

$$\begin{cases} \dot{S}(t) &= \lambda - \beta S(t)\dfrac{I(t)}{N(t)} - \delta S(t) + \zeta R(t) \\ \dot{I}(t) &= \left(\beta \dfrac{S(t)}{N(t)} - \delta - \alpha - \gamma\right) I(t) \\ \dot{R}(t) &= \gamma I(t) - \delta R(t) - \zeta R(t) \end{cases} \tag{1.1}$$

Using ODEs, sojourn times are implicitly assumed to be exponentially distributed [30], and thus Markovian or memoryless [31, 32]. In (1.1), the generation interval G – i.e., the period of time between a primary infection and one of its secondary infections [33] – is thereby exponentially distributed:

$$G \sim \mathscr{E}\big(\delta + \alpha + \gamma\big),$$

where $\mathscr{E}$ represents the exponential distribution; and the mean generation interval, denoted $\overline{G}$, is here equal to the inverse of the rate of leaving the infected compartment [30]:

$$\overline{G} = \frac{1}{\delta + \alpha + \gamma} \tag{1.2}$$

In this model, the duration of infectiousness (sojourn time in I) follows the same distribution as that of the generation interval G. The dynamics of symptoms is not modelled in (1.1) but note that it is often more frequent to know the dates of symptom onset rather than the actual dates of infection. Thus, the serial interval – i.e., the amount of time elapsed between the symptom onset of a primary infection and that of one of its secondary infections – is often used instead of the generation interval to estimate the mean $\overline{G}$.

[30]: Wallinga et al. (2007), 'How generation intervals shape the relationship between growth rates and reproductive numbers'

[31]: Forien et al. (2021), 'Estimating the state of the COVID-19 epidemic in France using a model with memory'

[32]: Sofonea et al. (2021), 'Memory is key in capturing COVID-19 epidemiological dynamics'

[33]: Svensson (2007), 'A note on generation times in epidemic models'

[30]: Wallinga et al. (2007), 'How generation intervals shape the relationship between growth rates and reproductive numbers'

The term between parentheses in the expression of $\dot{I}(t)$ in (1.1) is the (Malthusian) growth rate at time t of the epidemic, $r(t)$:

$$r(t) = \beta\frac{S(t)}{N(t)} - \delta - \alpha - \gamma, \tag{1.3}$$

which measures the speed of the infection at the population level [34]. Here, the transmission rate is assumed to be constant over time but the model can be readily extended to integrate time-varying changes in β such as seasonal forcing or disease control measures.

[34]: Park et al. (2019), 'A practical generation-interval-based approach to inferring the strength of epidemics from their speed'

Emergence and equilibrium analysis

In the absence of the pathogen, the system (1.1) converges towards a disease-free equilibrium where $S(\infty) = \lambda/\delta$. Alternatively, when a small quantity of the pathogen is introduced (at $t = 0$ for simplicity) into the otherwise fully susceptible population ($S(0)/N(0) \approx 1$), the fate of this host-pathogen system is governed by the sign of $r(0) = \beta - (\delta + \alpha + \gamma)$. When $r(0) < 0$, infections cannot reproduce themselves; therefore the pathogen goes extinct and the host population converges to the previous disease-free equilibrium. In contrast, when $r(0) > 0$, an outbreak breaks out and eventually stabilises to the following epidemiological attractor (endemic equilibrium):

$$\begin{cases} S(\infty) & = \dfrac{\lambda(\delta+\gamma+\zeta)(\delta+\alpha+\gamma)}{(\beta-\alpha)(\delta\gamma+(\delta+\alpha)(\delta+\zeta))-\zeta\alpha\gamma} \\ I(\infty) & = \dfrac{\lambda(\delta+\zeta)\big(\beta-(\delta+\alpha+\gamma)\big)}{(\beta-\alpha)(\delta\gamma+(\delta+\alpha)(\delta+\zeta))-\zeta\alpha\gamma} \\ R(\infty) & = \dfrac{\lambda\gamma\big(\beta-(\delta+\alpha+\gamma)\big)}{(\beta-\alpha)(\delta\gamma+(\delta+\alpha)(\delta+\zeta))-\zeta\alpha\gamma} \end{cases} \tag{1.4}$$

Notation reminder

- S: susceptible hosts
- I: infected/infectious hosts
- R: recovered hosts
- β: transmission rate
- γ: recovery rate
- α: virulence
- λ: influx of S
- δ: natural mortality rate
- ζ: rate of immunity waning

The threshold criterion $r(0) > 1$ is equivalent to:

$$\mathcal{R}_0 = \frac{\beta}{\delta + \alpha + \gamma} > 1, \tag{1.5}$$

where $\mathcal{R}_0$ is the basic reproduction number of the pathogen. More intuitively, $\mathcal{R}_0$ corresponds to the expected number of secondary infections caused by one primary infected individual (index case) in an otherwise fully susceptible population [25, 35–37]. The basic reproduction number provides insights on the potential spread of an emerging epidemic and plays thus a key role in infectious disease epidemiology. Unlike $r(t)$, however, $\mathcal{R}_0$ is not a rate but a dimensionless metric; therefore, it does not tell us anything about the speed of the epidemic.

[25]: Anderson et al. (1991), *Infectious diseases of humans: dynamics and control*
[35]: Anderson et al. (1982), 'Coevolution of hosts and parasites'
[36]: Diekmann et al. (1990), 'On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations'
[37]: Diekmann et al. (2000), *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*

Notation reminder

- $\mathcal{R}_0$: basic reproduction number
- $\mathcal{R}$: effective reproduction number
- r: growth rate of the epidemic
- $\overline{G}$: mean generation interval

Transient state of the epidemic

Through the course of an outbreak, the pool of susceptible hosts decreases and $\mathcal{R}_0$ becomes a poor predictor of the fate of the epidemic. Indeed, the success or failure of the pathogen is no longer governed by its performance in the initial (disease-free) environment. Instead, one may use the effective reproduction number $\mathcal{R}(t)$ that accounts for the current state of the epidemic and especially the remaining availability of susceptible hosts:

$$\mathcal{R}(t) = \mathcal{R}_0 \frac{S(t)}{N(t)}. \tag{1.6}$$

$\mathcal{R}(t)$ informs us on the strength of an epidemic [34, 38]. Equation (1.6) yields the following relationship between $\mathcal{R}(t)$, the growth rate of the epidemic $r(t)$ and the mean generation interval $\overline{G}$ [25, 30]:

[34]: Park et al. (2019), 'A practical generation-interval-based approach to inferring the strength of epidemics from their speed'
[38]: Dushoff et al. (2021), 'Speed and strength of an epidemic intervention'
[25]: Anderson et al. (1991), *Infectious diseases of humans: dynamics and control*
[30]: Wallinga et al. (2007), 'How generation intervals shape the relationship between growth rates and reproductive numbers'

$$\mathcal{R}(t) = 1 + \overline{G} \times r(t). \tag{1.7}$$

Note that this equation only holds for an exponentially distributed generation interval.

Reproduction numbers are among the most estimated quantities during real-time epidemic monitoring, particularly with the intention to design and assess control interventions that aim to reduce the spread of the pathogen – i.e., achieving $\mathcal{R}(t) < 1$. Yet, reproduction numbers can be difficult to estimate. Secondary infections can be counted using contact tracing data but this would always yield an underestimation of $\mathcal{R}(t)$. The speed of the epidemic $r(t)$ is also a key metric [34, 38]. Estimating $r(t)$ – and in particular its initial value $r(0)$, during the exponential growth phase of the epidemic – is typically performed from incidence time series and the generation interval distribution from contact tracing data [34]. $\mathcal{R}(t)$ can be estimated from previous inference of $r(t)$ and of the generation interval distribution using a moment generating function approach, in particular to deal with other distributions for the generation interval [30].

[34]: Park et al. (2019), 'A practical generation-interval-based approach to inferring the strength of epidemics from their speed'
[38]: Dushoff et al. (2021), 'Speed and strength of an epidemic intervention'
[34]: Park et al. (2019), 'A practical generation-interval-based approach to inferring the strength of epidemics from their speed'
[30]: Wallinga et al. (2007), 'How generation intervals shape the relationship between growth rates and reproductive numbers'

When the proportion of susceptible hosts drops to the critical value $S(t)/N(t) = 1/\mathcal{R}_0$, the epidemic reaches its peak before starting to decay ($\mathcal{R}(t) < 1$). This is the phenomenon under the notion of herd immunity, the immunity level that leads to the decay of the epidemic (here, $1-1/\mathcal{R}_0$), as introduced by [25]. The density of hosts infected after the peak has

[25]: Anderson et al. (1991), *Infectious diseases of humans: dynamics and control*

passed is called the overshoot of the epidemic [39]. Such a decline in the outbreak happens when a large enough fraction of the contacts of an infected host are immune to the disease. As shown with equations (1.4), the pathogen may however remain in the population without dying out in the long-term, circulating endemically, as long as susceptible hosts are still introduced in the system – e.g., births, net migration or immunity waning – [40].

[39]: Nguyen et al. (2023), 'Fundamental bound on epidemic overshoot in the SIR model'

[40]: Cobey (2020), 'Modeling infectious disease dynamics'

Figure 1.4 shows an example of simulation of system (1.1) (parameter values: $\lambda = 10$, $\delta = 0.1$, $\beta = 0.7$, $\gamma = 0.09$, $\alpha = 0.01$ and $\zeta = 0.001$) and illustrates several of the epidemiological elements presented above. Integration is solved numerically with the function `lsoda` from the R package `deSolve` [41].

[41]: Soetaert et al. (2010), 'Solving differential equations in R: package deSolve'

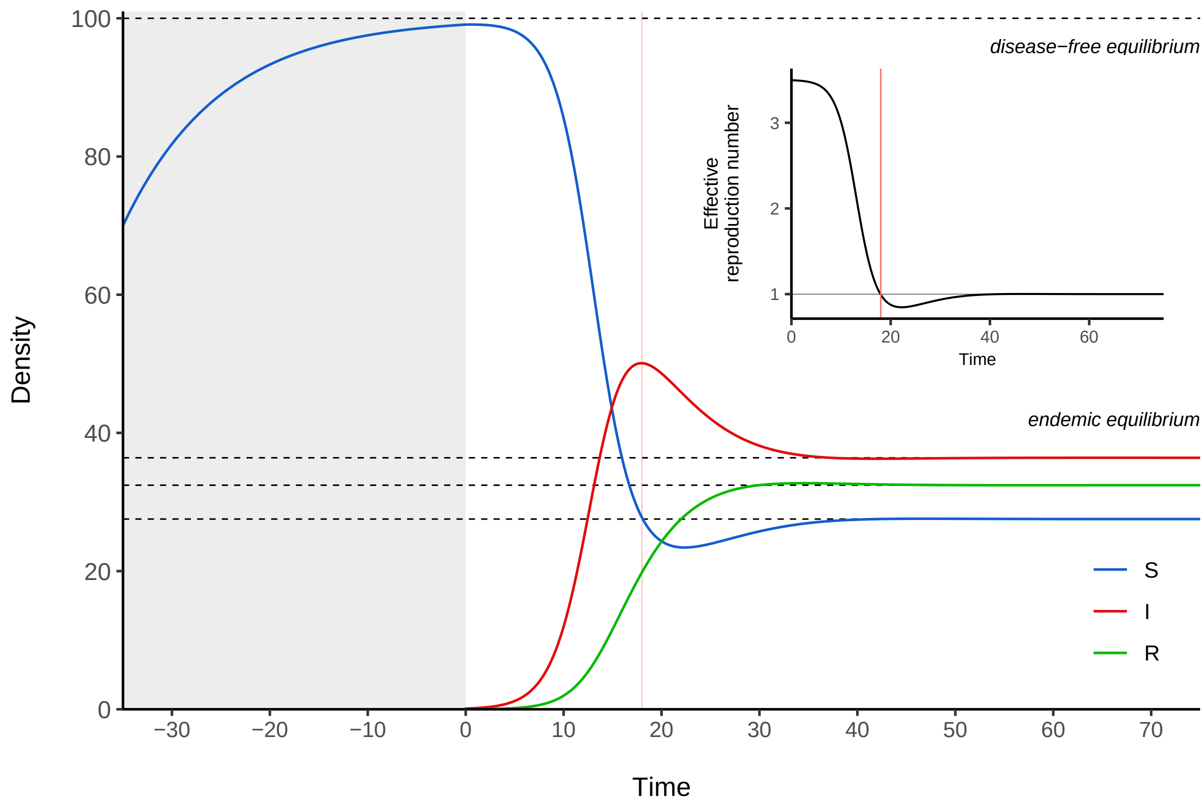


Figure 1.4: Simulation of the epidemiological dynamics of an SIRS model. I use model (1.1) with parameter values: $\lambda = 10$, $\delta = 0.1$, $\beta = 0.7$, $\gamma = 0.09$, $\alpha = 0.01$ and $\zeta = 0.001$. Prior to the introduction of the pathogen (grey background), the fully susceptible host population grows towards its disease free-equilibrium (top horizontal dashed line). At $t = 0$, the pathogen is introduced at very low frequency into the population ($I(0) = 0.1$). An epidemic then arises (basic reproduction number $\mathcal{R}_0 = 3.5 > 1$), reaches a peak (red vertical line) and eventually stabilizes at an endemic equilibrium (three lowest horizontal dashed lines). The small chart shows the dynamics of the effective reproduction number from $t = 0$ – when the effective reproduction number matches $\mathcal{R}_0$. The epidemic starts to decline when the population reaches herd immunity (the effective reproduction number falls below the threshold value 1).

1.2.2 Evolution

In this section, I begin with a short overview of the different forces driving evolution. I then focus on phenotypic evolution. In particular, introducing pathogen polymorphism (i.e., the coexistence of at least two variants) in the deterministic SIRS model (1.1), I present and discuss the frameworks of adaptive dynamics and evolutionary epidemiology.

Evolutionary forces

Evolution corresponds to a gradual change in heritable characteristics over generations in a population. Four evolutionary forces are discussed here: (i) mutation, (ii) natural selection, (iii) genetic drift and (iv) migration/dispersal.

Mutations are random and heritable alterations of the genetic information (DNA or RNA nucleotide sequence within a genome). Mutation rates vary widely across species – several orders of magnitude, between around 10^{-11} and 10^{-4} base substitutions per site per generation [42] – but vary also between strains of the same species (e.g., '*mutator*' bacteria have an increased adaptability to rapidly changing environmental pressures [43]). In particular, bacteria and viruses may exhibit really high mutation rates [42]. Mutation rates are generally higher for RNA genomes compared to DNA genomes, and higher for single-stranded genomes compared to double-stranded genomes [44]. Mutation rate also depends on the existence and the efficacy of proofreading and repair mechanisms (note that coronaviruses exhibit proofreading activities [45]). If mutations may arise spontaneously, the exposure to environmental mutagens – e.g. physical mutagens, such as UV, X rays or radioactivity, chemicals such as alkylating or DNA intercalating agents – can increase (sometimes dramatically) the background level of mutations. Crucially, *de novo* mutations are the ultimate source of genetic variation that fuels evolution. Different genotypes may lead to the same phenotype, as many mutations are neutral or nearly neutral [46] (e.g., owing to genetic code redundancy). Some mutations, on the other hand, result in another phenotype, associated with either a beneficial or a detrimental effect, increasing or decreasing the ability to survive and/or reproduce, respectively. Note that a mutation may affect different traits (pleiotropic mutation).

[42]: Gago et al. (2009), 'Extremely high mutation rate of a hammerhead viroid'

[43]: Pal et al. (2007), 'Coevolution with viruses drives the evolution of bacterial mutation rates'

[42]: Gago et al. (2009), 'Extremely high mutation rate of a hammerhead viroid'

[44]: Duffy et al. (2008), 'Rates of evolutionary change in viruses: patterns and determinants'

[46]: Kimura et al. (1968), 'Evolutionary rate at the molecular level'

Operating upon preexisting polymorphism, **natural selection** is a sorting mechanism driving adaptive evolution according to the strength of selective pressures. It corresponds to the differential survival and/or reproduction of phenotypes over generations [47]. Hence, natural selection is a directional force that increases the frequency of adaptive (beneficial) mutations in the population and purges deleterious ones. Crucially, survival and reproduction depends on the environment, so that natural selection is also environment-dependant. From the point of view of the pathogen, hosts represent the environment. A population experiencing a given environment is sometimes expected to go extinct; however, an evolutionary rescue may prevent extinction if adaptation is faster [48, 49].

[47]: Darwin (1859), *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*

[48]: Gonzalez et al. (2013), 'Evolutionary rescue: an emerging focus at the intersection between ecology and evolution'

[49]: Gandon et al. (2013), 'What limits the evolutionary emergence of pathogens?'

In contrast to natural selection, **genetic drift** is a stochastic process for which changes in the distribution of alleles in the population arise by chance rather than because of selective differences. For instance, for infectious diseases, such stochasticity may be introduced by superspreaders – i.e., individuals who infect a disproportionately large number of secondary cases – [50]. Without any deterministic force, allele frequencies undergo random walk processes since genes are randomly sampled from one generation to the next [51]. Populations with greater effective sizes are less subject to genetic drift than populations with smaller effective sizes (a theoretical population of infinite effective size is not affected at all by genetic drift). Small effective sizes are notably found in populations that

[50]: Markov et al. (2023), 'The evolution of SARS-CoV-2'

[51]: Lacy (1987), 'Loss of genetic diversity from managed populations: interacting effects of drift, mutation, immigration, selection, and population subdivision'

have gone through genetic bottlenecks. This is commonly experienced by pathogens upon transmission, as only a small amount of the genetic diversity found in an infectious host is transmitted to a susceptible host [50]. Like natural selection, genetic drift leads to the loss or the fixation of alleles and thus reduces the genetic diversity. This is particularly true for the loss of rare mutations, which, even if adaptive, are likely to be discarded through genetic drift. Genetic drift drives the evolution of alleles regardless of their potential adaptive values [52] and sometimes leads to the fixation of deleterious alleles, countering natural selection.

[50]: Markov et al. (2023), 'The evolution of SARS-CoV-2'

[52]: Wright (1955), 'Classification of the factors of evolution.'

While the first three evolutionary forces (mutation, natural selection and genetic drift) come into play in both open and closed populations, **migration** requires spatial heterogeneity. Gene flows tend to homogenize allele distributions between interconnected populations.

In this thesis, I focus on pathogen evolution and leave aside the evolution of the host. In particular, I focus on the competition between two strains (i.e., two phenotypes). *De novo* mutations are assumed not to occur, or at least not to lead to any other phenotype. Under the assumption that the two populations have both large sizes, I only use a deterministic framework to study the effect of natural selection (or of migration) and I thus neglect the effect of demographic stochasticity (genetic drift).

Adaptive dynamics in an SIRS model

The ability for pathogens to survive and reproduce depends on numerous phenotypic, or life-history, traits such as transmission, virulence, recovery or immunity escape [53]. In the long term, the evolutionary stable strategy (ESS) corresponds to the fittest phenotype, i.e., a strategy that cannot be invaded by any other variant. The ESS wins the evolutionary race and is a kind of evolutionary trap. Finding this singular strategy is classically tackled using an adaptive dynamics approach.

[53]: Day et al. (2020), 'On the evolutionary epidemiology of SARS-CoV-2'

Evolutionary invasion analysis Back to our previous monomorphic SIRS model (1.1), I now introduce genetic structure by modelling a polymorphic pathogen population. In particular, I track the competition between two strains: (i) the wildtype strain – hereafter denoted with the letter w – vs. (ii) a mutant strain, or variant – hereafter denoted by the letter m – (**Figure 1.5**). The host population, on the other hand, remains monomorphic and I only focus on pathogen evolution. The infected stage I is thus divided between hosts infected by the wildtype (I_w) and hosts infected by the variant (I_m), such that $I(t) = I_w(t) + I_m(t)$. I assume that the variant may differ phenotypically from the wildtype in its transmission rate $\beta_m = \beta_w + \Delta\beta$, its virulence $\alpha_m = \alpha_w + \Delta\alpha$ or its recovery rate $\gamma_m = \gamma_w + \Delta\gamma$. I assume that over-infections – including co-infections with both strains – do not occur and I also assume full cross-immunity – i.e., recovery from one strain confers immunity to the other. The temporal dynamics of hosts infected by strain k ($k \in \{w, m\}$) is thus given by:

$$\dot{I}_k(t) = \underbrace{\left(\beta_k \frac{S(t)}{N(t)} - \delta - \alpha_k - \gamma_k\right)}_{r_k(t)} I_k(t), \tag{1.8}$$

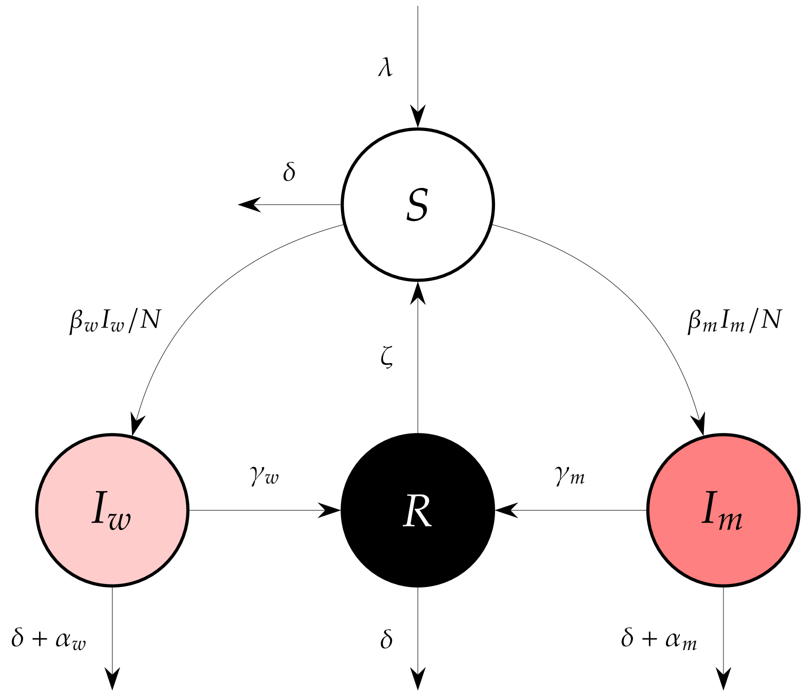


Figure 1.5: Flow chart of an SIRS model with two competitive pathogenic strains. I extend **Figure 1.3** to two pathogenic strains: the subscript w denotes the wildtype strain while the subscript m denotes the mutant strain (or variant). In terms of phenotypes, the wildtype and the variant may differ in transmission ($\beta_w \neq \beta_m$), recovery ($\gamma_w \neq \gamma_m$) or virulence ($\alpha_w \neq \alpha_m$).

with $r_k(t)$, the growth rate of strain k. Crucially, $r_k(t)$ is the absolute fitness of strain k, which depends on both the traits (β_k, δ, α_k and γ_k) and the environment (the availability of susceptible hosts $S(t)/N(t)$).

To simplify the analysis, the adaptive dynamics approach assumes that mutations are rare, so that a new variant only emerges after the previously established strain has stabilized to its epidemiological attractor [22–24]. This framework relies thus on a separation of timescale between epidemiological dynamics (fast) and evolutionary dynamics (slow). To determine whether or not the variant can invade the resident population (wildtype), I compute the invasion fitness of the variant, that is the growth rate of the variant r_m in the environment determined by the resident strain [54]. Setting equation (1.8) (with $k = w$) to 0 yields the proportion of susceptible hosts at the endemic equilibrium of the wildtype strain w:

$$\frac{S(t)}{N(t)} = \frac{\delta + \alpha_w + \gamma_w}{\beta_w} = \frac{1}{\mathcal{R}_{0,w}}$$

[22]: Lion et al. (2023), 'Extending eco-evolutionary theory with oligomorphic dynamics'
[23]: Geritz et al. (1998), 'Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree'
[24]: Dieckmann (2002), 'Adaptive dynamics of pathogen-host interactions'

[54]: Lion et al. (2018), 'Beyond R0 maximisation: on pathogen evolution and environmental dimensions'

For the variant to invade the resident population, its invasion fitness must be positive:

$$\underbrace{r_m\Big|_{\frac{S}{N}=\frac{1}{\mathcal{R}_{0,w}}} = \frac{\beta_m}{\mathcal{R}_{0,w}} - \delta - \alpha_m - \gamma_m}_{\text{Invasion fitness}} > 0.$$

After some rearrangements, this invasion fitness criterion is here equivalent to:

$$\mathcal{R}_{0,m} > \mathcal{R}_{0,w},$$

where $\mathcal{R}_{0,w}$ and $\mathcal{R}_{0,m}$ are the basic reproduction numbers of the wildtype and mutant strain, respectively. Finding analytically the best strategy in this model boils down to use the $\mathcal{R}_0$ maximization criterion. Equation

Notation reminder

- S/N: fraction of susceptible hosts
- β_w, β_w: transmission rates
- γ_w, γ_m: recovery rates
- α_w, α_m: virulence
- δ: natural mortality rate
- $\mathcal{R}_{0,w}$, $\mathcal{R}_{0,m}$: basic reproduction numbers

(1.5) shows that the basic reproduction number is an increasing function of the transmission rate (β) and a decreasing function of the rates of leaving the infectious state (α and γ, for the traits of the pathogen). Higher levels of transmission is beneficial for the pathogen, allowing it to spread more rapidly in the population; higher levels of virulence is detrimental for both the pathogen and its host (as the pathogen does not survive if the host dies); and higher recovery rates – and thus, lower durations of infectiousness – decrease the number of opportunities of pathogen transmissions. Therefore, the best strategy for the pathogen is to remain in the infectious state as long as possible ($\alpha + \gamma \to 0^+$) – with benign coexistence as the expected long-term evolutionary endpoint – coupled with the highest possible transmission rate ($\beta \to +\infty$) (**Figure 1.6**).

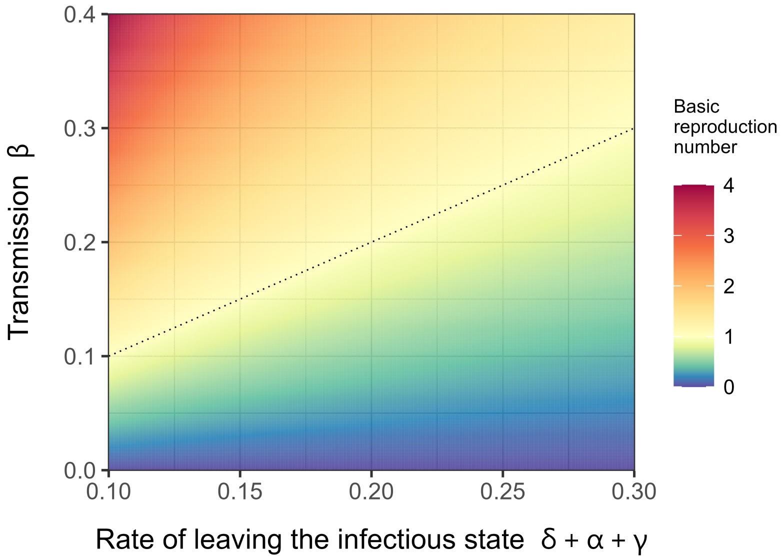


Figure 1.6: Landscape of the basic reproduction number in an SIRS model. I use the expression of $\mathcal{R}_0$ from equation (1.5). The dotted line indicate $\mathcal{R}_0 = 1$, below which all pathogenic strains go extinct. As the environmental feedback – the dynamics of susceptible hosts (resource) – is one-dimensional, adaptation is expected to maximize $\mathcal{R}_0$, that is pathogens are expected to evolve towards higher transmission rates and lower rates of leaving the infectious state.

Note however that the $\mathcal{R}_0$ maximization criterion is rather the exception than the rule: it holds when the environment is very simple and the environmental feedback (here, the dynamics of $S(t)$) only one-dimensional. In more complex models, however, evolution hardly maximizes $\mathcal{R}_0$ and computing the invasion fitness is then required [54]. More broadly, the strain that survives in the long-term would be the one that tolerates the worst environment (pessimization principle) [55]; this perspective is thus that evolution minimizes $S(t)$ (single resource) to the lowest density for which the disease still persist [54].

[54]: Lion et al. (2018), 'Beyond R0 maximisation: on pathogen evolution and environmental dimensions'

[55]: Diekmann (2004), 'A beginner's guide to adaptive dynamics'

[54]: Lion et al. (2018), 'Beyond R0 maximisation: on pathogen evolution and environmental dimensions'

On another matter, many observed examples of virulence maintenance has increasingly challenged the old wisdom that predicts that pathogens evolve inevitably towards avirulence [56]. These observations may be explained by relaxing the assumption that all life-history traits vary independently.

[56]: Alizon et al. (2009), 'Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future'

Phenotypic trade-off Phenotypic traits were previously assumed to vary independently from each other. However, many traits are not independent, i.e., positive or negative correlations exist between them. A famous example is the transmission-virulence trade-off. Based on the idea that, to be transmitted, pathogens need to exploit – and therefore harm – their host, this hypothesis suggests a positive covariance between these two traits [35]. Hence, selection for higher transmission can indirectly select for higher virulence [57]. This trade-off has been an alternative to the avirulence theory to explain the selection for intermediate levels of virulence and has been supported by several empirical studies [56]. In theoretical biology, the transmission-virulence trade-off is generally modeled by assuming that transmission is a monotonous increasing and concave, or saturating, function of virulence (see example in **Figure 1.7**). Trade-off functions add biological constraints and reduce therefore the set of feasible phenotypic combinations. Testing empirically the trade-off hypothesis may however turn out to be really complicated, and multiple trade-offs may exist [58].

[35]: Anderson et al. (1982), 'Coevolution of hosts and parasites'

[57]: Day et al. (2004), 'A general theory for the evolutionary dynamics of virulence'

[56]: Alizon et al. (2009), 'Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future'

[58]: Alizon et al. (2015), 'Adaptive virulence evolution: the good old fitness-based approach'

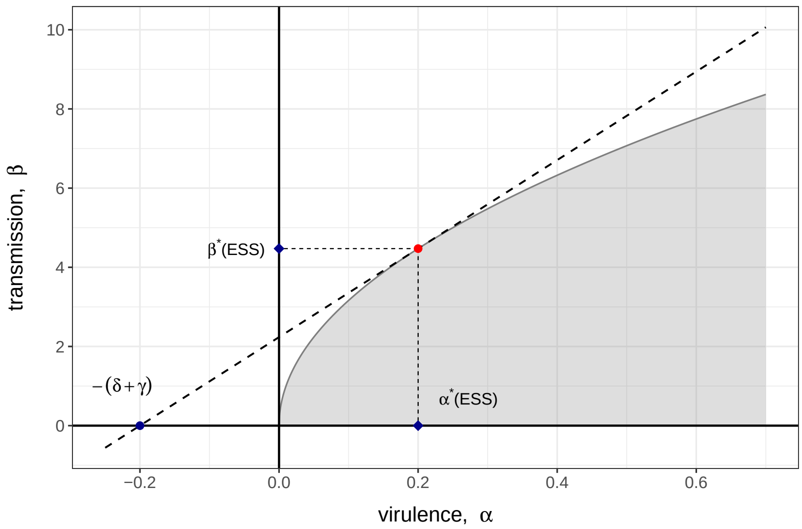


Figure 1.7: Trade-off between transmission and virulence. The transmission-virulence trade-off corresponds to the boundary (grey line) of the set of possible phenotypic combinations (shaded area). I use here the following concave function: $\beta(\alpha) = 10\sqrt{\alpha}$; fixed parameter values are: $\delta = 0.1$ and $\gamma = 0.1$. Equation (1.9) gives a useful geometric representation of the evolutionary stable strategy (ESS) which can be found graphically with the tangent of the curve that passes through the coordinates $(-(\delta+\gamma)\,;\,0)$ (dashed line).

Back to our previous example (1.8), I now assume, for the sake of simplicity, that the variant and the wildtype strains only differ in terms of transmission and virulence ($\gamma_m = \gamma_w = \gamma$) and that a trade-off function between β and α exists. The strategy maximizing $\mathcal{R}_0(\alpha)$ must verify:

$$\frac{\mathrm{d}\mathcal{R}_0(\alpha)}{\mathrm{d}\alpha} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\frac{\beta(\alpha)}{\delta+\alpha+\gamma}\right) = 0 \qquad \text{and} \qquad \frac{\mathrm{d}^2\mathcal{R}_0(\alpha)}{\mathrm{d}\alpha^2} < 0,$$

which yields:

$$\frac{\mathrm{d}\beta(\alpha)}{\mathrm{d}\alpha} = \frac{\beta(\alpha)}{\delta+\alpha+\gamma}. \tag{1.9}$$

This expression gives a useful geometric representation, allowing in particular the ESS to be found graphically (**Figure 1.7**). Taking the following trade-off function between transmission and virulence:

$$\beta(\alpha) = b\alpha^v$$

with $b \in \mathbb{R}_+^\star$ and $v \in]0;1[$, the virulence $\alpha^\star$ that maximizes $\mathcal{R}_0$ (ESS) is given by:

$$\alpha^\star = \frac{v}{1-v}(\delta+\gamma).$$

I show in **Appendix A** (teaching material) a similar evolutionary invasion analysis in an SIR model. Adaptive dynamics provides however no information about the transient (short-term) evolutionary dynamics when the competition takes place out of equilibrium.

Transient dynamics of pathogen phenotypic evolution in an SIRS model

In contrast to adaptive dynamics, evolutionary epidemiology does not necessarily assume that mutations are rare. This paradigm is particularly useful to study pathogen competition in more complex scenarios, while the epidemic has not yet reached an equilibrium (**Figure 1.8**).

Coupling epidemiological and evolutionary dynamics The epidemiological SIRS model (1.1) is now extended to account for both the wildtype strain and the variant as described in **Figure 1.5**:

$$\begin{cases} \dot{S}(t) &= \lambda - \overline{\beta}(t)S(t)\dfrac{I(t)}{N(t)} - \delta S(t) + \zeta R(t) \\ \dot{I}(t) &= \underbrace{\left(\overline{\beta}(t)\dfrac{S(t)}{N(t)} - \delta - \overline{\alpha}(t) - \overline{\gamma}(t)\right)}_{\overline{r}(t)} I(t) \\ \dot{R}(t) &= \overline{\gamma}(t)I(t) - \delta R(t) - \zeta R(t) \end{cases} \tag{1.10}$$

Notation reminder

- N: host population
- S: susceptible hosts
- I, I_w, I_m: infected/infectious hosts
- R: recovered hosts
- q: frequency of the variant
- $\beta, \overline{\beta}, \beta_w, \beta_m$: transmission rates
- $\gamma, \overline{\gamma}, \gamma_w, \gamma_m$: recovery rates
- $\alpha, \overline{\alpha}, \alpha_w, \alpha_m$: virulence
- λ: influx of S
- δ: natural mortality rate
- ζ: rate of immunity waning
- r, r_w, r_m: growth rates of the epidemic (absolute pathogen fitness)

where the overlines refer to mean values of the phenotypic traits after averaging over the distribution of strain frequencies:

$$\begin{cases} \overline{\beta}(t) &= \big(1-q(t)\big)\beta_w + q(t)\beta_m \\ \overline{\alpha}(t) &= \big(1-q(t)\big)\alpha_w + q(t)\alpha_m \\ \overline{\gamma}(t) &= \big(1-q(t)\big)\gamma_w + q(t)\gamma_m \end{cases} \tag{1.11}$$

with $q(t)$, the frequency of the variant at time t:

$$q(t) = \frac{I_m(t)}{I(t)}. \tag{1.12}$$

$\overline{r}(t)$ is the mean growth rate of the epidemic. Using the temporal dynamics of $I_m(t)$ (equation (1.8) with $k = m$) and of $I(t)$ (in system (1.10)) yields the following fundamental equation for the rate of change of the variant frequency [59]:

$$\dot{q}(t) = q(t)\big(r_m(t) - \overline{r}(t)\big). \tag{1.13}$$

[59]: Day et al. (2006), 'Insights from Price's equation into evolutionary epidemiology'

Thus, to increase in frequency, the absolute fitness of the variant $r_m(t)$ has to be higher than the mean fitness $\overline{r}(t)$. More broadly, equation (1.13) is not specific to this particular model and is known in population genetics and evolutionary game theory as a version of the replicator equation [22, 60, 61]. Using (1.13), the rate of change of the average trait values $\overline{\beta}(t)$, $\overline{\alpha}(t)$ and $\overline{\gamma}(t)$ are given respectively by the covariance between the trait and the fitness r (Price's equation without mutation [59]):

[22]: Lion et al. (2023), 'Extending eco-evolutionary theory with oligomorphic dynamics'
[60]: Taylor et al. (1978), 'Evolutionary stable strategies and game dynamics'
[61]: Schuster et al. (1983), 'Replicator dynamics'

[59]: Day et al. (2006), 'Insights from Price's equation into evolutionary epidemiology'

$$\begin{cases} \dot{\overline{\beta}}(t) = \mathbb{C}\text{ov}\,(\beta, r) &= \mathbb{V}(\beta)\dfrac{S(t)}{N(t)} - \mathbb{C}\text{ov}(\beta, \alpha) - \mathbb{C}\text{ov}(\beta, \gamma) \\ \dot{\overline{\alpha}}(t) = \mathbb{C}\text{ov}\,(\alpha, r) &= \mathbb{C}\text{ov}(\alpha, \beta)\dfrac{S(t)}{N(t)} - \mathbb{V}(\alpha) - \mathbb{C}\text{ov}(\alpha, \gamma) \\ \dot{\overline{\gamma}}(t) = \mathbb{C}\text{ov}\,(\gamma, r) &= \mathbb{C}\text{ov}(\gamma, \beta)\dfrac{S(t)}{N(t)} - \mathbb{C}\text{ov}(\gamma, \alpha) - \mathbb{V}(\gamma) \end{cases} \tag{1.14}$$

Notation reminder

- S/N: fraction of susceptible hosts
- q: frequency of the variant
- $\Delta\beta$: difference in transmission ($\Delta\beta = \beta_m - \beta_w$)
- $\Delta\gamma$: difference in recovery ($\Delta\gamma = \gamma_m - \gamma_w$)
- $\Delta\alpha$: difference in virulence ($\Delta\alpha = \alpha_m - \alpha_w$)
- r_w, r_m: growth rates of the epidemic (absolute pathogen fitness)
- $\mathcal{S}$: selection gradient of the variant

By extending in (1.13) the expressions for the growth rates $r_m(t)$ and $\overline{r}(t)$, I obtain here after some rearrangements:

$$\dot{q}(t) = \underbrace{q(t)\big(1 - q(t)\big)}_{\text{Genetic variance}} \underbrace{\mathcal{S}(t)}_{\text{selection gradient}}, \tag{1.15}$$

with:

$$\mathcal{S}(t) \;=\; r_m(t) - r_w(t) \;=\; \Delta\beta \frac{S(t)}{N(t)} - \Delta\alpha - \Delta\gamma. \tag{1.16}$$

The temporal dynamics of the frequency of the variant is thus given by the genetic variance times the selection gradient (aka the selection coefficient) $\mathcal{S}(t)$ [53, 62, 63], which is the difference in growth rates/absolute fitness between the variant and the wildtype. The genetic variance being always positive, the direction of selection is governed by the sign of $\mathcal{S}(t)$ [53], which depends on the phenotypic differences between the variant and the wildtype and, as soon as $\Delta\beta \neq 0$, on the availability of susceptible hosts. It is then more convenient to track the logit-frequency of the variant instead, that is the log odds ln(frequency of the variant / frequency of the wildtype) [64–67]:

[53]: Day et al. (2020), 'On the evolutionary epidemiology of SARS-CoV-2'
[62]: Day et al. (2007), 'Applying population-genetic models in theoretical evolutionary epidemiology'
[63]: Gandon et al. (2022), 'Targeted vaccination and the speed of SARS-CoV-2 adaptation'

[53]: Day et al. (2020), 'On the evolutionary epidemiology of SARS-CoV-2'

[64]: Chevin (2011), 'On measuring selection in experimental evolution'
[65]: Otto et al. (2021), 'The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic'
[66]: Boyle et al. (2022), 'Selective sweeps in SARS-CoV-2 variant competition'
[67]: Volz (2023), 'Fitness, growth and transmissibility of SARS-CoV-2 genetic variants'

$$\frac{\mathrm{d}\,\mathrm{logit}\big(q(t)\big)}{\mathrm{d}t} = \mathcal{S}(t). \tag{1.17}$$

Thus, $\mathcal{S}(t)$ quantifies the rate of change of the variant frequency on the logit scale and provides a useful metric for the speed of pathogen adaptation [53, 62, 63].

[53]: Day et al. (2020), 'On the evolutionary epidemiology of SARS-CoV-2'
[62]: Day et al. (2007), 'Applying population-genetic models in theoretical evolutionary epidemiology'
[63]: Gandon et al. (2022), 'Targeted vaccination and the speed of SARS-CoV-2 adaptation'

Crucially, the proportion of susceptible hosts differs from the proportion considered in evolutionary invasion analysis, so that the current fitness value differs from the invasion fitness. Especially, both the direction and the magnitude of adaptation change throughout the course of the epidemic [68]. Although the ESS wins in the long-term, another strategy may still outcompete the ESS in the short-term (see example in **Appendix A**). Computing the selection gradient enables to track the direction and strength of selection over time (**Figure 1.8-B**). For example in model (1.10), the selection gradient given in equation (1.16) shows that the threshold proportion of susceptible hosts for which the variant frequency remains constant is:

[68]: Berngruber et al. (2013), 'Evolution of virulence in emerging epidemics'

$$\frac{S(t)}{N(t)} = \frac{\Delta\alpha + \Delta\gamma}{\Delta\beta}.$$

A variant with higher transmission ($\Delta\beta > 0$) and higher rate of leaving the infectious compartment ($\Delta\alpha + \Delta\gamma > 0$) is thus selected for as soon as $S(t)/N(t) > (\Delta\alpha + \Delta\gamma)/\Delta\beta$, which typically holds at the beginning of the epidemic. At some point during the course of the epidemic however, the fraction of susceptible hosts falls below this threshold and the variant is now selected against.

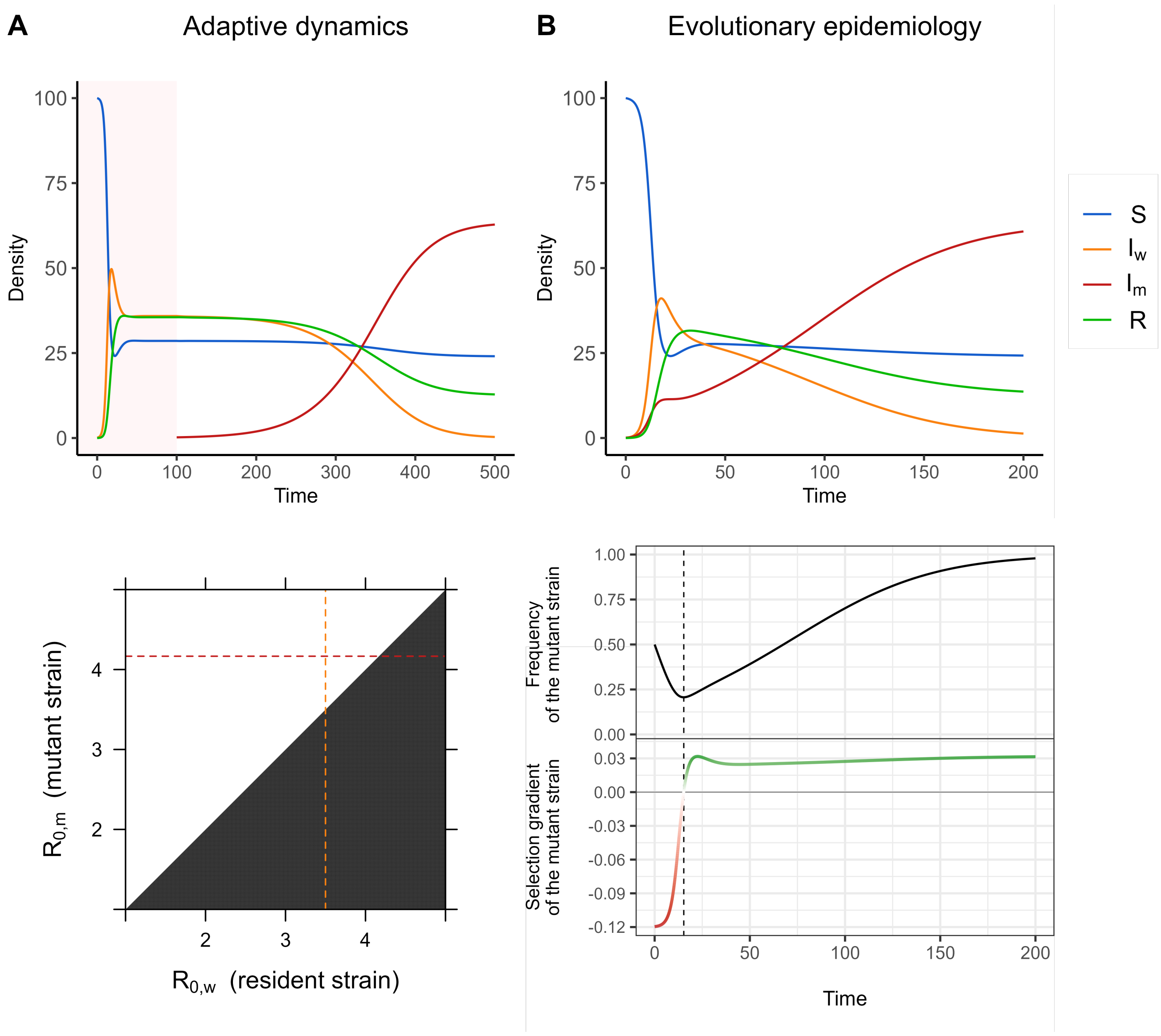


Figure 1.8: Adaptive dynamics vs. evolutionary epidemiology. The polymorphic model (1.10) is simulated with parameter values: $\lambda = 10$, $\delta = 0.1$, $\beta_w = 0.7$, $\beta_m = 0.5$, $\alpha_w = \alpha_m = 0$, $\gamma_w = 0.1$, $\gamma_m = 0.02$ and $\zeta = 10^{-3}$. At $t = 0$, $S(0) = \lambda/\delta = 100$, $I_w(0) = 0.1$ and $R(0) = 0$. (A) Within the framework of adaptive dynamics, mutations are assumed to be rare and a variant only emerges after the previous strain has reached its epidemiological attractor. Top panel: epidemiological dynamics of the wildtype strain before the emergence of the mutant strain (pink background) and after the introduction of a rare variant (white background, low initial density 0.2). Bottom panel: the mutant strain can invade the resident (wildtype) strain because its basic reproduction number is higher: $\mathcal{R}_{0,m} > \mathcal{R}_{0,w}$ (white=invasion, black=no invasion). (B) Within the framework of evolutionary epidemiology, mutations are not necessarily rare and a variant can emerge while the system has not yet reached any epidemiological attractor (I take here $I_m(0) = 0.1$). The long-term outcome of the competition is the same as in A, but evolutionary epidemiology also aims to track the short-term (transient) dynamics of the system. Top panel: epidemiological dynamics. Bottom panel: the selection gradient of the mutant strain is first negative (in red), the mutant strain is counter-selected and decreases in frequency; at some point the logit-frequency of the mutant reaches a minimum (vertical dashed line), its selection gradient being equal to 0 (in white); eventually, the selection gradient of the mutant strain is positive (in green), the mutant strain is selected for and increases in frequency.

Weak selection assumption and separation of timescale Under the assumption of weak selection, phenotypic differences between the variant and the wildtype ($\Delta\beta$, $\Delta\alpha$ and $\Delta\gamma$) are assumed to be small and of order ε (denoted $\mathcal{O}(\varepsilon)$, with $\varepsilon \ll 1$). Thus, the dynamics of the (logit-)frequency of the variant in equations (1.15)-(1.17) is also $\mathcal{O}(\varepsilon)$. Rewriting the temporal dynamics of infected hosts in (1.10) by expanding the mean traits values gives:

Notation reminder

- q: frequency of the variant
- β_w, β_m: transmission rates
- $\Delta\beta$: difference in transmission ($\Delta\beta = \beta_m - \beta_w$)
- γ_w, γ_m: recovery rates
- $\Delta\gamma$: difference in recovery ($\Delta\gamma = \gamma_m - \gamma_w$)
- α_w, α_m: virulence
- $\Delta\alpha$: difference in virulence ($\Delta\alpha = \alpha_m - \alpha_w$)
- r_w, r_m: growth rates of the epidemic (absolute pathogen fitness)
- $\mathcal{R}_{0,w}$, $\mathcal{R}_{0,m}$: basic reproduction numbers

$$\begin{aligned}\dot{I}(t) &= \left(\beta_w \frac{S(t)}{N(t)} - \delta - \alpha_w - \gamma_w\right) I(t) + q(t)\left(\Delta\beta \frac{S(t)}{N(t)} - \Delta\alpha - \Delta\gamma\right) I(t) \\ &= \underbrace{\left(\beta_w \frac{S(t)}{N(t)} - \delta - \alpha_w - \gamma_w\right) I(t)}_{\mathcal{O}(1)} + \mathcal{O}(\varepsilon)\,. \end{aligned} \tag{1.18}$$

Thus, epidemiological dynamics are much faster than evolutionary dynamics. A separation of timescale argument (see **Box 1.2.1**) can therefore be used to reduce the number of ODEs of the system. In particular, I focus on the slow dynamics (1.17) (evolution) whose analysis is simplified by assuming that the fast dynamics (1.18) (epidemiology) reaches instantaneously its quasi-equilibrium. Hence, setting the right-hand side of (1.18) to 0 yields when $I(t) \neq 0$:

$$\frac{S(t)}{N(t)} = \frac{\delta + \alpha_w + \gamma_w}{\beta_w} + \mathcal{O}(\varepsilon) = \frac{1}{\mathcal{R}_{0,w}} + \mathcal{O}(\varepsilon)\,.$$

Substituting $S(t)/N(t)$ in (1.17) by its quasi-equilibrium value gives the simplified dynamics:

$$\frac{\mathrm{d}\,\mathrm{logit}\big(q(t)\big)}{\mathrm{d}t} = \frac{\Delta\beta}{\mathcal{R}_{0,w}} - \Delta\alpha - \Delta\gamma + \mathcal{O}\left(\varepsilon^2\right).$$

The variant increases in frequency and wins the competition if $\mathrm{d}\,\mathrm{logit}\big(q(t)\big)/\mathrm{d}t > 0$. Neglecting the term $\mathcal{O}\left(\varepsilon^2\right)$, this is equivalent after some rearrangements to:

$$\mathcal{R}_{0,m} > \mathcal{R}_{0,w}.$$

I thus recover the analytical result of adaptive dynamics. The difference is that, here, selection is assumed to be weak but mutations are not assumed to be rare while in adaptive dynamics mutations are assumed to be rare but with no assumption on the size of the phenotypic differences.

Box 1.2.1 ***Slow-fast dynamics and separation of time scale***

A separation of timescale argument is a useful tool enabling to study fast and slow processes separately (singular perturbation theory, as proposed in Tikhonov's theorem) [69–71]. Let x (fast) and y (slow) be two real variables such that:

$$\begin{cases} \dfrac{dx}{dt} &= f(x,\ y) \\ \dfrac{dy}{dt} &= \varepsilon g(x,\ y) \end{cases}$$

with f and g, two continuous functions, and $\varepsilon \ll 1$, a very small positive parameter emphasizing that y has a much slower dynamics than x [72, 73].

[69]: Tikhonov (1952), 'Systems of differential equations containing small parameters in the derivatives. [In Russian]'
[70]: Rinaldi et al. (2000), 'Geometric analysis of ecological models with slow and fast processes'
[71]: Verhulst (2007), 'Singular perturbation methods for slow–fast dynamics'

[72]: Gjini et al. (2017), 'A slow-fast dynamic decomposition links neutral and non-neutral coexistence in interacting multi-strain pathogens'
[73]: Jardón-Kojakhmetov et al. (2021), 'A geometric analysis of the SIR, SIRS and SIRWS epidemiological models'

The analysis of the fast variable is simplified by neglecting the dynamics of the slow variable. Taking $\varepsilon = 0$*, the slow variable* y *becomes considered fixed to a constant value, which yields the critical fast dynamics:*

$$\begin{cases} \dfrac{dx}{dt} &= f(x,\ y) \\ \dfrac{dy}{dt} &= 0 \end{cases}$$

Conversely, the analysis of the slow variable is simplified by assuming that the fast variable reaches instantaneously a (quasi-)equilibrium value. Using the change of timescale $\tau = \varepsilon t$*, an equivalent alternative system is:*

$$\begin{cases} \varepsilon\dfrac{dx}{d\tau} &= f(x,\ y) \\ \dfrac{dy}{d\tau} &= g(x,\ y) \end{cases}$$

Taking $\varepsilon = 0$ *yields the critical slow subsystem:*

$$\begin{cases} 0 &= f(x,\ y) \\ \dfrac{dy}{d\tau} &= g(x,\ y) \end{cases}$$

Suppose that $0 = f(x,\ y)$ *is solved by the* $x = \phi(y)$*, with* ϕ*, a continuous function satisfying* $x(t) \longrightarrow \phi(y)$ *as* $t \longrightarrow +\infty$ *and* $f\left(\phi(y),\ y\right) = 0$ *((quasi-)equilibrium);* $\left(\phi(y),\ y\right)$ *is classically known as the 'slow manifold' [71–73]. One then only needs to focus on the reduced problem:*

$$\frac{dy}{d\tau} = g\left(\phi(y),\ y\right)$$

[71]: Verhulst (2007), 'Singular perturbation methods for slow–fast dynamics'
[72]: Gjini et al. (2017), 'A slow-fast dynamic decomposition links neutral and non-neutral coexistence in interacting multi-strain pathogens'
[73]: Jardón-Kojakhmetov et al. (2021), 'A geometric analysis of the SIR, SIRS and SIRWS epidemiological models'

1.2.3 Host structure

In the SIRS ODE systems (1.1) and (1.10), the pathogen habitat – namely the infected compartment I – is not structured. Yet, host structure is a key feature of host-pathogen systems and have different ecological and evolutionary impacts on the population dynamics.

Introducing host structure can for instance be used to deal with more complex or realistic disease natural histories. For example, exposed hosts are hosts that are infected but not yet infectious. Using compartmental epidemic models, the infected compartment is thus now divided between two chronologically successive compartments: (i) the exposed compartment E and then (ii) the infectious compartment I, which allows to take the potential latent period into account – i.e., the lag between infection and the onset of contagiousness – as in classical SEIR models. More sophisticated models may include additional stages to describe a particular disease natural history, such as asymptomatic or pre-symptomatic compartment (e.g., COVID-19 [53]). Besides, a fraction of the hosts may also be resistant to the pathogen, e.g., vaccination campaigns and adaptive immunity structure the host population between naive and primed hosts [63, 74]. In microbiology, life cycles are often more original and complex. For example, bacteriophages (or phages) are viruses that infect bacteria. Upon bacterial infection, phages lyse (kill) their host to release free viral particles in the environment where they infect new susceptible

[53]: Day et al. (2020), 'On the evolutionary epidemiology of SARS-CoV-2'

[63]: Gandon et al. (2022), 'Targeted vaccination and the speed of SARS-CoV-2 adaptation'
[74]: Day et al. (2022), 'Pathogen evolution during vaccination campaigns'

cells (horizontal transmissions). Alternatively, temperate phages can also integrate themselves into the bacterial genome (lysogeny) and be hereditarily transmitted to daughter cells through cellular division (vertical transmission) [75]. More broadly, pathogens may also be able to infect different types of host [76, 77], potentially from different species. Hosts may also be stratified by age or developmental stages, or structured in spatially separated patches interconnected by migration [59, 78]. In such structured models, the $\mathcal{R}_0$ computation is less straightforward but is achieved through the next-generation theorem [79].

[75]: Gandon (2016), 'Why be temperate: lessons from bacteriophage λ'

[76]: Gandon (2004), 'Evolution of multi-host parasites'

[77]: Regoes et al. (2000), 'Evolution of virulence in a heterogeneous host population'

[59]: Day et al. (2006), 'Insights from Price's equation into evolutionary epidemiology'

[78]: Lion (2018), 'Class structure, demography, and selection: reproductive-value weighting in nonequilibrium, polymorphic populations'

[79]: Diekmann et al. (2010), 'The construction of next-generation matrices for compartmental epidemic models'

1.3 Statistical inference

In the previous section, I have shown how deterministic models of host-pathogen systems can be constructed based on non-linear ODEs and, especially, how this theoretical framework provides useful insights on the evolutionary epidemiology of infectious diseases. Given a set of parameter values and initial conditions, it is easy to simulate the corresponding epidemiological and evolutionary dynamics over time. However, from a model tailored to a particular host-pathogen life cycle, one may also be interested in getting biologically relevant ranges of parameter values, or in comparing different (usually nested) models to determine which one would be the most appropriate. Hopefully, when data are available, it is possible under a range of assumptions to fit models to data in order to accurately estimate model parameters. Such parameters are for instance key life-history traits whose estimation allow to characterize variant phenotypically. In the following, I first present some typical time series data in evolutionary epidemiology. Second, I briefly present the frequentist (and Bayesian) approach to estimate model parameters from data. Third, I address the problem of parameter identifiability. And eventually, using a frequentist approach, I show with some examples some statistical methods and tools I used to fit (non-)linear models.

1.3.1 Time series in evolutionary epidemiology

Most work in mathematical epidemiology focuses on the analysis of time series of infected individuals to understand the demographic dynamics of infectious diseases. For a given infectious disease, traditional demographic data are for instance the incidence – i.e., number of new cases per unit time –, the prevalence – i.e., the proportion of cases in the population at a specific time – or the number of deaths or hospitalizations per unit time. Such data may be stratified by age, sex, vaccination status, geographical locations, etc... In most cases, absolute densities are difficult to assess directly and epidemiological data can be biased. Typically in public health, only a fraction of the population is actually tested and symptomatic individuals are more likely to be tested than the others. I show in **Table 1.2**, an example of a small part of a publicly available dataset that deals with the time series of daily new COVID-19 tests and daily new confirmed cases in the UK.

While demographic data have been used for a very long time, genetic data are much more recent. Novel methodological advances in molecular

Table 1.2: Example of time series of tests and new confirmed cases per day. Daily number of COVID-19 tests with new cases tested positive in the UK (from 07/04/2020), downloaded from the website Our World in Data.

Entity	Code	Day	new_tests_ 7day_smoothed	annotations	Daily new cases due to COVID-19 (rolling 7-day average, right-aligned)
United Kingdom	GBR	2020-04-07	17876	tests performed	4116.857
United Kingdom	GBR	2020-04-08	18521	tests performed	4116.857
United Kingdom	GBR	2020-04-09	19013	tests performed	4116.857
United Kingdom	GBR	2020-04-10	15713	tests performed	4116.857
United Kingdom	GBR	2020-04-11	15963	tests performed	4116.857
United Kingdom	GBR	2020-04-12	16216	tests performed	4661.429
...	...	...	...	...	...

biology from the end of the 20th century are revolutionizing this field of research, as genetic and genomic data become available. In particular, high-throughput sequencing methods are increasingly used to determine pathogen genotypes from collected samples and, very recently, the COVID-19 pandemic have led to an unprecedented sequencing effort along with rapid sharing of sequences. This allows notably to monitor the temporal dynamics of strain frequencies and thus to track pathogen evolution – for SARS-CoV-2, see for example CoVariants [80] or Nextsrain [81], with data from GISAID [21]. I show in **Table 1.3** an example of a small part of a publicly available dataset (from the technical briefing 5 of Public Health England [19]) that deals with the time series of the regional weekly number and percentage of the detection of SARS-CoV-2 Alpha variant among cases tested positive in England (the whole dataset is plotted in **Figure 1.9**).

[80]: Hodcroft (2021), *CoVariants: SARS-CoV-2 Mutations and Variants of Interest.*

[81]: Hadfield et al. (2018), 'Nextstrain: real-time tracking of pathogen evolution'

[21]: Khare et al. (2021), 'GISAID's role in pandemic response'

[19]: Public Health England (2020), *Investigation of novel SARS-COV-2 variant 202012/01: technical briefing 5*

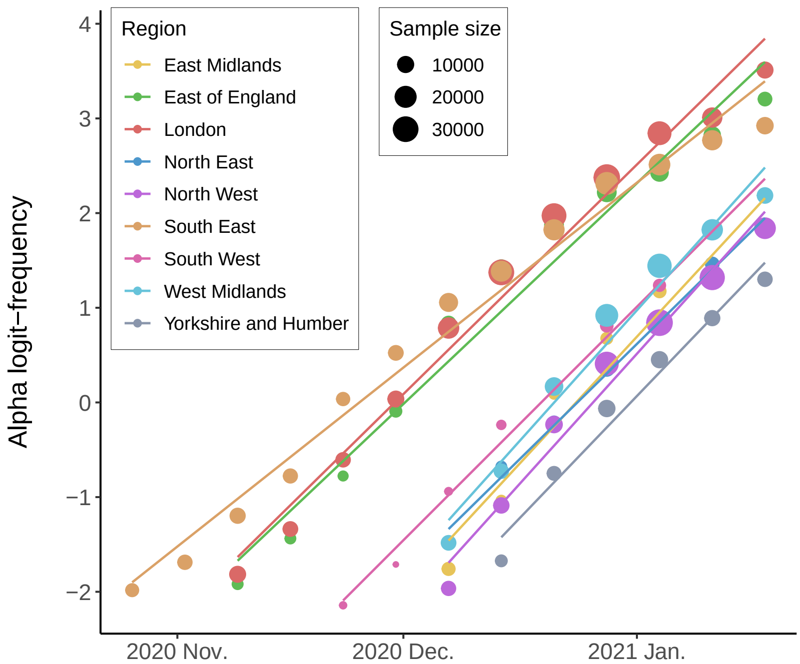


Figure 1.9: Rise of the SARS-CoV-2 Alpha variant in England late 2020 early 2021. Temporal dynamics of the weekly logit-frequency of the Alpha variant during its sweep across the nine regions of England. I use publicly available time series data from the technical briefing 5 of Public Health England where qPCR positive results with S gene target failure (SGTF) is used as a proxy for the Alpha variant.

Table 1.3: Example of time series of the proportion of a variant detection. Weekly regional number and percentage of Pillar 2 COVID-19 cases in England tested by TaqPath laboratories (download dataset). For this period of time, S gene target failure (SGTF) is used as a proxy for the SARS-CoV-2 Alpha variant while S gene detection (S-gene) refers to the previous lineage. This dataset comes from the technical briefing 5 of Public Health England.

Region week	week	n_Confirmed S-gene	n_Confirmed SGTF	percent_Confirmed S-gene	percent_Confirmed SGTF	n_Total
East Midlands	07/09/2020	635	7	98.9	1.1	642
East Midlands	14/09/2020	720	3	99.6	0.4	723
East Midlands	21/09/2020	964	17	98.3	1.7	981
East Midlands	28/09/2020	1685	15	99.1	0.9	1700
East Midlands	05/10/2020	2895	49	98.3	1.7	2944
...	...	...	...	...	...	...

1.3.2 Frequentist (and Bayesian) approach

In a statistical population, parameters of interest are typically considered as unknown quantities. Most often, data are only collected from a statistical sample, that is a (hopefully representative) subset of the population. Data can then be used to fit statistical models in order to infer the parameters of the population. From one study to another, the collected data vary in quantity and/or quality; in particular, data can be incomplete, biased or inaccurate owing to observational errors. Inferential statistics allow to estimate model parameters but also to quantify the uncertainty of these estimates. Two paradigms are classically presented: frequentist (rooted in frequentist probability) and Bayesian (based on Bayes' theorem). In this thesis, I (almost always) use a frequentist approach, so I just say a few words about the Bayesian approach at the end.

Frequentist approach Let $Y_{1:n}$ be a sequence of n independent random variables with assumed probability (mass or density) function f_Y (statistical model) and let $y_{1:n}$ be a realization of $Y_{1:n}$ (data). I also denote $\Theta \subset \mathbb{R}^p$, the space of the p parameters of interest, $\theta \in \Theta$, a vector of size p of these parameters, and σ, a vector of nuisance parameters. In model-based statistical inference, the likelihood is a key element. Under a statistical model, the likelihood function $\mathscr{L}$ is the (density) probability to observe the data $y_{1:n}$ conditionally on parameter values θ and σ:

$$\mathscr{L}\left(y_{1:n} \mid \theta, \sigma\right) = f_Y\left(y_{1:n}, \theta, \sigma\right).$$

Statistical inference based on a frequentist approach often relies on the method of maximum likelihood estimation (MLE), introduced by Fisher [82, 83]. Under the assumptions of a statistical model, MLE estimates, denoted $\hat{\theta}$, are expected to be the most likely point values of the parameters – i.e., point estimates that maximize the likelihood function – given the collected data. The MLE estimator (also denoted $\hat{\theta}$ for simplicity) of the parameters of interest θ is thus defined as:

[82]: Fisher (1912), 'On an absolute criterion for fitting frequency curves'
[83]: Fisher (1922), 'On the mathematical foundations of theoretical statistics'

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\left(\mathscr{L}\left(y_{1:n} \mid \theta, \sigma\right)\right).$$

It is often more convenient to work with the log-likelihood $\ln(\mathscr{L})$ instead. Assuming that the random variables $Y_{1:n}$ are independent with assumed probability (mass or density) functions f_{Y_i}, respectively, the likelihood function is given by:

$$\mathcal{L}\left(y_{1:n} \mid \theta, \sigma\right) = \prod_{i=1}^{n} f_{Y_i}\left(y_i, \theta, \sigma\right),$$

and the log-likelihood by:

$$\ln\left(\mathcal{L}\left(y_{1:n} \mid \theta, \sigma\right)\right) = \sum_{i=1}^{n} \ln\left(f_{Y_i}\left(y_i, \theta, \sigma\right)\right).$$

In the simplest cases, it may be possible to find analytically the solutions of θ and σ that maximize the log-likelihood by solving the following system of partial differential equations for each $\theta_i \in \theta$ and $\sigma_i \in \sigma$:

$$\begin{cases} \dfrac{\partial \ln\left(\mathcal{L}\left(y_{1:n} \mid \theta, \sigma\right)\right)}{\partial \theta_i} = 0 \\ \dfrac{\partial \ln\left(\mathcal{L}\left(y_{1:n} \mid \theta, \sigma\right)\right)}{\partial \sigma_i} = 0 \end{cases}$$

along with negative second derivatives to ensure that the extremum is a maximum.

Under the assumption of normality for the error terms, let $Y_{1:n}$ be a sequence of independent Gaussian random variables, such that: $\forall i \in [1,\ n]$, $Y_i \sim \mathcal{N}\left(\mu_i(\theta),\ \sigma^2\right)$, where μ_i and σ^2 are the mean and variance of Y_i, respectively. The log-likelihood is thus:

$$\ln\left(\mathcal{L}\left(y_{1:n} \mid \theta, \sigma^2\right)\right) = -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu_i(\theta)\right)^2.$$

For this Gaussian case, maximizing over θ is independent of σ and implies to minimize the sum of squares $\sum_{i=1}^{n}\left(y_i - \mu_i(\theta)\right)^2$, that is the Euclidean norm $\|y - \mu(\theta)\|^2$, as typically performed in the least squares method.

Besides, the (log-)likelihood is also useful to compare different models that have been fitted to the same data when one is interested in selecting the simplest (i.e., most parsimonious) model that sufficiently explains the data. This enables to avoid over-fitting and to keep models as simple as possible to be informative. For nested models, such selection can be tackled using likelihood ratio tests. More generally, one can compare the likelihoods while penalizing each model by its complexity. Among others, the Akaike's Information Criterion (AIC) [84] is for example defined as:

$$\mathrm{AIC} = -2\ln\left(\mathcal{L}\left(y_{1:n} \mid \hat{\theta}, \hat{\sigma}\right)\right) + 2p,$$

with p, the number of independently adjusted parameters. Although it is not a formal hypothesis test, selecting the model with the lowest AIC score is a reasonable approach.

Bayesian approach Unlike the frequentist approach, Bayesian inference does not assume that parameters are fixed unknown values to estimate, but assumes that parameters are random variables (due to the uncertainty about their true value) with unknown probability distribution to estimate.

Notation reminder

- θ: parameter(s) of interest
- $\hat{\theta}$: MLE estimator/estimation of θ
- σ: nuisance parameter(s)
- $\hat{\sigma}$: MLE estimator/estimation of σ
- $\mathcal{L}$: likelihood
- f_Y: probability function for the data
- n: number of data
- p: number of estimated parameters
- $y_{1:n}$: data

[84]: Akaike (1974), 'A new look at the statistical model identification'

In addition to the data, the Bayesian approach allows to account for prior knowledge about the parameters (informative prior). A prior may also be non-informative. As its name suggests, the Bayesian paradigm is based on Bayes' theorem:

Bayes' theorem

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)\mathbb{P}(B \mid A)}{\mathbb{P}(B)}$$

$$\underbrace{\mathbb{P}(\theta, \sigma \mid y_{1:n})}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}(\theta, \sigma)}^{\text{Prior}} \times \overbrace{\mathcal{L}(y_{1:n} \mid \theta, \sigma)}^{\text{Likelihood}}}{\underbrace{\mathbb{P}(y_{1:n})}_{\text{Marginal}}}.$$

The posterior distribution to estimate is calculated from *a priori* information on the parameters (prior) and the likelihood $\mathcal{L}(y_{1:n} \mid \theta)$. The denominator, a (possibly high-dimensional) integral potentially impossible to compute, can be ignored as it constitutes a normalising constant and cancels out for any posterior ratio. Hence: $\mathbb{P}(\theta, \sigma \mid y_{1:n}) \propto \mathbb{P}(\theta, \sigma) \times \mathcal{L}(y_{1:n} \mid \theta, \sigma)$. The posterior distribution is typically sampled using Monte Carlo Markov Chains (MCMC) algorithms, which allows then to obtain summary statistics and credible intervals for the estimated parameters.

1.3.3 Identifiability

It is not always guaranteed that model parameters can be estimated. Indeed, the problem of parameter estimation involves the notion of identifiability and partially observed dynamical systems frequently exhibit non-identifiability issues [85].

[85]: Wieland et al. (2021), 'On structural and practical identifiability'

[85]: Wieland et al. (2021), 'On structural and practical identifiability'
[86]: Cunniffe et al. (2023), 'Identifiability and Observability in Epidemiological Models'
[87]: Cobelli et al. (1980), 'Parameter and structural identifiability concepts and ambiguities: a critical review and analysis'
[88]: Raue et al. (2014), 'Comparison of approaches for parameter identifiability analysis of biological systems'
[89]: Bellu et al. (2007), 'DAISY: A new software tool to test global identifiability of biological and physiological systems'
[90]: Raue et al. (2009), 'Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood'
[9]: Kermack et al. (1927), 'A contribution to the mathematical theory of epidemics'
[91]: Hamelin et al. (2023), 'About the identifiability and observability of the SIR epidemic model with quarantine'
[86]: Cunniffe et al. (2023), 'Identifiability and Observability in Epidemiological Models'
[92]: Tuncer et al. (2018), 'Structural and practical identifiability analysis of outbreak models'

Structural identifiability (see **Box 1.3.1**) deals with the uniqueness of the solutions for the parameters θ, that is the capacity to theoretically recover unknown parameters from ideal (i.e., infinite and noise-free) input-output measurements [85–88]. Structural identifiability is an intrinsic property of the structure of a model. Obviously, two parameters that always appear as a product or a sum in a model can never be separately identifiable as an infinite number of combinations yield identical results (parameters can compensate each other). For example, considering the force of infection $(1 - c)\beta I(t)/N(t)$ with $c \in [0,\ 1]$, the efficacy of control measures, parameters β and c cannot be separately identifiable (although the product may be). Nevertheless, detecting non-identifiable parameters within non-linear systems is usually non-trivial. Formal approaches exist to investigate such structural identifiability – e.g., the differential algebra software DAISY [89]. Note that structural identifiability is an *a priori* problem. It can thus be investigated before conducting experiments and guide the design of experimental planning [90] – e.g., suggesting additional measurements so that the combination of information makes an otherwise non-identifiable parameter identifiable. Identifiability of the classical SIR model proposed by [9] was investigated quite recently (early 2000s) and the COVID-19 pandemic has led to an increased interest in parameter estimation and identification using compartmental epidemic models [91]. In the classical SIR model, the transmission rate β and the recovery rate γ are identifiable using prevalence observations and knowing the initial conditions $S(0)$, $I(0)$ and $R(0)$ (and thus N) [86, 92]; however, when only an unknown fraction of the infected hosts are reported, the transmission rate of this model is not separately identifiable from the reporting rate [86].

[86]: Cunniffe et al. (2023), 'Identifiability and Observability in Epidemiological Models'

Box 1.3.1 ***Structural identifiability***

Consider the following dynamical system:

$$\begin{cases} \dot{x}(t) & = f\left(x(t), u(t), x_0, \theta\right) \\ y(t) & = h\left(x\left(t, u(t), x_0, \theta\right)\right) \end{cases}$$

where x is the state of the system (process model) with initial condition $x(0) = x_0$ and whose temporal dynamics is described by an ODE governed by function f; u refers to inputs and y to the observations (outputs) of x through the functional mapping h (measurement model), which can increase the dimensionality of θ [88]. Indeed, the process model is typically not observable directly (latent, or "hidden", processes) but the outputs of the measurement model (the observations/data) are expected to indirectly reflect the states of the process model. In most cases with dynamical compartmental models, data are collected from the sampling of a limited number of compartments. By definition, the above system is globally structurally identifiable given an initial state x_0 if and only if:

[88]: Raue et al. (2014), 'Comparison of approaches for parameter identifiability analysis of biological systems'

$$\forall t > 0, \quad h\left(x\left(t, u(t), x_0, \theta\right)\right) = h\left(x\left(t, u(t), x_0, \theta'\right)\right) \implies \theta = \theta'$$

[86] and locally structurally identifiable if there exists a neighborhood of θ $\mathcal{V} \subset \Theta$ such that:

[86]: Cunniffe et al. (2023), 'Identifiability and Observability in Epidemiological Models'

$$\begin{aligned} &\forall t > 0,\ (\theta, \theta') \in \mathcal{V}^2, \\ &h\left(x\left(t, u(t), x_0, \theta\right)\right) = h\left(x\left(t, u(t), x_0, \theta'\right)\right) \implies \theta = \theta'. \end{aligned}$$

A model can also be partially identifiable when only a subset of θ, or some functions of the parameters, can be accurately estimated.

Structural non-identifiability depends thus only on the structure of the model and the empirical observations are assumed to be infinite and noise-free. Yet, this is never the case in practical situations. In that sense, practical non-identifiability also depends on the actual data (quantity and quality) and the numerical optimization algorithm used for the estimation problem [86, 90, 92]. Therefore, a structurally identifiable model (prerequisite) can still be practically non-identifiable due to too poor observations.

[86]: Cunniffe et al. (2023), 'Identifiability and Observability in Epidemiological Models'
[90]: Raue et al. (2009), 'Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood'
[92]: Tuncer et al. (2018), 'Structural and practical identifiability analysis of outbreak models'

A useful approach to investigate both the structural and the practical identifiability is the profile likelihood [90], which can be performed with either real or simulated data. To construct the profile likelihood of a given parameter of interest, one iteratively fixes the value of this parameter in order to cover a relevant range of values while maximizing each time the log-likelihood function over all the other parameters to estimate; one then plots the maximized log-likelihood as a function of the fixed values of the parameter of interest. A profile that peaks at a global maximum likelihood is the signature of an identifiable parameter. In contrast, a flat profile is the signature of a non-identifiable parameter [88, 90]. The curvature of the profile likelihood may also be used to compute CIs [90].

[90]: Raue et al. (2009), 'Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood'
[88]: Raue et al. (2014), 'Comparison of approaches for parameter identifiability analysis of biological systems'
[90]: Raue et al. (2009), 'Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood'
[90]: Raue et al. (2009), 'Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood'

Based on likelihood ratio test (and Wilks' theorem), the two values of a parameter of interest down 1.92 log-likelihood units (half the chi-square value 3.84 with 1 degree of freedom) from the maximum provide an approximate pointwise confidence interval to a confidence level 95%. A flat profile likelihood yields an infinite CI (non-identifiability).

I show below (cf.**§1.3.4–Non-linear optimizations**) some examples of profile likelihood based on simulated time series of the density of infected hosts from the SIRS model (1.1).

1.3.4 Fitting models: examples with SIRS models

In the previous section, I used non-linear systems of ODEs to simulate epidemiological and evolutionary dynamics. Within a frequentist framework (MLE approach), I now seek to fit such models to time series data in order to estimate model parameters of interest. Fitting non-linear dynamical models can be a difficult task. Below, I first present how non-linear optimizations can be used to obtain MLE estimates, which I exemplify based on the previous SIRS model (1.1). Second, variables can sometimes be linearized, so that the estimation problem is tackled using statistical linear models which exhibit convenient properties.

Non-linear optimizations based on the density of infected hosts

Notation reminder

- S: susceptible hosts
- I: infected/infectious hosts
- R: recovered hosts
- β: transmission rate
- γ: recovery rate
- α: virulence
- λ: influx of S
- δ: natural mortality rate
- ζ: rate of immunity waning
- $y_{1:n}$: time series data
- σ: SD of observation errors

I consider a first case based on the monomorphic SIRS model (1.1). Taking some parameter values ($\lambda = 10$, $\delta = 0.1$, $\beta = 0.7$, $\alpha = 0$, $\gamma = 0.09$ and $\zeta = 10^{-3}$) and initial conditions ($S(0) = 100$, $I(0) = 10^{-2}$ and $R(0) = 0$), I run a simulation from which I generate simulated data $y_{1:n}$ corresponding to the time series of infected individuals such that:

$$y_{1:n}(t) \overset{\text{i.i.d.}}{\sim} \text{Log-}\mathcal{N}\left(\ln(I(t)),\, \sigma^2\right),$$

where Log-$\mathcal{N}(\mu, \sigma^2)$ refers to the log-Normal distribution with mean μ and standard deviation (SD) σ on the log scale (here I take $\sigma = 0.1$, **Figure 1.10-A**). I assume now that only the values of the parameters β and γ are unknown and that I want to estimate them. The parameter space is only two-dimensional, all the other parameters and all initial conditions being fixed to their true values. With just two parameters to estimate simultaneously, the likelihood surface can be directly visualized by evaluating the likelihood at a grid of points within relevant ranges of values for β and γ (**Figure 1.10-B**). MLE estimates are given by the parameter values associated with the highest point of the likelihood surface.

As the dimensionality of the parameter space Θ increases, the likelihood is most conveniently explored using non-linear optimization algorithms to find the maximum. Examples of such non-linear optimization algorithms include the Nelder-Mead (aka downhill simplex) method [93], quasi-Newton methods (e.g., Broyden-Fletcher-Goldfarb-Shanno) or simulated annealing. Although optimization algorithms are powerful tools, they may as well bring some numerical difficulties. Optimization algorithms typically start from a given set of arbitrary initial parameter values and iteratively navigate in the parameter space while evaluating a function to

[93]: Nelder et al. (1965), 'A simplex method for function minimization'

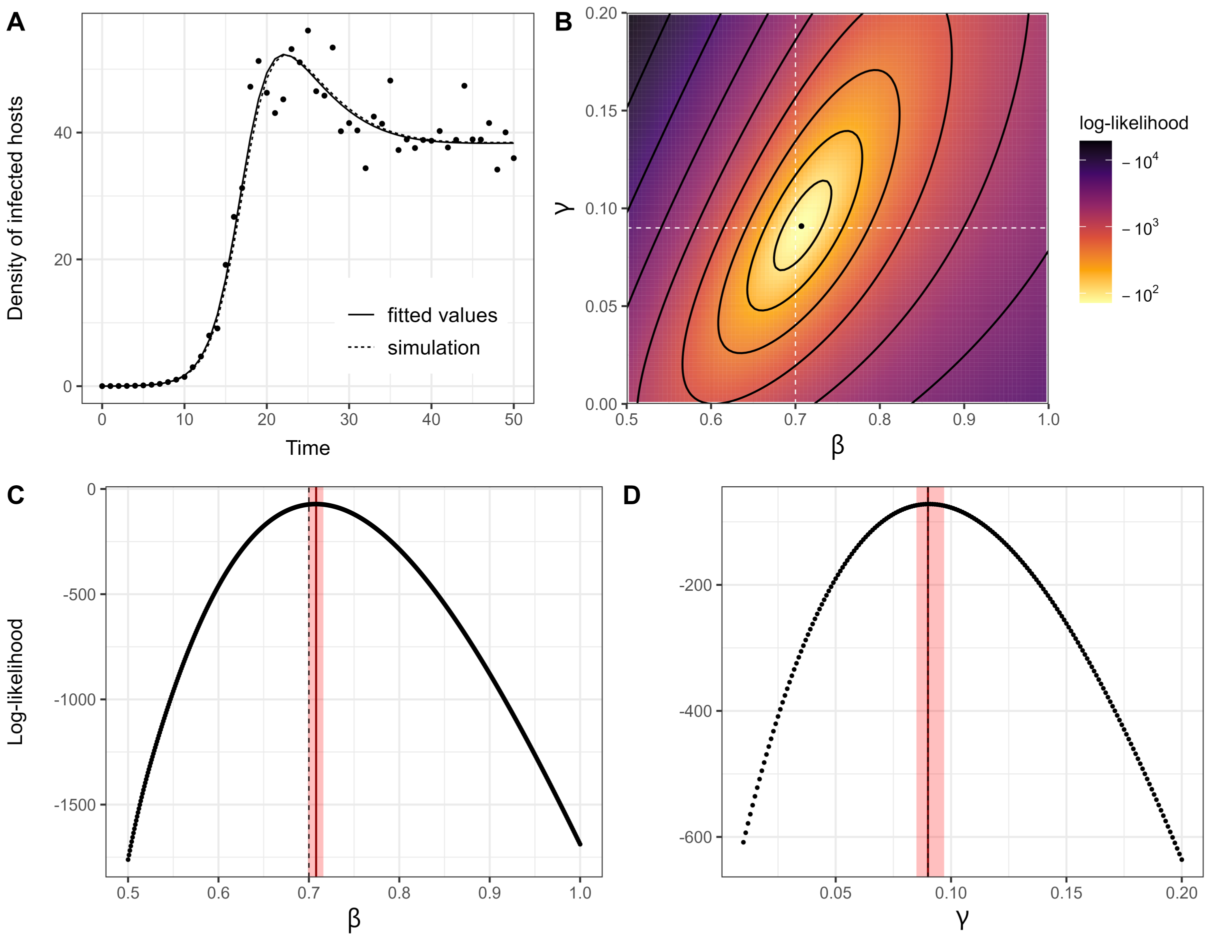


Figure 1.10: MLE of two parameters in an SIRS model from the time series of infected individuals. Model (1.1) is simulated with parameter values: $\lambda = 10$, $\delta = 0.1$, $\beta = 0.7$, $\gamma = 0.09$, $\alpha = 0$ and $\zeta = 0.001$; and initial conditions: $S(0) = \lambda/\delta = 100$, $I(0) = 0.01$ and $R(0) = 0$. I seek to estimate parameters β and γ (all the other parameters are fixed to their true values as well as all the initial conditions). (A) Simulated dynamics of infected hosts (solid line), simulated data (points), generated by multiplying $I(t)$ with an i.i.d. log-normal noise (mean 0 and SD 0.1 on the log scale) to mimic measurement errors, and fitted values (dashed line) based on MLE estimates ($\hat{\beta} = 0.707$, $\hat{\gamma} = 0.091$). (B) Contour plot of the log-likelihood surface as a function of parameters β and γ; the white dashed lines indicate true values and the point indicates MLE estimates (which maximize the log-likelihood). (C) Profile likelihood of parameter β. (D) Profile likelihood of parameter γ. For each point in C and D, the log-likelihood is maximized over the other parameter for 100 uniformly drawn starting points using the R function `optim` (maximum number of iterations 2000, absolute and relative tolerance 10^{-6}); the vertical dashed lines indicate true parameter values and the solid vertical lines (dark red) the best MLE estimates; the red shaded areas refer to parameter values within 1.92 units from the maximum log-likelihood (approximate 95% CI).

be optimized (and/or its gradient) until convergence. Most often, due to the presence of local maxima, where the algorithm can be stuck, likelihood maximization should be repeated from a larger or smaller number of starting points to ensure convergence to a global maximum. Furthermore, imposing parameter bounds through parameter transformations enables to reduce the exploration space ($\mathbb{R}^p$, by default) to the relevant parameter space ($\Theta \subset \mathbb{R}^p$). In particular, biological parameters are typically positive real values and probabilities should lie between 0 and 1. Based on prior knowledge of the biology of the system, it is also often possible to refine parameter bounds even further.

In our previous example, the whole density of infected hosts was measured. Using non-linear optimizations (here, the Nelder-Mead method performed by the R function `optim` from the basic package `stats`), I compute the profile likelihood of parameters β and γ (**Figure 1.10-C** and **D**), which clearly shows that under these conditions these two parameters are indeed identifiable. However, in most cases, only a fraction of the infected hosts is reported. For the sake of simplicity, I assume that the reporting rate ρ of the I compartment is constant over time such that:

Notation reminder

- $y_{1:n}$: data (density of infected hosts)
- I: infected/infectious hosts
- β: transmission rate
- γ: recovery rate
- ρ: reporting rate of infected hosts
- σ: SD of observation errors

$$y_{1:n}(t) \overset{\text{i.i.d.}}{\sim} \text{Log-}\mathcal{N}\left(\ln(\rho I(t)), \sigma^2\right).$$

With the same values of parameters and initial conditions as before, and taking $\rho = 0.05$, I now want to estimate parameters β, γ, ρ and σ. I show in **Figure 1.11** optimization results (Nelder-Mead alhorithm) from 1000 sets of uniformly drawn starting points. I then compute the profile likelihood of each of these four parameters under two scenarios: (1) $\rho = 0.05$ and (ii) $\rho = 0.2$. I show in **Figure 1.12** that all four parameters are identifiable under these conditions. But note that computing profile likelihood is a bit more difficult numerically when the reporting rate is low; in particular the recovery rate γ is estimated with greater uncertainty when $\rho = 0.05$.

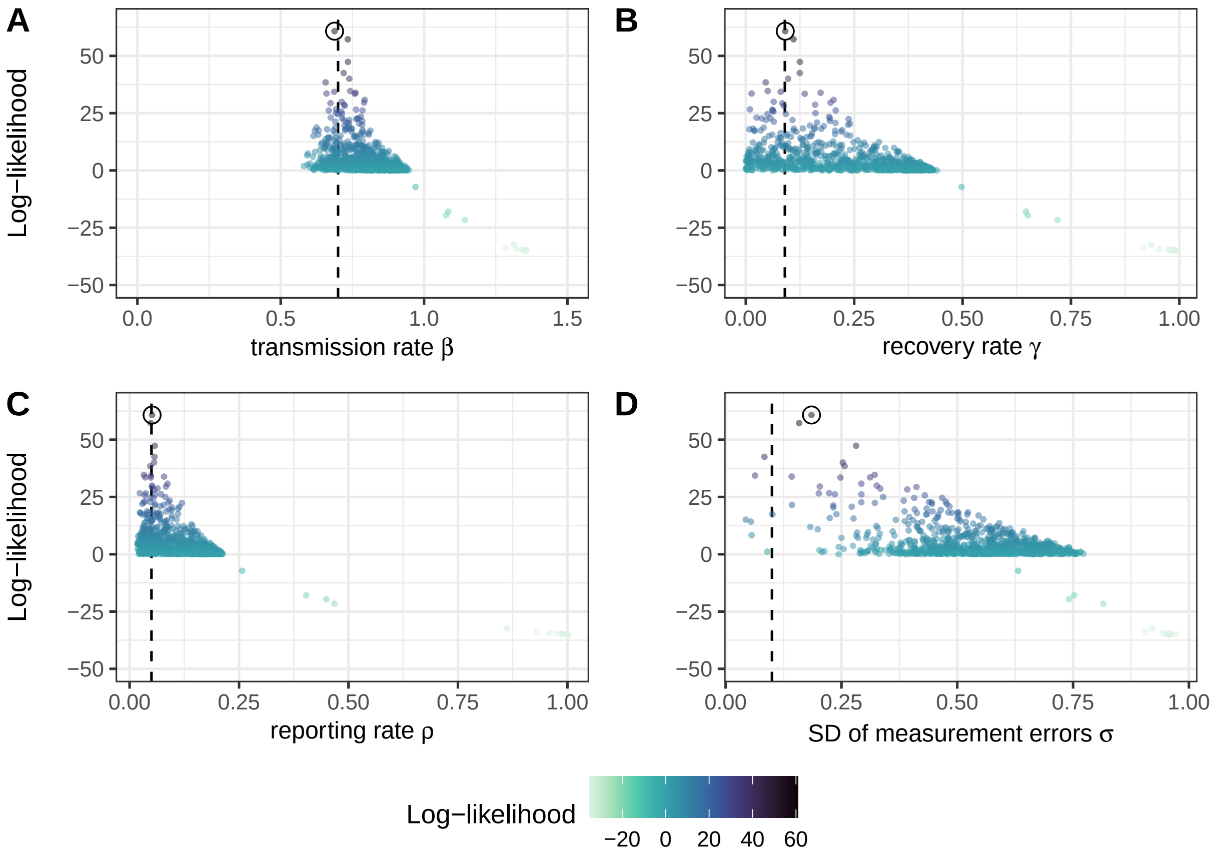


Figure 1.11: Results of non-linear optimizations to obtain MLE estimates of four parameters of an SIRS model from the time series of infected individuals. I simulate model (1.1) (see **Figure 1.10** for parameter values and initial conditions) and generate simulated data by multiplying the time series of the density of infectious hosts by the reporting rate ρ ($\rho = 0.05$) and by an i.i.d. log-normal noise (with mean 0 and SD $\sigma = 0.1$ on the log scale). Non-linear optimizations are run starting from 1000 uniformly drawn sets of initial parameter values to find the best MLE estimates of (A) the transmission rate β, (B) the recovery rate γ, (C) the reporting rate ρ and (D) the SD of measurement errors σ. All other parameters and initial conditions are fixed to their true value. Non-linear optimizations are tackled using the R function `optim` (Neder-Mead algorithm, maximum number of iterations 2000, relative and absolute tolerance 10^{-8}). The vertical dashed lines indicate true parameter values. Points associated with the highest log-likelihood are enclosed in a circle.

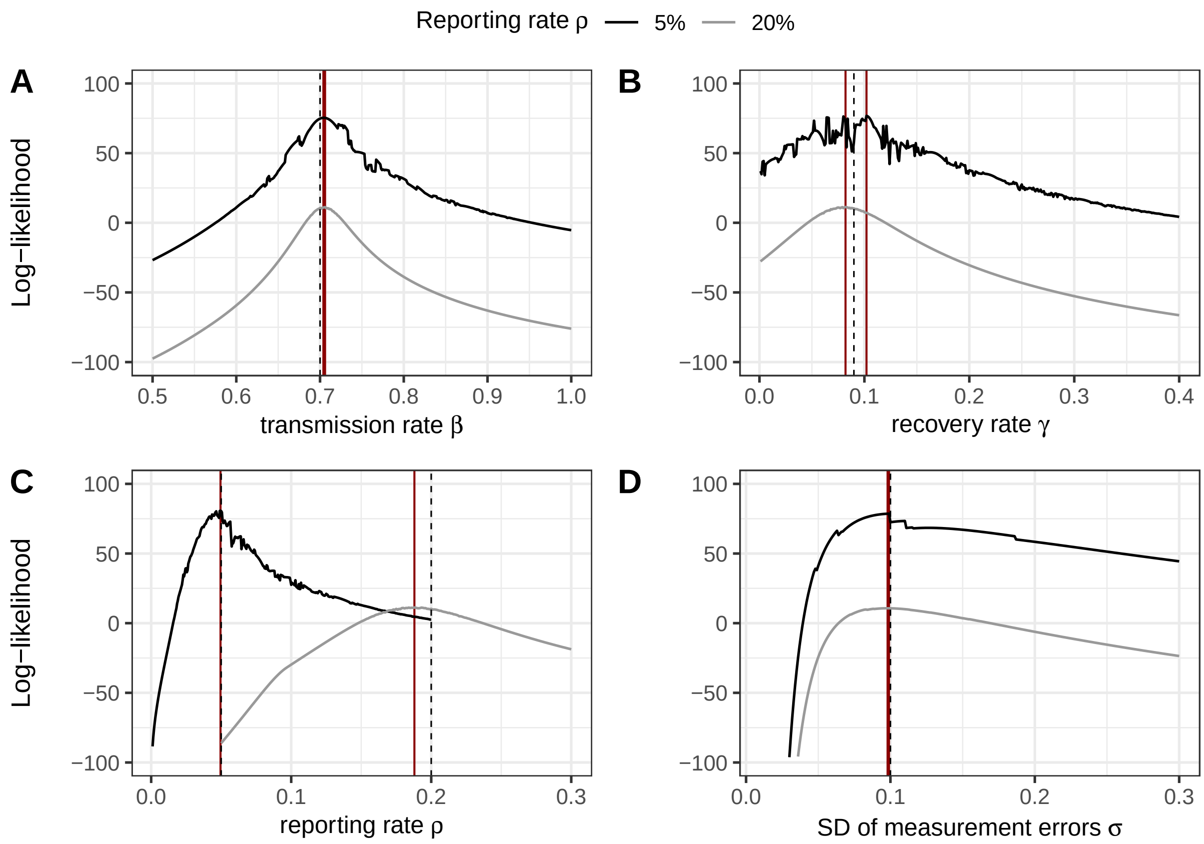


Figure 1.12: Profile likelihood for four parameters of an SIRS model from the time series of infected individuals. I simulate model (1.1) (see **Figure 1.10** for parameter values and initial conditions) and generate simulated data by multiplying the time series of the density of infectious hosts by the reporting rate ρ ($\rho = 0.05$ or $\rho = 0.2$) and by an i.i.d. log-normal noise (with mean 0 and SD $\sigma = 0.1$ on the log scale). Non-linear optimizations are run to construct the profile likelihood of four parameters: (A) the transmission rate β, (B) the recovery rate γ, (C) the reporting rate ρ and (D) the SD of measurement errors σ. In each profile likelihood, for each fixed value of the parameter of interest, the log-likelihood is maximized over the other parameters to estimate starting from 400 uniformly drawn starting points. Non-linear optimizations are tackled using the R function `optim` (Nelder-Mead algorithm, maximum number of iterations 2000, relative and absolute tolerance 10^{-8}). The vertical dashed lines indicate true parameter values and the solid vertical lines (dark red) the best MLE estimates.

Linear models based on the variant logit-frequency

Linear models – and its extensions – are much easier to deal with than non-linear models. Crucially, analytical solutions to the problem of likelihood maximization exist and there is thus no need to explore the likelihood surface. For example, fitting to the data the classical Gaussian linear model:

$$\underbrace{Y_{1:n}}_{\substack{\text{Response}\\\text{variable}}} = X\theta + E,$$

with X the incidence (or design) matrix and $E \sim \mathcal{N}_n\left(0,\ \sigma^2 I_n\right)$, the vector of residuals (with assumptions of normality, independence and homoscedasticity), only requires to solve the normal equations: $\hat{\theta} = \left(X^\top X\right)^{-1} X^\top Y_{1:n}$. This is the classical procedure for (multiple) linear regression and AN(C)OVA (analysis of (co)variance). Extensions of the (Gaussian) linear model include, *inter alia*, linear mixed-effects models, dealing with non-independent data structures, and generalized linear models, dealing with observations whose probability density function (PDF) belongs to an extended family of probability distributions (exponential family, e.g., Binomial, Poisson).

For the polymorphic ODE system (1.10), equation (1.17) shows that the selection gradient $\mathcal{S}(t)$ of a variant is given by the slope of its frequency on the logit scale. Under the assumption that the selection gradient $\mathcal{S}$ is constant over time, integrating (1.17) yields:

Notation reminder

- q: frequency of the variant
- $\mathcal{S}$: selection gradient of the variant
- $\Delta\gamma$: difference in recovery ($\Delta\gamma = \gamma_m - \gamma_w$)

$$\text{logit}(q(t)) = \text{logit}\big(q(0)\big) + \mathcal{S} \times t,$$

so that a simple linear regression enables to estimate the selection gradient $\mathcal{S}$. The logit function has been widely used to linearize the sigmoid-shaped curves of frequencies over time, especially to quantify the strength of selection. As an example, I simulate model (1.10) with parameter values: $\lambda = 10$, $\delta = 0.1$, $\beta_w = \beta_m = 0.7$, $\alpha_w = \alpha_m = 0$, $\gamma_w = 0.09$, $\gamma_m = 9 \times 10^{-3}$ and $\zeta = 10^{-3}$; and initial conditions: $S(0) = 100$, $I_w(0) = 10^{-2}$, $I_m(0) = 10^{-4}$ and $R(0) = 0$ (**Figure 1.13-A**). In this example, the wildtype strain and the variant only differ in terms of recovery rates ($\Delta\beta = \Delta\alpha = 0$ and $\Delta\gamma < 0$). Equation (1.17) reduces then to:

$$\frac{\mathrm{d}\,\text{logit}\big(q(t)\big)}{\mathrm{d}t} = \mathcal{S} = -\Delta\gamma.$$

Assuming that I only have the time series of the strain frequencies, I generate simulated data by adding random i.i.d. noise (standard Normal distribution) to the simulated trajectory of the logit-frequency of the variant. Fitting a Gaussian linear model to the simulated time series data enables to estimate the slope (**Figure 1.13-B**) which yields the following estimation for the phenotypic difference $\Delta\gamma$: -0.075 (95% CI [-0.090, -0.060], true value $\Delta\gamma = -0.081$). If the wildtype strain and the variant also differ in terms of virulence ($\Delta\alpha \neq 0$), $\Delta\gamma$ and $\Delta\alpha$ are not

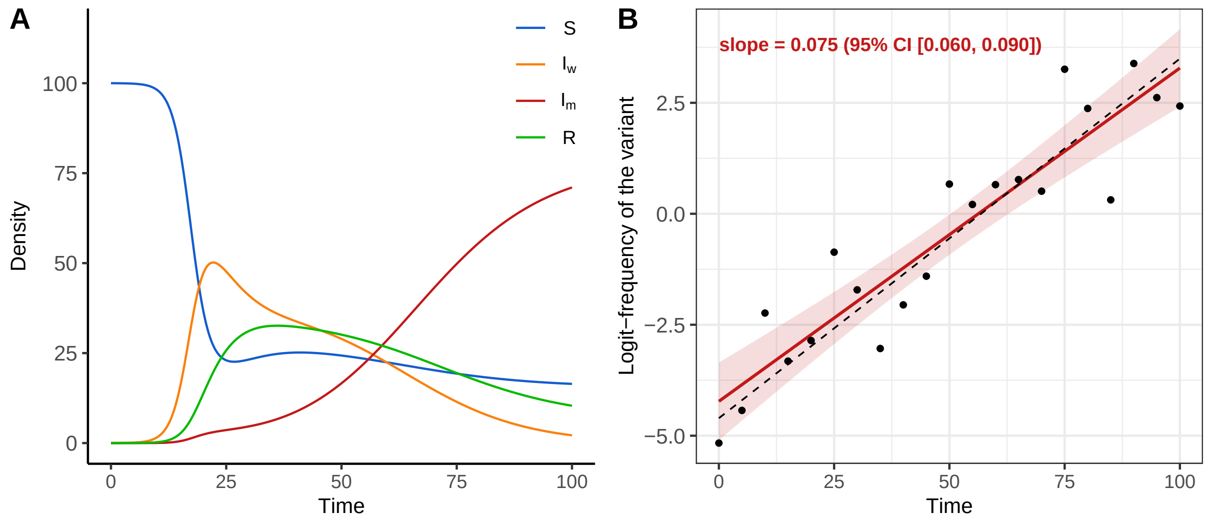


Figure 1.13: Fitting a linear model to estimate the selective advantage of a variant. The polymorphic model (1.10) is simulated with parameter values: $\lambda = 10$, $\delta = 0.1$, $\beta_w = \beta_m = 0.7$, $\alpha_w = \alpha_m = 0$, $\gamma_w = 0.09$, $\gamma_m = \gamma_w/10 = 9 \times 10^{-3}$ and $\zeta = 10^{-3}$; and initial conditions: $S(0) = \lambda/\delta = 100$, $I_w(0) = 10^{-2}$, $I_m(0) = 10^{-4}$ and $R(0) = 0$. Therefore, the phenotypes of the variant and of the wildtype only differ in terms of recovery rates ($\Delta\gamma = -0.081$). The selection gradient of this variant is thus merely: $\mathcal{S} = -\Delta\gamma$ (cf. equations (1.15)-(1.17)). (A) Epidemiological dynamics over time. (B) Logit-frequency of the variant. Based on the original simulation (dashed line), I simulate data (points, $n = 21$) by adding i.i.d. Gaussian noise (with mean 0 and SD 1) on the logit scale to mimic measurement errors. A Gaussian linear regression is fitted to these simulated data – fitted values (dark red solid line) and 95% CI (red shaded envelope) – and thereby estimates the slope, which yields a MLE estimate of $\Delta\gamma$: -0.075 (95% CI $[-0.090, -0.060]$).

structurally separately identifiable based solely on the observations of the logit-frequency of the variant over time (but estimating the slope enables to estimate $-(\Delta\alpha + \Delta\gamma)$). The transmission difference $\Delta\beta$ is more special as it is weighted by the availability of susceptible hosts $S(t)/N(t)$. One can take advantage of this relationship to estimate $\Delta\beta$ but it would require extra information on the S compartment. If the epidemic is slow enough though, the pool of susceptible hosts S can be assumed constant throughout the time period under consideration as a first approximation.

1.3.5 Bootstrapped-based confidence intervals

MLE estimates are *point* estimates and no information is provided on their uncertainty. Quantifying such uncertainty, i.e. constructing confidence intervals (CIs) for each estimated parameter, can be tackled through a parametric approach, which relies asymptotically on the properties of an assumed probability distribution (e.g., Normal, Student, χ^2- distributions). However, it may be difficult to meet the underlying requirements of the assumptions of traditional parametric inferential methods. Bootstrap techniques are useful and powerful non- or semi-parametric methods that rely on resampling schemes – mimicking the sampling process – to estimate the sampling distribution of a statistical quantity of interest [94]. Given the resampled data, the inference of the empirical probability distribution enables to infer the true distribution of the statistical quantity given the original data.

[94]: Efron (1979), 'Bootstrap methods: Another look at the jackknife'

In this thesis, I often use a model-based approach where I resample from the residuals between the original data and the fitted values. The idea of this semi-parametric approach is to fit parametric models first and then resampling from the residuals to generate new bootstrapped data. Refitting the model to resampled data and reiterating this procedure a sufficiently large number of times enables to obtain the joint sampling distributions of the parameters of interest. Classically, residuals are assumed to be identically distributed (homoscedastic). Yet, i.i.d. setups are often violated in practical situations. For non-i.i.d. models, wild bootstrap can be used to capture any pattern of heteroscedasticity in the original error terms [95, 96]. To generate bootstrapped data using wild bootstrap, residuals are randomly perturbed by an i.i.d. sequence of n random weights $\{W_i\}_{i=1}^{n}$ satisfying $\mathbb{E}(W_i) = 0$ and $\mathbb{E}\left(W_i^2\right) = 1$ (e.g., Standard Normal, Mammen's 2-points or Rademacher distribution). Note that the bootstrap sample is thus not a subset of the original sample. Crucially, the new residuals are independent of the data and capture any heteroscedasticity found in the original data. In the context of time series, one can perform sieve bootstrap [97, 98]. This method approximates the actual underlying stationary processes by an autoregressive or moving-average model to construct sieve bootstrap samples. One can use sieve bootstrap to simulate new residuals and, again, generate bootstrapped data. In this thesis, I use both wild and sieve bootstrap.

[95]: Liu (1988), 'Bootstrap procedures under some non-iid models'
[96]: Kline et al. (2012), 'A score based approach to wild bootstrap inference'

[97]: Bühlmann (1997), 'Sieve bootstrap for time series'
[98]: Ulloa et al. (2013), 'Sieve bootstrap prediction intervals for contaminated non-linear processes'

1.4 Objectives of this thesis

Evolutionary epidemiology theory helps to understand the time-varying selection acting on new pathogen variants. In particular, the strength of selection acting on more transmissible and/or more virulent variants is expected to change with the availability of susceptible hosts in the population. Combining epidemiological (demographic) and evolutionary (strain frequencies) data, this theoretical framework is adequate to infer the life-history traits of new variants during epidemics in various situations. The first two research projects (Chapters two and three) stand at the interface between experimental/empirical data and theoretical models; the third project (Chapter four) relies primarily on a theoretical approach.

1.4.1 Objectives of Chapter two (project Alpha)

What is the phenotypic origin of an observed growth advantage? Upon the emergence and sweep of a novel pathogen variant, it is important to understand (i) whether it only arises by chance – i.e., its dynamics are solely driven by stochastic processes – or (ii) whether it is actually more adapted compared to the ancestral lineages and favored through natural selection. Yet, various phenotypic traits may be affected by adaptive mutations and result in similar increased pathogen fitness. For instance, equation (1.5) shows that an increase in the transmission rate or a decrease in the recovery rate (i.e., an increase in the duration of infectiousness) are two mechanistic hypotheses that can both lead to an increase in the basic reproduction number. It can therefore be acknowledged that a specific variant has a selective advantage whilst it still remains unclear which phenotypic traits are involved exactly. In public health, pathogen adaptation undermines our effort to control epidemics and, making matters worse, the unanticipated evolution of life-history traits may lead the pathogen to take advantage of poorly designed control strategies. Hence, tracking pathogen adaptation and identifying the traits responsible for a variant growth advantage is a worthy and important goal, in particular when it comes to optimizing the efficacy of control strategies.

In this chapter, I address this question for the rise of the SARS-CoV-2 variant of concern Alpha that emerged in England in late 2020, about one year after the beginning of the COVID-19 pandemic. I seek to characterize the Alpha variant phenotype and disentangle the origin of its selective advantage considering two phenotypic traits: (i) the transmission rate and (ii) the mean duration of infectiousness. For this purpose, I use the Stringency Index [99], a composite score tailored to measure the extent of non-pharmaceutical interventions (NPIs), i.e., control measures implemented to mitigate the spread of the epidemic. Crucially, NPIs that limit the contact rate between infectious and susceptible hosts unconditionally to infection (e.g., school closure, cancel public events) are predicted to diminish the relative selective advantage of variants with higher transmission rates but not of variants with longer durations of infectiousness. I seek to take advantage of these theoretical contrasting effects to estimate the transmission and recovery components of the selective advantage of the Alpha Variant during its sweep across the

[99]: Hale et al. (2021), 'A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)'

nine regions of England. Using a combination of information from epidemiological data (screening test results, fatality cases) and genetic data (variant frequency among positive cases), I develop a statistical approach based on the analysis of the temporal fluctuations of the selection gradient (here, in an SEIR model) driven by the variability of NPIs stringency.

1.4.2 Objectives of Chapter three (project Lambda)

Alongside public health, characterizing variant phenotypes is also a worthy goal in experimental microbiology, although the life cycles of microorganisms can be much more original and complex. In experimental microbiology, experimental life-history assays are often used to estimate pathogen phenotypes but typically focus on monomorphic pathogen populations. Microbiological systems are well suited for tackling in real life predictions from evolutionary epidemiology theory. While epidemics in animal, including human, populations, such as COVID-19, allow one to confront theory with real biological observations, evolution experiments using microorganisms provide an unparalleled way to conduct population-scale experiments in highly controlled and replicated settings. Evolution experiments are specifically designed to put *a priori* theoretical predictions to the test, whereas epidemics in nature are more used within an *a posteriori* approach. Experimental evolution based on the biology of phage-bacteria systems are particularly useful to test the validity of theoretical predictions. Microbiological systems conditions are more controlled than conditions in human epidemiological studies, and experimental data are often less messy and incomplete. Yet, the experimental validation of theoretical predictions in experimental microbiology is usually limited to qualitative comparisons.

In this chapter, built upon a previous dataset [68], I show how to deepen quantitatively the match between theoretical models and experimental data within a polymorphic pathogen population. This quantitative process enables to better understand the forces acting upon pathogen evolution but also to estimate key phenotypic traits. The previous study was an evolution experiment with the temperate phage λ; two viral strains with distinct life-history strategies (wildtype vs. virulent phage λcI857) were put in competition in continuous cultures of *Escherichia coli* (**Figure 1.14**) and the authors tracked both the epidemiology (prevalence) and the evolution (strain frequency among infected cells and in the culture medium) underlying the viral competition. Based on a qualitative match between numerical simulations and experimental time series, this study confirmed important theoretical predictions on the dynamics of selection on virulence: (i) variants with higher virulence – i.e., variants with larger propensity to lyse bacterial cells and transmitted mostly horizontally – are selected for when susceptible hosts are abundant, but the direction of selection is reversed as soon as the epidemic has reached high prevalence (epidemiological feedback), (ii) starting with a lower prevalence results in a higher increase in virulence during the course of the epidemic and (iii) the virulent strain is always more frequent among free viral particles than among lysogenic cells. In this chapter, I carry out a reverse approach, from experimental evolution back to theory, and seek to perform a quantitative analysis based on a combination of theoretical

[68]: Berngruber et al. (2013), 'Evolution of virulence in emerging epidemics'

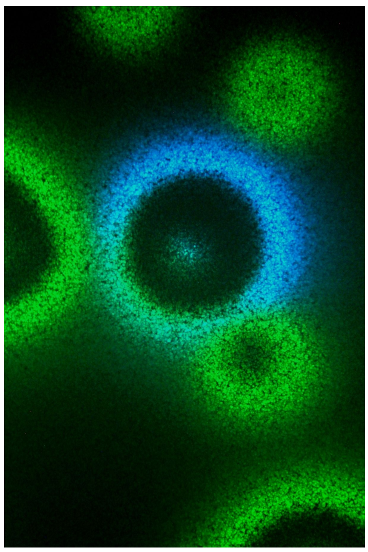


Figure 1.14: Spread of an epidemic of two strains of phage λ in a continuous culture of *E. coli*. The wildtype and the virulent strain λcl857 have different fluorescence color. Photo Thomas W. Berngruber – CEFE – CNRS.

analyses and statistical inference using the same experimental time series data. I thereby illustrate the virtuous interplay between theoretical, experimental and statistical approaches to better understand and forecast the evolutionary epidemiology of infectious diseases.

1.4.3 Objectives of Chapter four (project Omega)

How does migration shape pathogen evolution? In the previous projects, I focused on the evolutionary epidemiology of host-pathogen systems in closed populations. In particular, for the sake of simplicity, I assume in the first project that, during the sweep of the SARS-CoV-2 variant Alpha in England, all regions were independent of each other, without any inter-regional migration flows. Yet, in most cases, natural populations are hardly closed systems and spatially separated populations are interconnected through movements ("migration") of susceptible and infected individuals. Spatial transmission of diseases spread by direct contagion is directly associated with host movements and spatial heterogeneity plays a significant role in epidemiological dynamics. From an *epidemiological perspective*, migration alters the speed of pathogen propagation in host populations, or enables the epidemic to spread on larger scales by allowing pathogens to reach new susceptible hosts from other populations. From an *evolutionary perspective*, migration also interferes with the rise of variants and affect the transient (short-term) and long-term evolutionary dynamics of infectious diseases. In particular, the differential growth of a specific variant in different locations – e.g., the sweep of the SARS-CoV-2 Delta variant in India and in different regions of England [67] – challenges the hypothesis that the dynamics of the variant frequency was solely driven by differences in fitness. Such discrepancy might be explained by other mechanisms, such as host movements between populations. Yet, little is known about how migration shapes the evolutionary dynamics of the pathogen, especially in the short term.

[67]: Volz (2023), 'Fitness, growth and transmissibility of SARS-CoV-2 genetic variants'

In this chapter, I seek to better understand the interplay between migration and selection in pathogen evolution. Based on an SIRS model with a polymorphic pathogen population (competition between two strains), I extend the model within a closed population to a two-patch metapopulation where host populations are interconnected through commuter travel. In particular, I track the transient evolutionary dynamics of the competition between the two strains when the selection is homogeneous or heterogeneous among host populations.

Chapter two

Evolution, 2023, **77(10)**, 2213–2223
https://doi.org/10.1093/evolut/qpad133
Advance access publication July 19, 2023
Original Article

OXFORD

Phenotypic evolution of SARS-CoV-2: a statistical inference approach

Wakinyan Benhamou, Sébastien Lion, Rémi Choquet, Sylvain Gandon

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

Corresponding author: Wakinyan Benhamou, CEFE (UMR 5175), Campus du CNRS, 1919 route de Mende, 34293 Montpellier cedex 5, France.
Email: wakinyan.benhamou@cefe.cnrs.fr

R.C. and S.G. have contributed equally to this work.

Abstract

Since its emergence in late 2019, the SARS-CoV-2 virus has spread globally, causing the ongoing COVID-19 pandemic. In the fall of 2020, the Alpha variant (lineage B.1.1.7) was detected in England and spread rapidly, outcompeting the previous lineage. Yet, very little is known about the underlying modifications of the infection process that can explain this selective advantage. Here, we try to quantify how the Alpha variant differed from its predecessor on two phenotypic traits: The transmission rate and the duration of infectiousness. To this end, we analyzed the joint epidemiological and evolutionary dynamics as a function of the Stringency Index, a measure of the amount of Non-Pharmaceutical Interventions. Assuming that these control measures reduce contact rates and transmission, we developed a two-step approach based on *SEIR* models and the analysis of a combination of epidemiological and evolutionary information. First, we quantify the link between the Stringency Index and the reduction in viral transmission. Second, based on a novel theoretical derivation of the selection gradient in an *SEIR* model, we infer the phenotype of the Alpha variant from its frequency changes. We show that its selective advantage is more likely to result from a higher transmission than from a longer infectious period. Our work illustrates how the analysis of the joint epidemiological and evolutionary dynamics of infectious diseases can help understand the phenotypic evolution driving pathogen adaptation.

Keywords: life-history evolution, adaptation, competition, selection—natural, models/simulations, parasitism

Introduction

In December 2019, acute pneumonias of as yet "*unknown etiology*" were increasingly reported in Wuhan, the capital of the Hubei Province in Central China (Lu et al., 2020). Since then, the infectious agent responsible for this emerging zoonosis, a virus of the family *Coronaviridae* named SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), has spread worldwide, causing the pandemic COVID-19 (coronavirus disease 2019) (World Health Organization, 2020) that is still ongoing today.

The possibility of a rapid SARS-CoV-2 adaptation was initially met with considerable skepticism (Grubaugh et al., 2020b; Rausch et al., 2020). Indeed, compared to other single-stranded RNA viruses, the mutation rate of SARS-CoV-2 is relatively low (estimated at the onset of the pandemic around $6.8–9.8\times10^{-4}$ substitution.site^{-1}.year^{-1} [van Dorp et al., 2020; Vasilarou et al., 2021]). Besides, all the observed mutations in SARS-CoV-2 were initially thought to be neutral or slightly deleterious. The occasional rise of some mutations could be due to demographic stochasticity (Day et al., 2020; Grubaugh et al., 2020a), but the dramatic rise of specific mutations in different regions of the world challenged the hypothesis that none of these mutations were beneficial. In particular, the analysis of the emergence and the spread of several variants of concern (VOCs) across the world—for example, Alpha (lineage B.1.1.7), Delta (lineage B.1.617.2), or Omicron (lineage B.1.1.529) (see, e.g., CoVariants [Hodcroft, 2021] or Nextsrain [Hadfield et al., 2018])—demonstrated that these variants carry adaptive mutations that explain their faster rate of spread in the human population (Obermeyer et al., 2022). However, each of these mutations may act on various dimensions of the fitness landscape of the virus and affect different life-history traits. It is therefore much less clear *why* specific variants are favored. In other words: Which phenotypic trait(s) can explain this increase in viral fitness? Viral fitness is governed by multiple life-history traits like the transmission, the virulence or the recovery rates of the virus (Day et al., 2020). It is crucial to understand which traits are involved in the increase in fitness because they may have very different implications for epidemiological dynamics and public health. For instance, an increase in the transmission rate or in the duration of infectiousness both lead to an increase in viral fitness but they may have distinct consequences for the efficacy of Non-Pharmaceutical Interventions (NPIs), implemented to mitigate the epidemic. It is therefore very important to understand and track this adaptation to optimize our control strategies.

In the following, we will focus on the first of these VOCs: The lineage B.1.1.7, categorized as *Variant of Concern 202012/01* and afterwards named "*Alpha variant*." This variant emerged in early fall 2020 in the South-East region of England (Public Health England, 2020; Volz et al., 2021) and then spread rapidly across the country (Figure 1). The reproduction number of the Alpha variant (i.e., its expected number of secondary infections) was estimated to be 40–100% higher than

Received July 29, 2022; revisions received March 27, 2023; accepted July 05, 2023
Associate Editor: Rees Kassen; Handling Editor: Tracey Chapman

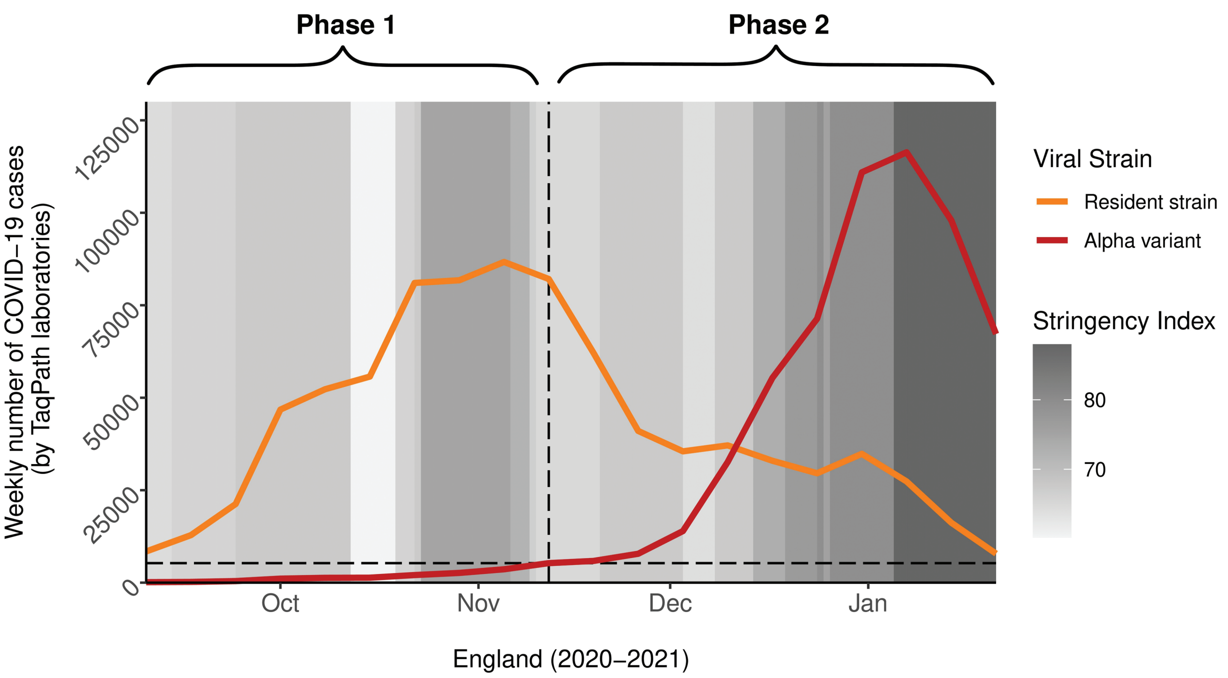


Figure 1. The two consecutive phases of the analysis of the spread of the Alpha variant. In phase 1 (before the emergence of the Alpha variant), we assume the epidemic is driven solely by the resident strain; in phase 2 (after the emergence of the Alpha variant), the epidemic results from the joint dynamics of the resident strain and the Alpha variant. In the first step of our analysis, we estimated the impact of the Stringency Index (a measure of the amount of NPIs implemented to mitigate the epidemic, from 0 [no control] to 100 [stringest control]) on the propagation of the resident strain during phase 1. In the second step of our analysis, knowing the impact of NPIs, we estimated the phenotypic differences between the resident strain and the Alpha variant during phase 2. The dates reported on the chart match the middle of each week (Thursday). We set the end of phase 1 when the Alpha variant reached 5% of the cases tested positive at the national scale (horizontal dashed line). For the sake of simplicity, we show data at the national scale, but the starting date of phase 2 varied among regions (see Supplementary Figure S1 and Methods, A two-step analysis).

for the previous lineage (Davies et al., 2021; Volz et al., 2021). Several studies aimed to unravel what phenotypic differences could explain this increased fitness. First, (Davies et al., 2021) explored various underlying biological mechanisms and suggested that a higher transmission rate per contact for the Alpha variant was the most parsimonious explanation, but that a longer duration of infectiousness—merely increasing the number of opportunities of transmission—could also explain the data very well. (Blanquart et al., 2022) developed another methodological approach considering three phenotypic traits: The overall reproduction number, the mean, and the standard deviation of the generation time distribution of the infection. They showed that the selective advantage of the Alpha variant was likely to be driven by a higher reproduction number with an unaltered mean generation time.

The present work is a new attempt to characterize the life-history traits of the Alpha variant, for which we consider two phenotypic traits: (a) the transmission rate and (b) the recovery rate (inverse of the mean duration of infectiousness). We propose a novel approach to estimate these two phenotypic traits based on the analysis of the time-varying fluctuations of the selection coefficient driven by the variability in the intensity of NPIs used to limit the spread of the virus. As pointed out by (Otto et al., 2021), the selection coefficient of the Alpha variant (i.e., the slope of the change in its logit-frequency) varied with the intensity of NPIs, measured by the "*Stringency Index*," a composite score published by the Oxford COVID-19 Government Response Tracker (OxCGRT) (Hale et al., 2021). In (Day et al., 2020) and (Otto et al., 2021), control measures that reduce contact rates between infectious and susceptible hosts are predicted to reduce the (relative) selective advantage of variants that have a higher transmission rate—in addition to slowing down the spread of the epidemic—but without affecting the selective advantage of variants that have a longer duration of infectiousness. In the following, we exploit these contrasting effects of NPIs on the selection coefficient to infer the transmission rate and the mean duration of infectiousness of the new variant.

We use a stepwise approach of two consecutive phases of the epidemic (Figure 1). First, we focus on the analysis of the epidemiological dynamics taking place just before the emergence of the Alpha variant (i.e., just before it reached 5% of the positive cases) and we infer the relationship between the Stringency Index and the effectiveness of the control measures (NPIs) on the viral propagation in the United Kingdom. Second, we derive a novel expression for the selection coefficient of a variant in a susceptible–exposed–infectious–recovered (*SEIR*) model. Knowing the impact of NPIs on the viral propagation from the first step, we use our expression of the selection coefficient to infer the effects of the mutations of the Alpha variant on (a) the transmission rate and (b) the mean duration of infectiousness from the analysis of the evolutionary dynamics taking place, in each region of England, just after the emergence of the variant (i.e., just after it reached 10% of the positive cases).

Methods

A two-step analysis

The analysis is performed in two steps considering two consecutive evo-epidemiological periods of time: Before and after the emergence of the Alpha variant in England (Figure 1). The

Table 1. Overview of the two-step analysis. This table summarizes the main features of the two phases of the analysis. For each one, we recall the aim, dates, circulating SARS-CoV-2 strains that we considered, fixed parameters, data, and fitted variables (model)—equation numbers are specified between brackets just after the corresponding variable. For both phases, we also use values of the Stringency Index in the United Kingdom. $\mathcal{R}_0$, γ (or γ_w), and β (or β_w) are the basic reproduction number and the per capita rates (per day) of recovery and transmission, respectively, of the resident strain w; κ is the per capita transition rate (per day) from the exposed to the infectious state (same for both strains); k and a are the parameters linking the Stringency Index to the efficacy of NPIs (same for both strains); $S(t)/N$ is the proportion of susceptible hosts in the population (assumed constant in the second phase); D refers to the cumulative density of COVID-19-related deaths. See Supplementary Tables S1 and S2 for a more detailed summary of the parameters involved in phases 1 and 2, respectively.

Dates	Strain(s)	Fixed parameters	Data	Fitted variable(s)
Phase 1—National frequency of the Alpha variant < 5 % AIM: Estimating the impact of NPIs (control parameters k and a) on the spread of the virus				
August 3, 2020 – November 8, 2020	Resident strain (WT)	• $\mathcal{R}_0 = 2.5$ • $\gamma = 0.1$ • $\beta = 0.25$ ($\approx \gamma\mathcal{R}_0$) • $\kappa = 0.2$ • $S(t_0^{\text{step 1}})/N = 0.9$	Daily new cases tested negative (United Kingdom) Daily new cases tested positive (United Kingdom) Daily new fatality cases (United Kingdom)	$T^-(t)$ (5) $T^+(t)$ (6) $D(t) - D(t-1)$ (3)
Phase 2—Regional frequency of the Alpha variant ≥ 10 % AIM: Knowing the impact of NPIs, estimating the phenotypic differences $\Delta\beta$ and $\Delta\gamma$				
Region-dependant (final week January 18, 2021)	Resident strain (WT) and Alpha variant	• k and a (estimations from phase 1) • $\beta_w = 0.25$ • $\gamma_w = 0.1$ • $\kappa = 0.2$ • $S/N = 0.75$ ($\approx$ final proportion of S at the end of the simulation of phase 1)	Weekly regional logit-frequencies of S Gene Target Failure among cases tested positive (England)	$\text{logit}(\tilde{f}_m(t))$ (11)

first step aims to estimate the force of infection in the presence of NPIs. In particular, we quantify $c(t)$, a function measuring the impact of NPIs at time t on the force of infection $\lambda(t)$. This first step takes place temporally before the emergence of the Alpha variant—that is, before it reaches 5% of the cases tested positive in England—and consists in modeling the epidemiological phase of the previous lineage, which we refer to as the resident strain, disregarding the pre-existing genetic diversity (Hodcroft et al., 2021). The second step consists in estimating the differences in contagiousness and in infectious duration in the presence of NPIs during the period when the two strains cohabit, that is, for each region, from the moment the frequency of the variant reaches 10% of cases tested positive. We combine information from screening and mortality data for the first step (using an epidemiological model), while we focus on the changes in frequency of the variant among positive cases for the second step. See Table 1 for an overview of this two-step approach.

For both steps, we consider a host population of size N. We note S, E, I, and R, respectively, the states (or compartments) of individuals that are Susceptible to the disease, Exposed (i.e., infected but not yet infectious), Infectious, and Recovered. For a given state, for instance S, and current time t (expressed in days), we note $S(t)$ the density of people in that state and $\dot{S}(t)$ its differentiation with respect to time. Let β be the per capita transmission rate (direct and horizontal) and γ the per capita recovery rate. Control measures implemented by governments such as social distancing, face coverings, lockdowns, or travel bans are NPIs that aim to curb the spread of the epidemic by alleviating the force of infection $\lambda(t) = \beta I(t)/N$. Given $c(t)$ the effectiveness of these measures—ranging from 0 (no control) to 1 (total control)—, the expression for the force of infection thus becomes: $\lambda(t) = (1 - c(t))\beta I(t)/N$. Directly estimating the control efficiency $c(t)$ is usually impossible; it results from a multitude of factors that may vary spatially and temporally and is not necessarily proportional to the severity of the measures in place. This is why we choose here to include the *Stringency Index* (which we noted $\psi(t)$), a composite score published by OxCGRT (Hale et al., 2021). This index is based on nine component indicators and rescaled to a value between 0 (no control) and 100 (the stringest) in order to reflect the strictness of public health policy. Eight component indicators are related to "*containment and closure*" (school and workplace closing, cancel public events, restrictions on gathering site, close public transport, stay-at-home requirements, and restrictions on internal movement and on international travel) and one is related to "*health system*" (public information campaign) (Hale et al., 2021). These measures, in contrast with post-symptomatic isolation or contact tracing (not explicitly taken into account in this score), are mainly limiting the number of contacts *un*conditionally to infection, that is mostly intended to reduce the transmission rate than to shorten the infectious period. We thus assume that NPIs included in the Stringency Index would only affect the transmission rate (and not the infectious period). Although somewhat imperfect, this index has the advantage of integrating many factors into one value, as well as being available per day online since the onset of the pandemic in many countries. We model the link between $c(t)$ and $\psi(t)$ through the following concave or convex relationship:

$$c(t) = k\left(\frac{\psi(t)}{100}\right)^a, \quad (1)$$

with $k \in [0; 1]$, the maximum achievable efficiency (when $\psi(t) = 100$), and with $a \in \mathbb{R}_+^*$, a "*shape*" parameter.

Step 1: Epidemiological analysis just before the emergence of the Alpha variant

We use a version of the well-known *SEIR* model (see Supplementary Figure S2) to estimate the parameters that govern the epidemiological dynamics before the arrival of the Alpha variant. We denote α, the additional per capita mortality rate induced by the viral disease (i.e., the virulence) and D, the compartment of (COVID-19-related) deceased individuals. We assume that the (potential) onset of symptoms and the onset of infectiousness occur simultaneously after a latent period of mean duration $1/\kappa$. Within the infectious compartment I, some hosts develop symptoms (I_S) with probability ω, while the others remain asymptomatic (I_A) with complementary probability. It is further assumed that individuals I_A systematically recover at a per capita rate γ, while individuals I_S are divided into two subcompartments depending on their fate: I_{Sd}, with probability p, for those who will eventually die from the disease (with virulence α), or, alternatively, I_{Sr}, for those who will eventually recover (at the same rate γ as asymptomatic hosts). We model these epidemiological trajectories using the following system of ordinary differential equations (ODEs):

$$\begin{cases} \dot{S}(t) = -(1-c(t))\beta S(t)\frac{I(t)}{N} \\ \dot{E}(t) = (1-c(t))\beta S(t)\frac{I(t)}{N} - \kappa E(t) \\ \dot{I}_A(t) = (1-\omega)\kappa E(t) - \gamma I_A(t) \\ \dot{I}_{Sr}(t) = (1-p)\omega\kappa E(t) - \gamma I_{Sr}(t) \\ \dot{I}_{Sd}(t) = p\omega\kappa E(t) - \alpha I_{Sd}(t) \\ \dot{R}(t) = \gamma\big(I_A(t) + I_{Sr}(t)\big) \\ \dot{D}(t) = \alpha I_{Sd}(t) \end{cases} \tag{2}$$

Following (Diekmann et al., 2010) for the construction of the Next Generation Matrix, the basic reproduction number $\mathcal{R}_0$—that is, the expected number of *infectees* from one *infector* in a fully susceptible population—is then given in the absence of NPI by:

$$\mathcal{R}_0 = \beta\left(\frac{1-\omega p}{\gamma} + \frac{\omega p}{\alpha}\right).$$

In the context of COVID-19, the product ωp—that is, the probability of developing symptoms and dying from the disease—is very low. We then approximate the basic reproduction number of the resident strain of SARS-CoV-2 as $\mathcal{R}_0 \approx \beta/\gamma$.

At each time point (each day), only a small fraction of the population is tested and hosts with symptoms are more likely to be tested than others. In order to take these biases into account, we use the following range of assumptions:

- Individuals S and I_A are tested with the same probability/reporting rate ρ;
- Individuals S and I_A can be tested several times;
- All new individuals I_S (symptomatic) are tested (reporting rate of 1);
- Screening of individuals E and R is neglected (reporting rate of 0);
- All new disease-related deaths are reported (reporting rate of 1).

Furthermore, screening efforts in the United Kingdom tended to be strengthened over time during this period (as shown for instance by the increasing number of negative tests in Supplementary Figure S3). As the reporting rate for individuals without symptoms S and I_A can no longer be considered constant, we also assume a linear increase with time:

- The reporting rate ρ for individuals S and I_A (without symptoms) increases linearly over time: $\rho(t) = \eta\, t + \mu$.

The reporting rate is not identifiable in an *SIR* model when only a fraction of the compartment I is observed (Hamelin et al., 2021). Thus, we also consider the disease-related deaths in the observation process. The combination of information, that is daily new cases tested negative and tested positive and daily new fatality cases, allows us to identify the reporting rate. Between two consecutive time points $t-1$ and t, the number of new fatality cases is given by:

$$D(t) - D(t-1) = \int_{t-1}^{t} \alpha I_{Sd}(t)\,\mathrm{d}t, \tag{3}$$

and, given $\int_{t-1}^{t} \omega\kappa E(t)\,\mathrm{d}t$, the *incidence* of symptomatic cases (i.e., new incomers in compartment I_S), we decomposed the number of performed tests $T(t)$ as follows:

$$T(t) = T^-(t) + T^+(t) = \underbrace{\rho(t)S(t)}_{T^-(t)} + \underbrace{\rho(t)I_A(t) + \int_{t-1}^{t} \omega\kappa E(t)\,\mathrm{d}t}_{T^+(t)}, \tag{4}$$

with $T^-(t)$ and $T^+(t)$, the number of cases tested negative and tested positive, respectively. Thus:

$$T^-(t) = (\eta\, t + \mu)\, S(t) \tag{5}$$

$$T^+(t) = (\eta\, t + \mu)\, I_A(t) + \int_{t-1}^{t} \omega\kappa E(t)\,\mathrm{d}t \tag{6}$$

Step 2: Evolutionary analysis

We now consider that two distinguishable pathogenic strains compete: the resident (or WT) strain, represented with the subscript w, and the mutant strain (or variant), represented with the subscript m. The total number of exposed hosts $E(t)$, where t is the current time, can therefore be decomposed into: $E(t) = E_m(t) + E_w(t)$. Likewise, for the infectious hosts $I(t)$: $I(t) = I_w(t) + I_m(t)$, and we denote $q_m(t) = I_m(t)/I(t)$, the frequency of the variant in I. We propose that the variant may differ phenotypically from the resident strain in its effective transmission rate $\beta_m = \beta_w + \Delta\beta$ and/or its recovery rate $\gamma_m = \gamma_w + \Delta\gamma$. In contrast, we neglect any difference in terms of latent period ($\kappa_m = \kappa_w = \kappa$), and we neglect the virulence of both strains ($\alpha_m = \alpha_w = 0$). For SARS-CoV-2, frequencies of the Alpha variant did not seem to depend on the age of hosts (Davies et al., 2021). Assuming furthermore that over-infections do not occur—including coinfections with both strains—and that (persistent) immunity acquired with either strain protects effectively against both, we start with the simple following *SEIR* model:

$$\begin{cases} \dot{S}(t) = -(1-c(t))\overline{\beta}(t)S(t)\frac{I(t)}{N} \\ \dot{E}(t) = (1-c(t))\overline{\beta}(t)S(t)\frac{I(t)}{N} - \kappa E(t) \\ \dot{I}(t) = \kappa E(t) - \overline{\gamma}I(t) \\ \dot{R}(t) = \overline{\gamma}(t)I(t) \end{cases} \tag{7}$$

where the overlines refer to mean values of the phenotypic traits after averaging over the distribution of strain frequencies:

$$\begin{cases}\overline{\beta}(t) = (1 - q_m(t))\beta_w + q_m(t)\beta_m \\ \overline{\gamma}(t) = (1 - q_m(t))\gamma_w + q_m(t)\gamma_m\end{cases}$$

As described in (Lion, 2018; Lion & Gandon, 2022), under the assumption of weak selection, the overall frequency of the variant $\tilde{f}_m(t)$ can be tracked using:

$$\frac{\mathrm{d}\tilde{f}_m(t)}{\mathrm{d}t} = \underbrace{\tilde{f}_m(t)(1 - \tilde{f}_m(t))}_{\text{Genetic variance}} \; \underbrace{\mathbf{v}(t)^\top \Delta\mathbf{R}(t)\mathbf{f}(t)}_{\text{s(t), selection coefficient}}, \tag{8}$$

with $\mathbf{v}(t)$ and $\mathbf{f}(t)$, the vectors of reproductive values and class frequencies, respectively, within the infected states (E and I), and $\Delta\mathbf{R}(t)$, the matrix of differences in transition rates between the mutant strain and the resident strain (for more details, see Supplementary Appendix). An easier way to study $s(t)$ in time series analyses is not to directly work with frequencies but with logit-frequencies instead, that is ln(frequency of the variant/frequency of the resident strain). Indeed, it may easily be shown that:

$$\frac{\mathrm{d}\ \mathrm{logit}(\tilde{f}_m(t))}{\mathrm{d}t} = s(t). \tag{9}$$

We then focus on the selection coefficient of the variant $s(t)$ (also known as the selection gradient). According to its value (weakly or strongly positive, weakly or strongly negative), this selection coefficient quantifies over time the success or the disadvantage of the variant over the resident strain through natural selection (Day & Gandon, 2006, 2007; Day et al., 2020). In Supplementary Appendix, we show that, using quasi-equilibrium approximations for fast variables, the selection coefficient of the variant $s(t)$ may be approximated as:

$$s(t) \approx \frac{2\kappa(1 - c(t))\Delta\beta\frac{S(t)}{N} - \left(\kappa - \overline{\gamma}(t) + \sqrt{\left(\kappa - \overline{\gamma}(t)\right)^2 + 4\kappa(1 - c(t))\overline{\beta}(t)\frac{S(t)}{N}}\right)\Delta\gamma}{2\sqrt{\left(\kappa - \overline{\gamma}(t)\right)^2 + 4\kappa(1 - c(t))\overline{\beta}(t)\frac{S(t)}{N}}}. \tag{10}$$

For the *SIR* model nested in the *SEIR* model (7), the selection coefficient is merely: $s(t) = (1-c(t))\Delta\beta S(t)/N - \Delta\gamma$ (Day & Gandon, 2006, 2007), which shows analytically the importance of the control through $c(t)$ to distinguish the scenario where the selective advantage of the variant stems from a higher transmission rate ($\Delta\beta > 0$; $\Delta\gamma = 0$) from the scenario with a longer duration of infectiousness ($\Delta\beta = 0$; $\Delta\gamma < 0$), or from an intermediate scenario ($\Delta\beta \neq 0$; $\Delta\gamma \neq 0$). In other words, it is particularly the variations in $c(t)$ that might help us to decouple the effects of these two phenotypic traits. Simply adding an exposed state E makes the selection gradient surprisingly much more difficult to express but the importance of the variations in $c(t)$ for this purpose (although less clear-cut) remains nevertheless relevant as suggested by (10).

Statistical inference

Programing

Numerical simulations and data analyses were carried out using R (R Core Team, 2021) version 4.1.1 (August 10, 2021). ODEs were solved numerically by the function "*ode*" (method "*ode45*") from the package "*deSolve*" (Soetaert et al., 2010).

Step 1

We used daily screening data between August 3, 2020 and November 8, 2020 in the United Kingdom (a period for which the Alpha variant was below 5% among cases tested positive in England); 7-day rolling average data were used in order to mitigate the effects of variation in testing activity, for example, during weekends. We also included daily COVID-19-related deaths in the United Kingdom ("*Daily deaths with COVID-19 on the death certificate by date of death*") as well as the Stringency Index.

The goal of this part is to compute $c(t)$ from the Stringency Index and thus to focus on the estimation of the parameters k and a. We used additional information from the literature to fix the value of some parameters of the model (2): We set the mean latent period to 5 days (Ding et al., 2021) and the mean duration of infectiousness to 10 days (Byrne et al., 2020), that is, an average infection period of 15 days; we also set the basic reproduction number $\mathcal{R}_0$ to 2.5 (Ferguson et al., 2020; Li et al., 2020) and the initial proportion of susceptible hosts to 0.9. Besides, we approximated the initial states within compartment I. This is summarized with further details in Supplementary Table S1. With these parameters fixed, the model (2) is identifiable (Supplementary Figure S4, following Raue et al., 2009). The remaining parameters of the first phase were estimated using weigthed least squares (WLS). Let $\theta = \left(k,\ a,\ E(t_0^{\text{step 1}}),\ \alpha,\ \omega,\ p,\ \eta,\ \mu\right)^\top$ be the vector of parameters to estimate, with $t_0^{\text{step 1}}$ the initial time point of the first step, and $\hat{\theta}$ its estimator such that:

$$\hat{\theta} = \underset{\theta}{\mathrm{argmin}} \sum_i \sum_t \frac{\left(y_i(t) - f_i(\theta, t)\right)^2}{f_i(\theta, t)},$$

where the subscript i refers to our three observation states, that is, daily new cases tested negative and tested positive and daily new fatality cases, which were modeled through functions f_i. $f_i(\theta, t)$ corresponds thus to the expected observations, while $y_i(t)$ corresponds to the real observations (data). With WLS, squared residuals are weighted by the inverse of the variance of the observations $y_i(t)$; these weights balanced the contrasting intrinsic contributions of each observation, for example, negative tests and deaths are not on the same order of magnitude. Assuming $y_i(t)$ to be Poisson-distributed—consistent with ODEs where sojourn times are exponentially distributed—then the variance of the observations is $f_i(\theta, t)$. This would correspond to the Pearson χ^2 function in (Berkson, 1980). Nonlinear optimizations were tackled with the R function "*optim*," from the basic package "*stats*", using the Nelder–Mead (or downhill simplex) method—maximum number of iterations *maxit* = 2,000, absolute and relative convergence tolerance *abstol* = *reltol* = 10^{-6}. This optimization procedure was iterated for 1,500 sets of uniformly drawn initial values (because of the presence of local minima) and was restricted to certain ranges of values through parameter transformations (cf. Supplementary Table S3). Only parameter estimates from the best fit, that is, successful completion with the lowest WLS value, were kept, and we refer to them as the best WLS estimates.

Parameter distributions were then computed using wild bootstrap (Kline & Santos, 2012; Liu, 1988), which allow in particular to take into account any heteroscedasticity in the residuals. To do this, 2,000 sets of bootstraped data were generated: Residuals were perturbed by an i.i.d. sequence of n random weights $\{W_i\}_{i=1}^{n}$ following Mammen's two-points distribution (that is, $(1-\sqrt{5})/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $(1+\sqrt{5})/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$), which satisfies $\mathbb{E}(W_i)=0$ and $\mathbb{E}(W_i^2)=1$ (Kline & Santos, 2012). Nonlinear optimizations were then reiterated, but starting only from the best WLS estimates and the corresponding set of initial values.

As a sensitivity analysis, ±10% and ±20% perturbations were applied to the fixed parameters of the first phase separately (β, κ, γ, and $S(t_0^{\text{step 1}})/N$), and nonlinear optimizations were each time reiterated starting from a set of 500 initial conditions (uniformly drawn, as before).

Step 2

We used weekly regional frequencies of S Gene Target Failure (SGTF) in England from the technical briefing 5 of Public Health England (PHE), which was investigating the new VOC 202012/01 variant between September 2020 and January 2021 (Public Health England, 2020). Briefly, qPCR from the ThermoFisher TaqPath kit (designed to target three genes: ORF1ab, N, and S) were performed after swab sampling in the wider population, that is, outside NHS hospitals and PHE labs. Due to the deletion ΔH69/V70 in the genome of the Alpha variant, a mismatch between one of the three molecular probes and the viral sequence encoding for the glycoprotein Spike (S) resulted in a failure of detection, or SGTF, a genomic signature that was then used as a proxy for this variant (Public Health England, 2020; Volz et al., 2021). As in the first step, we also included values of the Stringency Index in the United Kingdom.

Under the assumption that variations in $S(t)/N$ on short time scales may be neglected for a controlled epidemic ($S(t)/N \approx S/N$) and by neglecting the effect of the rise in frequency of the variant on the average phenotypic trait values, that is, $\bar{\beta}(t) \approx \beta_w$ and $\bar{\gamma}(t) \approx \gamma_w$ (weak selection approximation), we may integrate (10) in accordance with (9) to find an expression for the overall logit-frequency of the variant:

$$\begin{aligned}\operatorname{logit}(\tilde{f}_m(t)) &\approx \operatorname{logit}(\tilde{f}_m(t_0^{\text{step 2}}))\\ &+\kappa\int_{t_0^{\text{step 2}}}^{t}\left(\frac{(1-c(t))}{\sqrt{(\kappa-\gamma_w)^2+4\kappa(1-c(t))\beta_w S/N}}\right)\mathrm{d}t\ \ \Delta\beta\frac{S}{N}\\ &-\frac{1}{2}\left[(\kappa-\gamma_w)\int_{t_0^{\text{step 2}}}^{t}\left(\frac{1}{\sqrt{(\kappa-\gamma_w)^2+4\kappa(1-c(t))\beta_w S/N}}\right)\mathrm{d}t+\Delta t\right]\ \Delta\gamma,\end{aligned} \tag{11}$$

where $\Delta t = t - t_0^{\text{step 2}}$ is the period of time between the system at time t and its initial state.

$\Delta\beta$ appears as a product with S/N in (11), which implies that they are likely not to be separately identifiable. At the final time point of the first phase, our best fit ended up with a proportion of susceptible hosts around 0.75. Hence, we consistently set S/N to 0.75 for the second phase. As in the first phase, we also set: $\kappa = 0.2$, $\beta_w = 0.25$, and $\gamma_w = 0.1$. Phenotypic differences relative to the previous lineage ($\Delta\beta$ and $\Delta\gamma$ in (11)) were estimated using a linear mixed-effects model (MEM) to fit weekly logit-frequencies of SGTF among cases tested positive for COVID-19 as a proxy of the Alpha variant in the nine regions of England late 2020 early 2021. We assumed that these frequencies were representative of the infected population and that the regions were independent of each other, that is, no inter-region flows. In more detail, $\operatorname{logit}(\tilde{f}_m(t))$ was the response variable, $\Delta\beta$ and $\Delta\gamma$ were treated as fixed effects and the region (nine in total) was treated as a random effect on the intercept of the model. Hence, for the region i at time point t (i and t are now noted as indexes for clarity):

$$\underbrace{\operatorname{logit}(\tilde{f}_m)_{t,i}}_{\text{Response variable}} = intercept + \Delta\beta\ C_t^{\beta} + \Delta\gamma\ C_t^{\gamma} + Region_i + \varepsilon_{t,i},$$

with:

- $intercept$, the common fixed effect (reference);
- $C_t^{\beta} = \kappa\displaystyle\int_{t_0}^{t}\frac{(1-c(t))}{\sqrt{(\kappa-\gamma_w)^2+4\kappa(1-c(t))\beta_w S/N}}\,dt\ \frac{S}{N}$, the covariate associated with $\Delta\beta$ (fixed effect);
- $C_t^{\gamma} = -\frac{1}{2}\left[\left(\kappa-\gamma_w\right)\displaystyle\int_{t_0}^{t}\frac{dt}{\sqrt{(\kappa-\gamma_w)^2+4\kappa(1-c(t))\beta_w S/N}}+\Delta t\right]$, the covariate associated with $\Delta\gamma$ (fixed effect);
- $Region_i \sim \mathcal{N}\left(0,\ \nu^2\right)$, the random effect (with variance ν^2) of the region i on the intercept of the model;
- $\varepsilon_{t,i} \sim \mathcal{N}\left(0,\ \sigma^2\right)$, the residual error (with variance σ^2).

This MEM was implemented using the function "*lmer*" from the R package "*lme4*": $\operatorname{logit}(\tilde{f}_m) \sim \Delta\beta + \Delta\gamma + (1|Region)$, and 95% CIs of parameters $\Delta\beta$ and $\Delta\gamma$ were computed using the function "*confint*" from the package "*stats*." For each region, the initial date corresponds to the moment the Alpha variant reached 10% of cases tested positive, that is, above horizontal lines in Supplementary Figure S1-D. Below this threshold, the dynamics of the variant could not be considered deterministic. The parameters k and a that govern the link between the Stringency Index and the intensity of control (1) were set according to their best WLS estimates and joint distribution that were previously computed in the first step (cf. Step 1).

As for the first step, we investigated the robustness of our estimations. First, keeping our best WLS estimates for parameters k and a, linear MEMs were reiterated with ±10% and ±20% perturbations in the fixed parameters of the second phase separately (β_w, κ, γ_w, and S/N). Second, we used estimations of parameters k and a that we obtained after perturbing the fixed parameters of the first phase (cf. Step 1) to propagate these perturbations to the outcomes of the second step; the values of the fixed parameters of the second phase were updated each time in accordance.

Results

In the first step of the analysis, we develop an *SEIR* model (see equation (2) and Supplementary Figure S2) to capture the effect of the control measures $c(t)$ on the epidemiological dynamics. The effectiveness of these control measures has been quantified and monitored with the Stringency Index (Hale et al., 2021). As shown in the Methods section (A two-step analysis), the Stringency Index depends mainly on NPIs that decrease contacts with susceptible hosts, and we therefore assume that NPIs only affect transmission, but not the infectious period. We model the link between the effectiveness of the control measures $c(t)$ and the Stringency Index $\psi(t)$ at each time point t through the following function:

$$c(t) = k\left(\frac{\psi(t)}{100}\right)^{a}, \tag{1}$$

with k, the maximum achievable effectiveness, and a, a "*shape*" parameter; $\psi(t)$ takes values between 0 (no control) and 100. We generated daily new fatality cases (3), daily new cases tested negative (5) and daily new cases tested positive (6) that we fitted to observed data using weighted least squares (WLS) (see Methods, Step 1). The best WLS estimates for this model yielded $k = 1$ and $a = 3.78$. The adjusted model seemed to fit the general dynamics of the data even though somewhat locally perfectible (Supplementary Figure S3). We then quantified the uncertainty of our parameter estimates using wild bootstrap (Kline & Santos, 2012; Liu, 1988): We reiterated about 2,000 nonlinear optimizations on perturbed data in order to get 2,000 new sets of estimations (cf. Methods, Step 1). We thus obtained the joint distributions of the estimated parameters (see Supplementary Figure S5), and in particular parameters k and a that govern equation (1).

In the second step of the analysis, we seek to explain the rapid spread of the Alpha variant through an increase in the transmission rate and/or the recovery rate. We developed an *SEIR* model that takes into account the circulation of both the Alpha variant and the previous lineage, which we will refer to as the resident strain (Methods, Step 2: Evolutionary analysis). This model was used to derive an approximation of the temporal dynamics of the overall frequency $\tilde{f}_m(t)$ of the Alpha variant. Under the assumptions of weak selection and quasi-equilibrium of fast variables (for more details, see Supplementary Appendix), we obtained the following approximation for the selection coefficient s(t) of the Alpha variant:

$$s(t) = \frac{\mathrm{d}\,\mathrm{logit}(\tilde{f}_m(t))}{\mathrm{d}t} \approx \frac{\kappa + \bar{r}(t)}{\kappa + \bar{\gamma}(t) + 2\bar{r}(t)} \times \left[\frac{\Delta\beta}{\bar{\beta}(t)} \left(\bar{r}(t) + \bar{\gamma}(t) \right) - \Delta\gamma \right], \quad (12)$$

with $\mathrm{logit}(\tilde{f}_m(t)) = \ln(\tilde{f}_m(t)/(1 - \tilde{f}_m(t)))$ and where $\Delta\beta$ and $\Delta\gamma$ are the phenotypic differences between the Alpha variant and the resident strain in terms of transmission and recovery, respectively; $\bar{\beta}(t)$ and $\bar{\gamma}(t)$ refer to the average transmission and recovery rates across all genotypes; κ is the transition rate from the exposed state E to the infectious state I. Lastly, $\bar{r}(t)$ is the average growth rate of the epidemic:

$$\bar{r}(t) = q(t)\left((1 - c(t))\bar{\beta}(t)\frac{S(t)}{N} - \bar{\gamma}(t) \right),$$

with $q(t)$, the frequency of infectious individuals among infected hosts (i.e., $I(t)/(E(t)+I(t))$). It is important to note that NPIs affect the selection coefficient of the variant (12) through the growth rate of the epidemic $\bar{r}(t)$, which depends on the amount of control $c(t)$. Crucially, this impact is stronger if the Alpha variant is more transmissible (i.e., $\Delta\beta > 0$) (see also Day et al., 2020 and Otto et al., 2021). Interestingly, we found—as in (Otto et al., 2021)—a negative correlation between the selection coefficient of the Alpha variant in England and the Stringency Index: –0.88 at the national scale (95% CI [–0.98; –0.39]) and between –0.97 (London, 95% CI [–0.99; –0.86]) and –0.81 (South West, 95% CI [–0.97; –0.14]) at the regional level (Supplementary Figure S6). In the following, we approximated $\bar{r}(t)$ using the quasi-equilibrium expression of $q(t)$, we assumed that the proportion of susceptible hosts remained approximately constant during the second phase of the analysis ($S(t)/N \approx S/N$) and we neglected the effect of the rise in frequency of the variant on the average phenotypic trait values in (12) and $\bar{r}(t)$ (weak selection assumption).

Under these assumptions along with the previous best WLS estimates for the control parameters from the first step of the analysis ($k = 1$, $a = 3.78$), the fitted linear MEM (Supplementary Figure S7) led to the following estimations of the phenotypic differences (per day): $\Delta\beta = 0.15$ (95% CI [0.033; 0.258]) and $\Delta\gamma = -0.047$ (95% CI [–0.099; +0.001]) (Figure 2). With a significance level of 5%, likelihood-based comparisons of nested MEMs show a significant effect for $\Delta\beta$, but not for $\Delta\gamma$ (although with a p-value very close to the significance threshold) (Supplementary Table S4). In addition, we sought to propagate to the second phase the uncertainty of our estimates of the parameters k and a. Starting from each of the almost 2,000 pairs $\{k; a\}$ based on previous wild bootstrap computations, we obtained as many new estimators for $\{\Delta\beta; \Delta\gamma\}$. For $\Delta\beta$, 95% of them were between 0.147 and 0.153 (Supplementary Figure S9-A), for which each corresponding 95% CI remained positive (Supplementary Figure S8). In contrast, 95% of these 2,000 estimates were between –0.054 and –0.046 for $\Delta\gamma$ (Supplementary Figure S9-B), among which 61% of the corresponding 95% CIs included 0 (Supplementary Figure S8). These distributions led us to conclude that the Alpha variant has a higher transmission rate than the resident strain. With these estimates of $\Delta\beta$ and $\Delta\gamma$ and in the absence of NPI, the selection coefficient of the Alpha variant was computed, on average, around 0.77 per week (*SD* 0.02 per week).

We also explored the robustness of these estimations by applying ±10% and ±20% perturbations in the fixed parameters of our model (cf. Table 1) to investigate how they would affect our results. First, we kept the best WLS estimates for the control parameters ($k = 1$, $a = 3.78$), and we applied the perturbations to the fixed parameters of the second phase of the analysis. Our estimations of $\Delta\beta$ and $\Delta\gamma$ were not very sensitive to these perturbations (cf. Supplementary Figure S10). Second, we applied the perturbations in the fixed parameters of the first step in order to get new estimates of the control parameters k and a. We used these new estimates in the second phase of the analysis to estimate the phenotypic differences $\Delta\beta$ and $\Delta\gamma$. The parameter k was hardly affected by these perturbations but the parameter a was more sensitive, in particular when varying the transmission rate or the initial proportion of susceptible hosts (cf. Supplementary Figure S11-1). Next, we reiterated the second step with these new estimations of the pair $\{k; a\}$. All the 95% CIs of the estimates of $\Delta\beta$ remained positive after these perturbations. However, some perturbations led to more negative values of $\Delta\gamma$ (i.e., the 95% CIs of $\Delta\gamma$ included only negative values, Supplementary Figure S11-2). Note that this effect seems to be driven by the variations in the estimation of the parameter a (cf. Supplementary Figure S11). Taken together, the results of these analyses confirm the conclusion that the Alpha variant has a higher transmission ($\Delta\beta > 0$). An increase in the mean duration of infectiousness ($\Delta\gamma < 0$) seems less likely but cannot be completely ruled out.

Discussion

We developed a two-step approach to characterize the phenotypic variation of the Alpha variant relative to the previously dominant lineage. In the first step of the analysis, we focus on the epidemiological dynamics before the

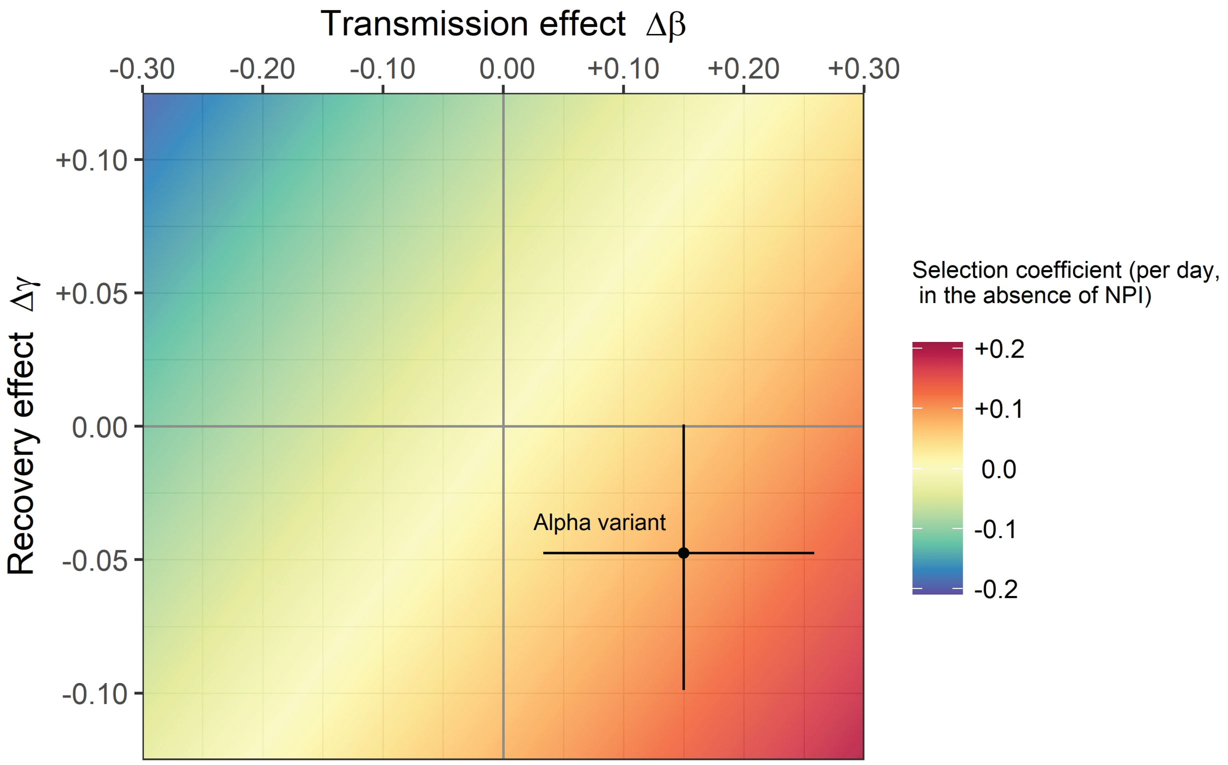


Figure 2. Phenotypic profile of the Alpha variant (transmission and recovery rates) relative to the resident strain. By definition, the phenotype of the resident strain is located at the origin of the graph ($\Delta\beta = 0$; $\Delta\gamma = 0$). Linear MEM estimates (black point, expressed per day) of phenotypic differences in transmission $\Delta\beta$ and in recovery $\Delta\gamma$ as well as 95% CIs (black cross) are based on the best WLS estimates of control parameters k and a from the analysis of phase 1 ($k = 1$ and $a = 3.78$). We obtained $\Delta\beta = 0.15$ (95% CI $[0.033; 0.258]$) and $\Delta\gamma = -0.047$ (95% CI $[-0.099; +0.001]$). For the fixed parameters, we set: $S/N = 0.75$, $\kappa = 0.2$, $\beta_w = 0.25$ and $\gamma_w = 0.1$. The colored background represents the values of the selection coefficient (in the absence of NPI) as a function of $\Delta\beta$ and $\Delta\gamma$; the selection coefficient is here around +0.11 per day (or +0.77 per week) for the Alpha variant. Estimates and 95% CIs based on the joint distributions of parameters k and a from wild bootstrap computations are represented in Supplementary Figure S8.

emergence of the Alpha variant and we used an *SEIR* model, a simplified representation of an age-structured model, to infer the effect of the Stringency Index on the reduction of transmission induced by these control measures. This led us to infer a convex increasing function that captures the effect of the Stringency Index on the reduction in the number of contacts with susceptible hosts (Supplementary Figure S12).

The second step of this approach is based on the analysis of the change in frequency of the Alpha variant after its emergence. Using evolutionary epidemiology theory (Day & Gandon, 2006, 2007; Day et al., 2020), we derive an expression for the gradient of selection in an *SEIR* model. The analysis of selection in such a class-structured environment (the virus is infecting both the *E* and the *I* hosts) is facilitated under the assumption of weak selection and the approximation of quasi-equilibrium for fast variables (Gandon & Lion, 2022; Lion, 2018; Lion & Gandon, 2022). We recover a classical result derived from simpler *SIR* models: The intensity of selection for higher transmission rates depends on the availability of susceptible hosts and the amount of NPIs aiming to reduce contact (e.g., social distancing or face coverings). In contrast, selection for longer durations of infectiousness is much less sensitive to these control measures. Using our independent estimation of the effectiveness of NPIs based on the Stringency Index, we inferred both $\Delta\beta$ and $\Delta\gamma$ of the Alpha variant from the temporal dynamics of its logit-frequency. This analysis suggests that the selective advantage of the Alpha variant was mainly driven by a higher transmission rate. An increase in the mean duration of infectiousness (i.e., a lower rate of recovery) seems less likely but cannot be completely ruled out. Interestingly, recent experimental studies of viral transmission confirm the transmission advantage of the Alpha variant. Viral shedding in breath aerosols was indeed found to be higher in individuals infected with the Alpha variant than with previous lineages (Lai et al., 2022).

Several specific mutations of the Alpha variant could explain these phenotypic differences. Preliminary genomic characterizations detected around 17 nonsynonymous substitutions or deletions compared to the previous lineage; about half were associated with the protein S gene, including mutations of immunological significance (Rambaut et al., 2020). In particular, the mutation N501Y, known to increase the affinity of the viral glycoprotein S for the human receptor ACE2 (Starr et al., 2020), and the mutation P681H, adjacent to a serine protease cleavage site that is required for cell infection (Hoffmann et al., 2020), are both likely to affect the within-host development of the virus in infected hosts. How this development affects key phenotypic traits like transmission and recovery rates in human host is difficult to explore experimentally. Our analysis can thus provide a complementary approach that may help to link genetic and phenotypic variation.

Yet, it is important to note that this analysis relies on several simplifying assumptions. For instance, we assumed that infectiousness began at the same time as the onset of symptoms, that is, the latent period and the incubation period coincide perfectly in time. Yet, transmission from a presymptomatic state is a distinctive feature of SARS-CoV-2 (Day et al., 2020; He et al., 2020; Rothe et al., 2020). Besides, our framework sticks to the *SEIR* class of models formalized by ODEs, with κ and γ, the (constant) rates of leaving the exposed and infectious compartments, respectively. This implicitly yields sojourn times in the different compartments that are exponentially distributed—and thus, markovian or memoryless (Forien et al., 2021; Sofonea et al., 2021). As a result, the generation time follows a hypoexponential distribution (generalized Erlang distribution) with mean $1/\kappa + 1/\gamma$ and variance $1/\kappa^2 + 1/\gamma^2$ (Wallinga & Lipsitch, 2007). Our analysis does not allow the mean and variance of this distribution to change independently but a variation in γ does affect the mean and the variance of the generation time. Several studies, however, have discussed the influence of the shape of the generation time distribution on both the epidemiological and evolutionary dynamics of the pathogen (Abbott et al., 2022; Blanquart et al., 2022; Day, 2003; Park et al., 2019, 2022; Wallinga & Lipsitch, 2007). We show in Supplementary Appendix S7 how to recover our results using the selection on the shape of the generation time distribution used by (Blanquart et al., 2022). In both analyses, variations in the intensity of NPIs are assumed to impact the effective reproduction number without altering the generation time distribution (which means they only impact transmission). Nevertheless, some control measures like contact tracing and postsymptomatic isolation may impact the duration of infectiousness, the generation time distribution, and the selection on the variant (Park et al., 2022).

Data availability and quality are major limiting factors in any statistical inference analysis. The Stringency Index provides a rough approximation of the intensity of control at the national scale. More precise and more local estimations of control would allow us to refine our estimations. In addition, we show in Supplementary Appendix S6 how the availability of data frequency among different types of hosts (i.e., the differentiation between the exposed and the infectious compartments) may provide another way to estimate $\Delta\beta$ and $\Delta\gamma$.

To conclude, we contend that it is important to exploit the joint epidemiological and evolutionary dynamics of SARS-CoV-2 to better understand its phenotypic evolution. This phenotypic evolution is undermining our efforts to control the epidemic. New variants are emerging and are affecting other phenotypic traits. In particular, the ability of new variants (e.g., Omicron) to escape immunity has a major impact on the epidemiological dynamics (Paton et al., 2022). Inference approaches using both epidemiological and evolutionary analysis could yield important insights on the adaptive trajectories on the phenotypic landscape of SARS-CoV-2, and possibly other pathogens.

Supplementary material

Supplementary material is available online at *Evolution*.

Data availability

Data that were used in this study, along with the scripts for the analyses (in R), are available in Dryad (DOI: 10.5061/dryad.ns1rn8q04) as well as in a GitHub repository.

Author contributions

Conceptualization, methodology, investigation, and writing (original draft and review & editing): W.B., S.L., R.C., and S.G.; formal analysis: W.B.; vizualization: W.B. and S.G.; supervision: S.L., R.C., and S.G.

Funding

This work was funded by grants ANR-16-CE35-0012 "STEEP" to S.L. and ANR-17-CE35-0012 "EVOMALWILD" to S.G. from the Agence Nationale de la Recherche. We also thank the MESRI (French Ministry of Research) and the École Normale Supérieure Paris-Saclay for the PhD scholarship of W.B.

Conflict of interest: The authors declare no conflict of interest.

Acknowledgment

We thank Troy Day and François Blanquart for inspiring discussions.

References

Abbott, S., Sherratt, K., Gerstung, M., & Funk, S. (2022). Estimation of the test to test distribution as a proxy for generation interval distribution for the Omicron variant in England. *medRxiv*. https://doi.org/10.1101/2022.01.08.22268920

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *Annals of Statistics*, *8*(3):457–487. https://www.jstor.org/stable/2240587

Blanquart, F., Hozé, N., Cowling, B. J., Débarre, F., & Cauchemez, S. (2022). Selection for infectivity profiles in slow and fast epidemics, and the rise of SARS-CoV-2 variants. *eLife*, *11*:e75791. https://doi.org/10.7554/eLife.75791

Byrne, A. W., McEvoy, D., Collins, A. B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E. A., McAloon, C., O'Brien, K., Wall, P., Walsh, K. A., & More, S. J. (2020). Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open*, *10*(8):e039856. https://doi.org/10.1136/bmjopen-2020-039856

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., van Zandvoort, K., Silverman, J. D., CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, Diaz-Ordaz, K., Keogh, R., Eggo, R. M., Funk, S., Jit, M., Atkins, K. E., & Edmunds, W. J. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, *372*(6538):eabg3055. https://doi.org/10.1126/science.abg3055

Day, T. (2003). Virulence evolution and the timing of disease life-history events. *Trends in Ecology & Evolution*, *18*(3):113–118. https://doi.org/10.1016/S0169-5347(02)00049-6

Day, T. & Gandon, S. (2006). Insights from Price's equation into evolutionary epidemiology. In *Disease evolution: Models, concepts, and data analysis*. Z. Feng, U. Dieckmann & S. Levin (Eds.), (Vol. 71). *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. (pp. 23–44). American Mathematical Society. https://doi.org/10.1090/dimacs/071/02

Day, T. & Gandon, S. (2007). Applying population-genetic models in theoretical evolutionary epidemiology. *Ecology Letters*, *10*:876—888. https://doi.org/10.1111/j.1461-0248.2007.01091.x

Day, T., Gandon, S., Lion, S., & Otto, S. P. (2020). On the evolutionary epidemiology of SARS-CoV-2. *Current Biology*, *30*(15):R841—R870. https://doi.org/10.1016/j.cub.2020.06.031

Diekmann, O., Heesterbeek, J. A. P., & Roberts, M. G. (2010). The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, *7*(47):873–885. https://doi.org/10.1098/rsif.2009.0386

Ding, Z., Wang, K., Shen, M., Wang, K., Zhao, S., Song, W., Li, R., Li, Z., Wang, L., Feng, G., Hu, Z., Wei, H., Xiao, Y., Bao, C., Hu, J., Zhu, L., Li, Y., Chen, X., Yin, Y., Wang, W., Cai, Y., Peng, Z., & Shen, H. (2021). Estimating the time interval between transmission generations and the presymptomatic period by contact tracing surveillance data from 31 provinces in the mainland of China. *Fundamental Research*, *1*(2):104–110. https://doi.org/10.1016/j.fmre.2021.02.002

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., van Elsland, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P. G., Walters, C., Winskill, P., Whittaker, C., Donnelly, C. A., Riley, S., & Ghani, A. C. (2020). *Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand*. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf

Forien, R., Pang, G., & Pardoux, É. (2021). Estimating the state of the COVID-19 epidemic in France using a model with memory. *Royal Society Open Science*, *8*(3):202327. https://doi.org/10.1098/rsos.202327

Gandon, S. & Lion, S. (2022). Targeted vaccination and the speed of SARS-CoV-2 adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(3):e2110666119. https://doi.org/10.1073/pnas.2110666119

Grubaugh, N. D., Hanage, W. P., & Rasmussen, A. L. (2020a). Making sense of mutation: What D614G means for the COVID-19 pandemic remains unclear. *Cell*, *182*:794–795. https://doi.org/10.1016/j.cell.2020.06.040

Grubaugh, N. D., Petrone, M. E., & Holmes, E. C. (2020b). We shouldn't worry when a virus mutates during disease outbreaks. *Nature Microbiology*, *5*:529–530. https://doi.org/10.1038/s41564-020-0690-4

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23):4121—4123. https://doi.org/10.1093/bioinformatics/bty407

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, *5*:529—538. https://doi.org/10.1038/s41562-021-01079-8

Hamelin, F., Iggidr, A., Rapaport, A., & Sallet, G. (2021). *Observability, identifiability and epidemiology a survey*. HAL. https://hal.archives-ouvertes.fr/hal-02995562/document

He, X., Lau, E. H. Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling, B. J., Li, F., & Leung, G. M. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, *26*:672—675. https://doi.org/10.1038/s41591-020-0869-5

Hodcroft, E. B. (2021). CoVariants: SARS-CoV-2 mutations and variants of interest. https://covariants.org/

Hodcroft, E. B., Zuber, M., Nadeau, S., Vaughan, T. G., Crawford, K. H. D., Althaus, C. L., Reichmuth, M. L., Bowen, J. E., Walls, A. C., Corti, D., Bloom, J. D., Veesler, D., Mateo, D., Hernando, A., Comas, I., González-Candelas, F., SeqCOVID-SPAIN consortium, Stadler, T., & Neher, R. A. (2021). Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature*, *595*:707—712. https://doi.org/10.1038/s41586-021-03677-y

Hoffmann, M., Kleine-Weber, H., & Pöhlmann, S. (2020). A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular Cell*, *78*(4):779–784. https://doi.org/10.1016/j.molcel.2020.04.022

Kline, P. M. & Santos, A. (2012). A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, *1*(1):23—41. https://doi.org/10.1515/2156-6674.1006

Lai, J., Coleman, K. K., Sheldon Tai, S.-H., German, J., Hong, F., Albert, B., Esparza, Y., Srikakulapu, A. K., Schanz, M., Sierra Maldonado, I., Oertel, M., Fadul, N., Louie Gold, T., Weston, S., Mullins, K., McPhaul, K. M., Frieman, M., & Milton, D. K. (2022). Evolution of SARS-CoV-2 shedding in exhaled breath aerosols. *Clinical Infectious Diseases*, *76*(5):786–794. https://doi.org/10.1093/cid/ciac846

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T. T. Y., Wu, J. T., Gao, G. F., Cowling, B. J., Yang, B., Leung, G. M., & Feng, Z. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *The New England Journal of Medicine*, *382*(13):1199—1207. https://doi.org/10.1056/NEJMoa2001316

Lion, S. (2018). Class structure, demography and selection: reproductive-value weighting in nonequilibrium, polymorphic populations. *The American Naturalist*, *191*(5):620–637. https://doi.org/10.1086/696976

Lion, S. & Gandon, S. (2022). Evolution of class-structured populations in periodic environments. *Evolution*, *76*(8):1674—1688. https://doi.org/10.1101/2021.03.12.435065

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics*, *16*(4):1696—1708. https://doi.org/10.1214/aos/1176351062

Lu, H., Stratton, C. W., & Tang, Y.-W. (2020). Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *Journal of Medical Virology*, *92*(4):401–402. https://doi.org/10.1002/jmv.25678

Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L., Bosso, M., Park, D. J., & Babadi, M. (2022). Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science*, *376*(6599):1327–1332. https://doi.org/10.1126/science.abm1208

Otto, S. P., Day, T., Arino, J., Colijn, C., Dushoff, J., Li, M., Mechai, S., Van Domselaar, G., Wu, J., Earn, David, J. D., & Ogden, N. H. (2021). The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Current Biology*, *31*(14):R918–R929. https://doi.org/10.1016/j.cub.2021.06.049

Park, S. W., Bolker, B. M., Funk, S., Metcalf, C. J. E., Weitz, J. S., Grenfell, B. T., & Dushoff, J. (2022). The importance of the generation interval in investigating dynamics and control of new SARS-CoV-2 variants. *Journal of The Royal Society Interface*, *19*(191):20220173. https://doi.org/10.1098/rsif.2022.0173

Park, S. W., Champredon, D., Weitz, J. S., & Dushoff, J. (2019). A practical generation-interval-based approach to inferring the strength of epidemics from their speed. *Epidemics*, *27*:12–18. https://doi.org/10.1016/j.epidem.2018.12.002

Paton, R. S., Overton, C. E., & Ward, T. (2022). The rapid replacement of the Delta variant by Omicron (B.1.1.529) in England. *Science Translational Medicine*, *14*(652):eabo5395. https://doi.org/10.1126/scitranslmed.abo5395

Public Health England (2020). Investigation of novel SARS-CoV-2 variant 202012/01 (Technical briefing 5). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959426/Variant_of_Concern_VOC_202012_01_Technical_Briefing_5.pdf

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D. L., & Volz, E. (2020). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineagein-the-uk-defined-by-a-novel-set-of-spike-mutations/563

Raue, A., Kreutz, C., Maiwald, T., Bachman, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, *25*(15):1923—1929. https://doi.org/10.1093/bioinformatics/btp358

Rausch, J. W., Capoferri, A. A., Katusiime, M. G., Patro, S. C., & Kearney, M. F. (2020). Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(40):24614—24616. https://doi.org/10.1073/pnas.2017726117

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke, C., Guggemos, W., Seilmaier, M., Drosten, C., Vollmar, P., Zwirglmaier, K., Zange, S., Wölfel, R., & Hoelscher, M. (2020). Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *The New England Journal of Medicine*, *382*:970–971. https://doi.org/10.1056/NEJMc2001468

Soetaert, K., Petzoldt, T., & Setzer, R. W. (2010). Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, *33*(9):1—25. https://doi.org/10.18637/jss.v033.i09

Sofonea, M. T., Reyné, B., Elie, B., Djidjou-Demasse, R., Selinger, C., Michalakis, Y., & Alizon, S. (2021). Memory is key in capturing COVID-19 epidemiological dynamics. *Epidemics*, 35:100459. https://doi.org/10.1016/j.epidem.2021.100459

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., & Bloom, J. D. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, *182*(5):1295—1310. https://doi.org/10.1016/j.cell.2020.08.012

van Dorp, L., Richard, D., Tan, C. C. S., Shaw, L. P., Acman, M., & Balloux, F. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nature Communications*, *11*(5986). https://doi.org/10.1038/s41467-020-19818-2

Vasilarou, M., Alachiotis, N., Garefalaki, J., Beloukas, A., & Pavlidis, P. (2021). Population genomics insights into the first wave of COVID-19. *Life*, *11*(2):129. https://doi.org/10.3390/life11020129

Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., Gonçalves, S., Jorgensen, D., Myers, R., Hill, V., Jackson, D. K., Gaythorpe, K., Groves, N., Sillitoe, J., Kwiatkowski, D. P., The COVID-19 Genomics UK (COG-UK) consortium, Flaxman, S., Ratmann, O., Bhatt, S., Hopkins, S., Gandy, A., Andrew, R., & Ferguson, N. M. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*, *593*:266—269. https://doi.org/10.1038/s41586-021-03470-x

Wallinga, J. & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, *274*:599–604. https://doi.org/10.1098/rspb.2006.3754

World Health Organization (WHO) (2020). *WHO situation report on 11 February 2020*. (No. 22). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=fb6d49b1_2

—— Supplementary Figures and Tables ——

Phenotypic evolution of SARS-CoV-2: a statistical inference approach

Wakinyan Benhamou iD, Sébastien Lion iD, Rémi Choquet† iD and Sylvain Gandon† iD

CEFE, CNRS, Univ Montpellier, EPHE, IRD, Montpellier, France
†: equal contribution

July 5, 2023

Supplementary figures (pp. 3-15)

Supplementary tables (pp. 16-18)

References for supplementary figures and tables (p. 19)

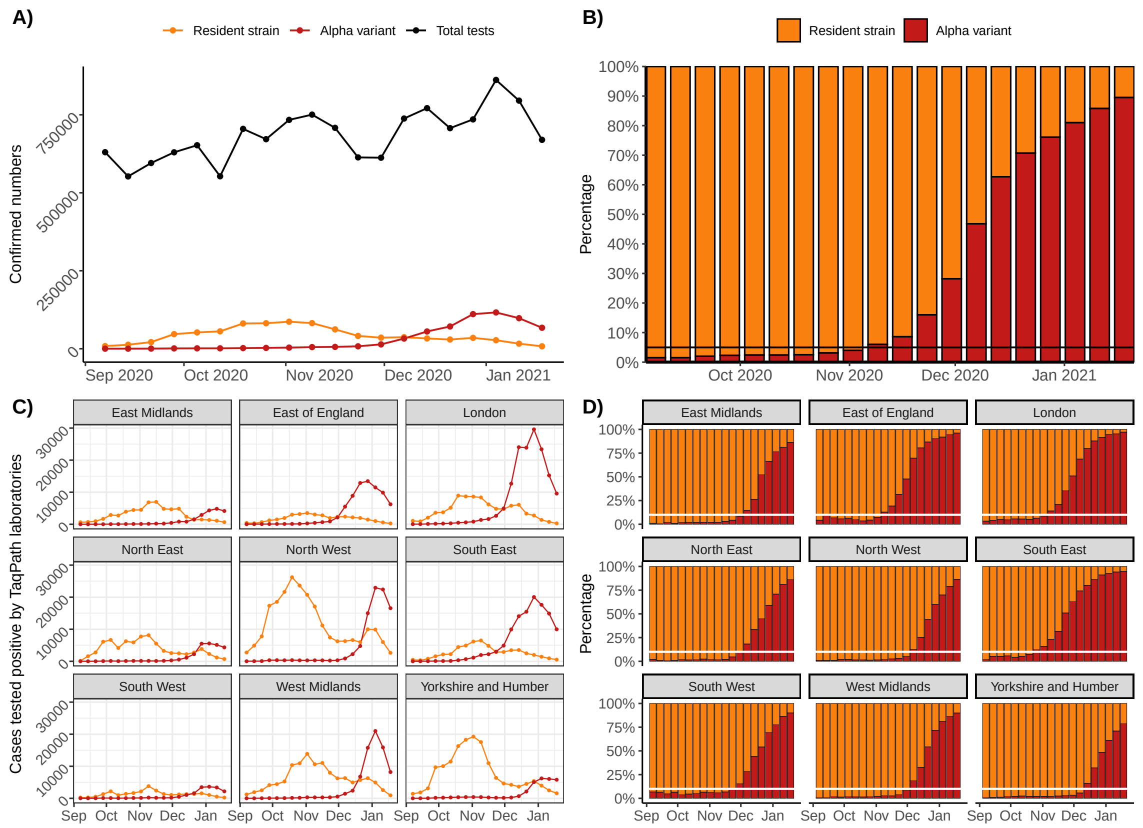


Figure S1: **Epidemiological and genetic data from the COVID-19 outbreak in England between September 2020 and January 2021.** Weekly numbers of TaqPath Pillar 2 COVID-19 positive tests associated with the resident strain of SARS-CoV-2 or with the Alpha variant (lineage B.1.1.7) at the national scale (**A**, total numbers of tests (not only TaqPath tests) are shown in black) and at regional scale (**C**). Weekly percentages of each strain within the TaqPath Pillar 2 COVID-19 positive tests at the national scale (**B**) and at the regional scale (**D**). SGTF from qPCR was used as a proxy of the Alpha variant. In this study, we considered two consecutive evo-epidemiological phases: (i) the phase that preceded the emergence of the Alpha variant, and (ii) the phase that followed it. The first phase took place just before the frequency of the variant reached 5% of the cases tested positive at the national scale (horizontal black line in **B**). Then, for each region, the second phase started at the date the variant reached at least the threshold value 10%, indicated in **D** with horizontal white lines. Below, the number of cases associated with the Alpha variant was quite low and the dynamics of its frequency was widely driven by stochastic processes – see for example the erratic dynamics below the horizontal white line for the second region ('East of England') in **D**.

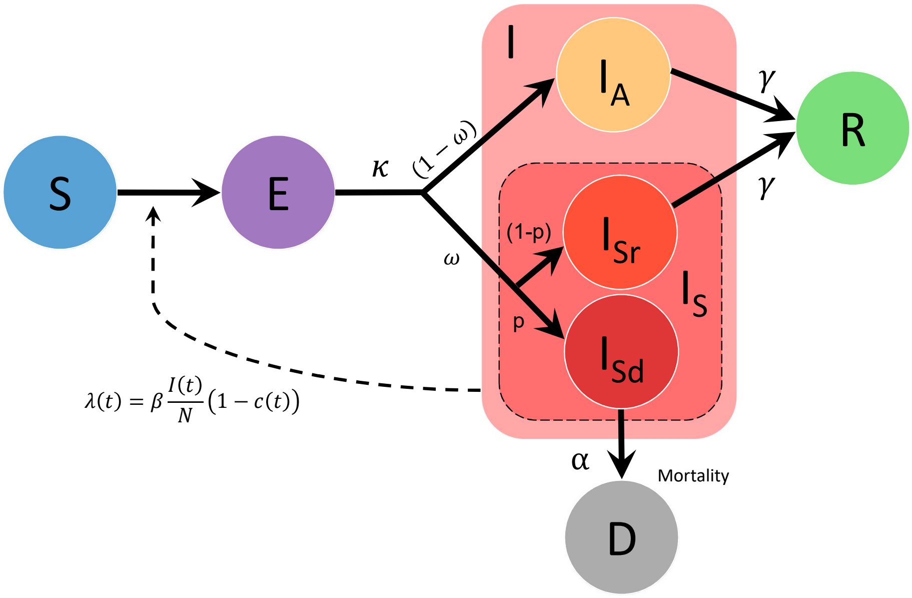


Figure S2: **Flow chart of the epidemiological *SEIR* model used in the first step of our analysis (phase 1).** Hosts may be: S (Susceptible to the infection), E (Exposed, that is infected but not yet infectious), I (Infected, with I_A: Asymptomatic; I_S: Symptomatic (with subscript r for those for will eventually recover and d for those who will eventually die from the disease)), R (Recovered and immunised) and D (Deceased). Transitions are represented by solid line arrows associated with the corresponding parameters (see Table S1 for definitions). The dashed line arrow symbolises the role of compartment I in the force of infection $\lambda(t)$ – i.e. transition rate from compartment S to compartment E. $c(t)$ is the efficacy of NPIs.

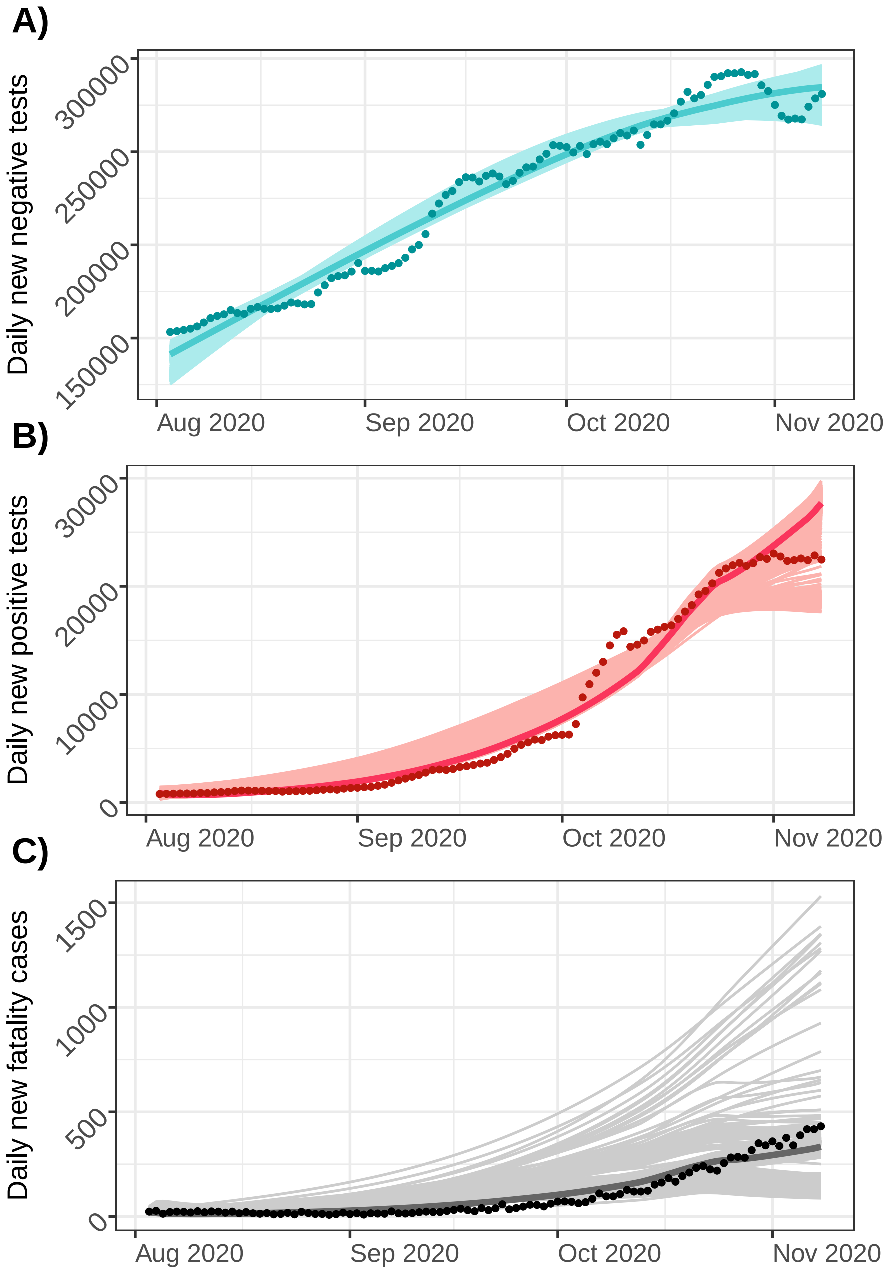


Figure S3: **Model fitting to data for the first phase (from 2020-08-03 to 2020-11-08 in the UK).** Observed data – (**A**) daily new cases tested negative, (**B**) daily new cases tested positive and (**C**) daily new fatality cases – are shown as dark points. Fits based on the best WLS estimates are represented as dark lines and about 2000 model fits resulting from wild bootstrap are represented as light lines.

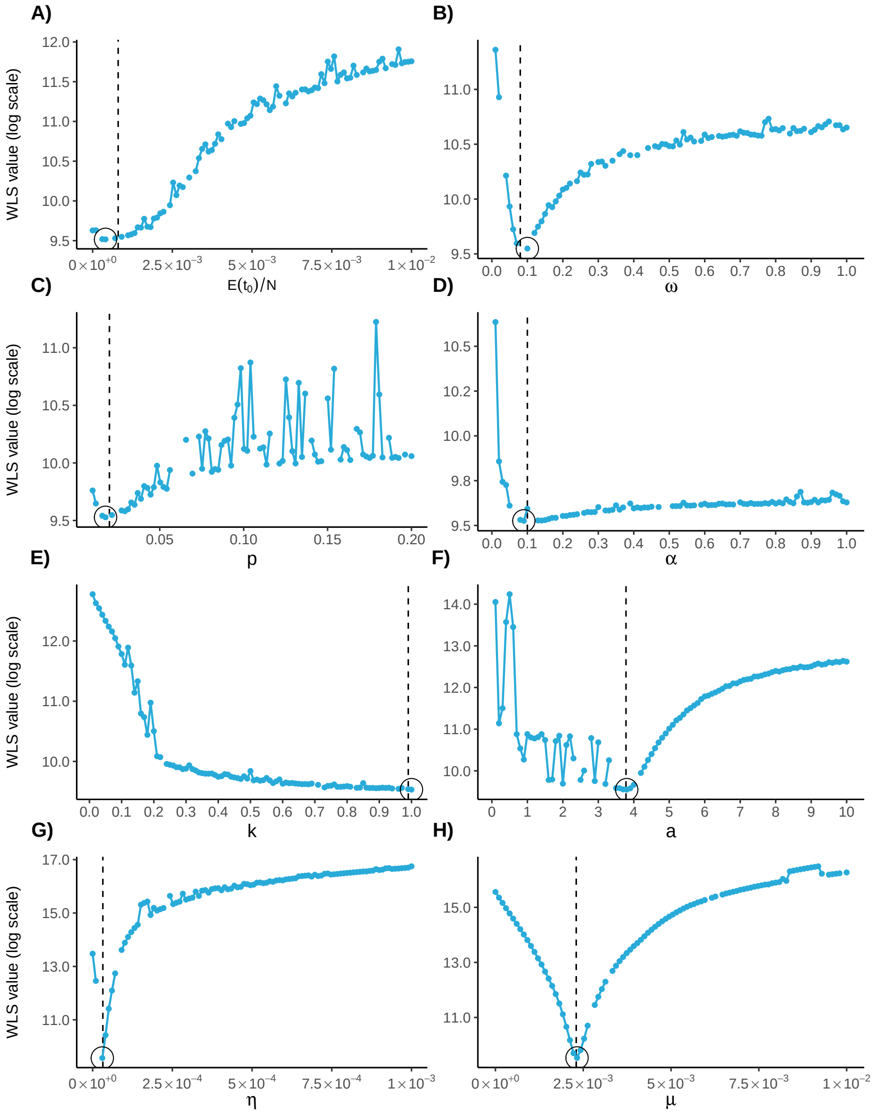


Figure S4: **Identifiability profiles of the first phase.** Simulated data for the first phase – i.e. (i) new cases tested negative, (ii) new cases tested positive and (iii) new fatality cases – were generated with the parameter values indicated by the vertical dashed lines: $E(t_0)/N = 8.0 \times 10^{-4}$ (where t_0 refers to the initial time point of the first phase), $\omega = 0.08$, $p = 0.02$, $\alpha = 0.1$, $k = 0.99$, $a = 3.78$, $\eta = 3.2 \times 10^{-5}$, $\mu = 2.3 \times 10^{-3}$; i.i.d. Gaussian noise was added to each series of simulated data, with standard deviation (i) 5000, (ii) 500 and (iii) 5. For the fixed parameters, we set: $\kappa = 0.2$, $\gamma = 0.1$, $\mathcal{R}_0 = 2.5$, $\beta = \gamma\mathcal{R}_0 = 0.25$ and $S(t_0)/N = 0.9$. We used real values for the Stringency Index (from 2020-08-02 to 2020-11-08 in the UK). Profiles were built following (Raue et al., 2009): a parameter of interest is set to a given value (on the x-axis) and the others are estimated to obtain a WLS value (y-axis, here on log scale); this is then reiterated with different values of the parameter of interest to cover the desired range (x-axis). As initial conditions, we only started from the parameter values that we used to simulate the data. Discontinuities in the profiles result merely from convergence failures. Points associated with the lowest WLS value are enclosed in a circle and their good match with the values used to simulate the data (vertical dashed lines) confirms that these parameters should be identifiable. Though, the shallowness of some profiles suggest that precise estimations may be numerically difficult.

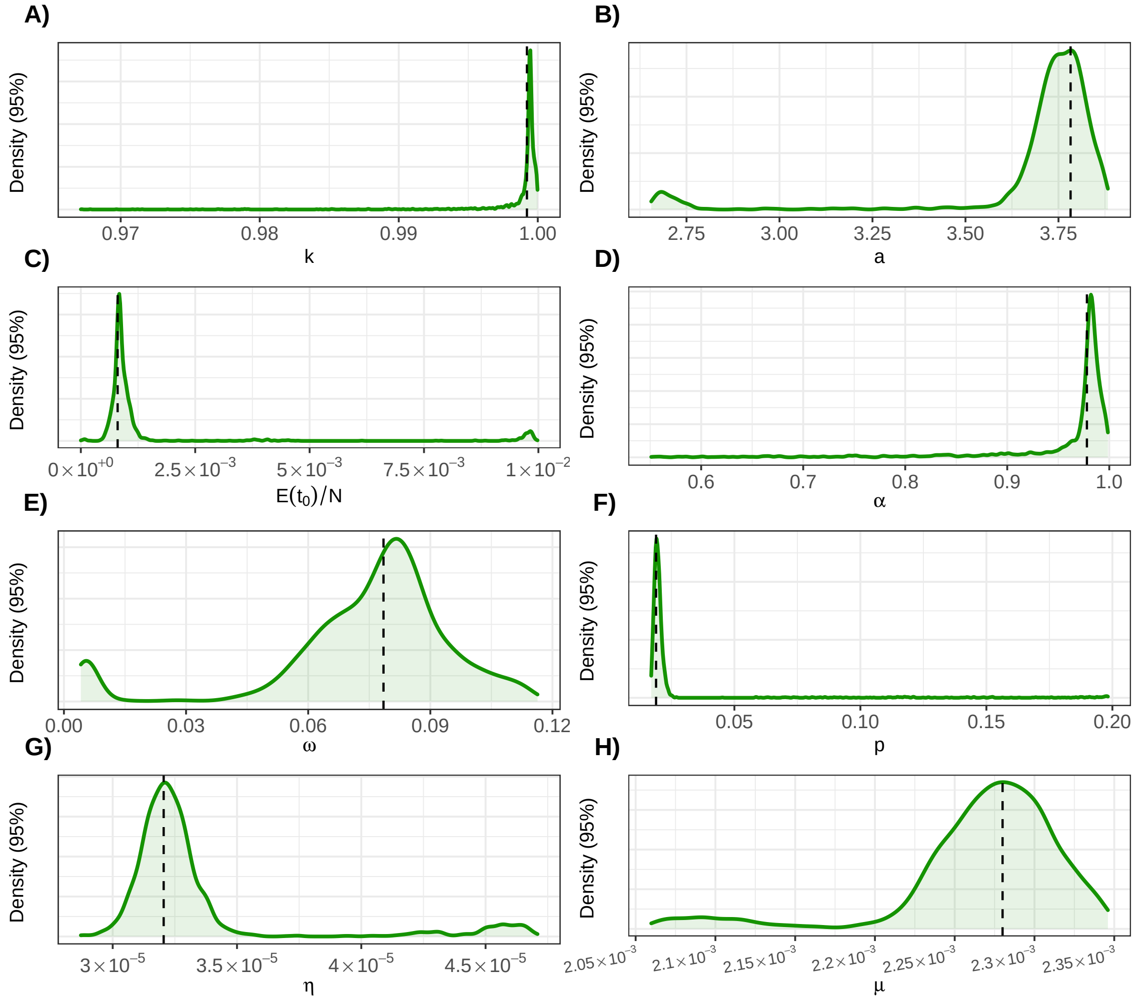


Figure S5: **95% distributions of the parameters estimated in the first step.** The distributions of these parameters (see Table S1 for definitions) were computed using wild bootstrap (Kline and Santos, 2012; Liu, 1988); *bootstraped* data were generated with residuals perturbed by an i.i.d. sequence of n random weights $\{W_i\}_{i=1}^{n}$ following Mammen's 2-points distribution (that is, $(1-\sqrt{5})/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $(1+\sqrt{5})/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$), which satisfies $\mathbb{E}(W_i)=0$ and $\mathbb{E}(W_i^2)=1$. We only represented values between the 2.5% quantile and the 97.5% quantile. The vertical dashed lines indicate the best (minimum) WLS estimates computed from the original data. Using Rademacher distribution (that is 1 or -1 equiprobably) instead yields very similar results (not shown).

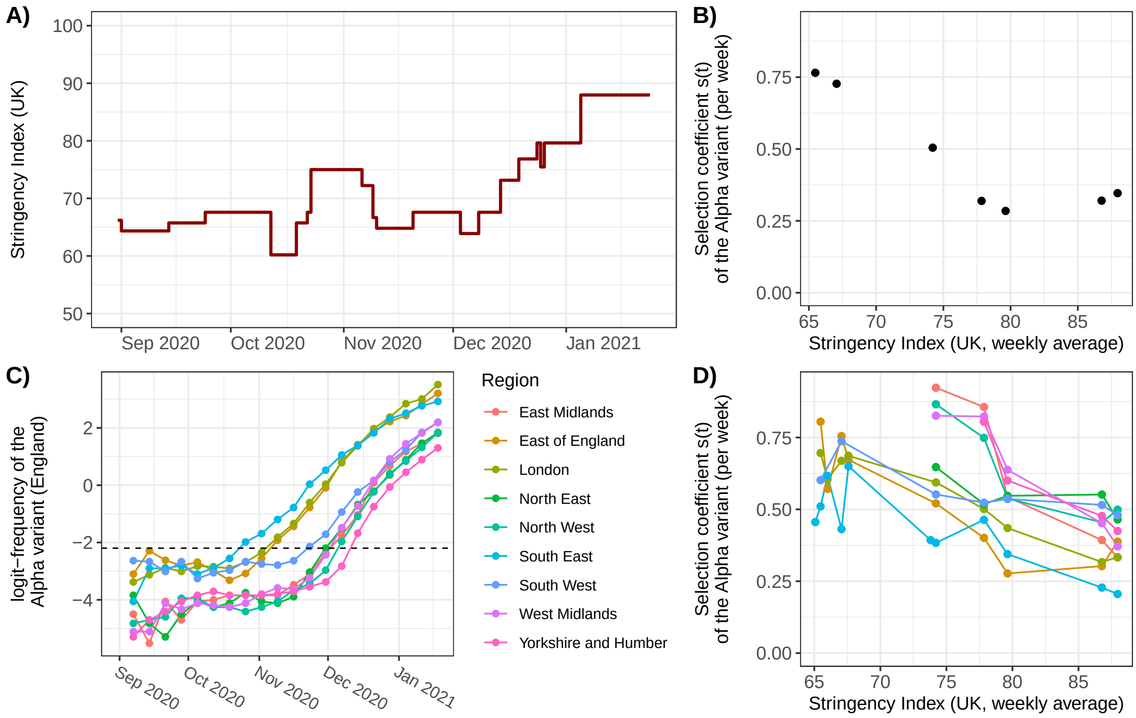


Figure S6: **The selection coefficient of the Alpha variant in England is negatively correlated with the Stringency Index in the UK (fall - winter 2020-2021).** (**A**) Daily values of the Stringency Index in the UK; (**C**) Temporal dynamics of the logit-frequency of SGTF, used as a proxy of the Alpha variant, for each region of England (the black horizontal dashed line indicates the threshold frequency 10%); (**B-D**) Selection coefficient $s(t)$ (per week) of the Alpha variant – i.e. slope of its logit-frequency over time - at the national scale (**B**) and by region (**D**) against the Stringency Index (weekly average) – only frequencies greater than or equal to 10% (above the threshold in **C**) were considered. The correlation between $s(t)$ and the Stringency Index at the national scale is -0.884 (95% CI [-0.983; -0.390]) with a significance test yielding a *p-value* of 8.33×10^{-3}. Correlations at the regional scale are: East Midlands: -0.948 (95% CI [-0.997; -0.400], *p-value* = 0.0142), East of England: -0.868 (95% CI [-0.972; -0.483], *p-value* = 2.39×10^{-3}), London: -0.969 (95% CI [-0.994; -0.857], *p-value* = 1.61×10^{-5}), North East: -0.885 (95% CI [-0.987; -0.263], *p-value* = 0.0189), North West: -0.899 (95% CI [-0.993; -0.080], *p-value* = 0.0381), South East: -0.846 (95% CI [-0.959; -0.499], *p-value* = 1.04×10^{-3}), South West: -0.809 (95% CI [-0.971; -0.142], *p-value* = 0.0276), West Midlands: -0.968 (95% CI [-0.998; -0.588], *p-value* = 6.8×10^{-3}), Yorkshire and Humber: -0.931 (95% CI [-0.999; 0.289], *p-value* =0.0694).

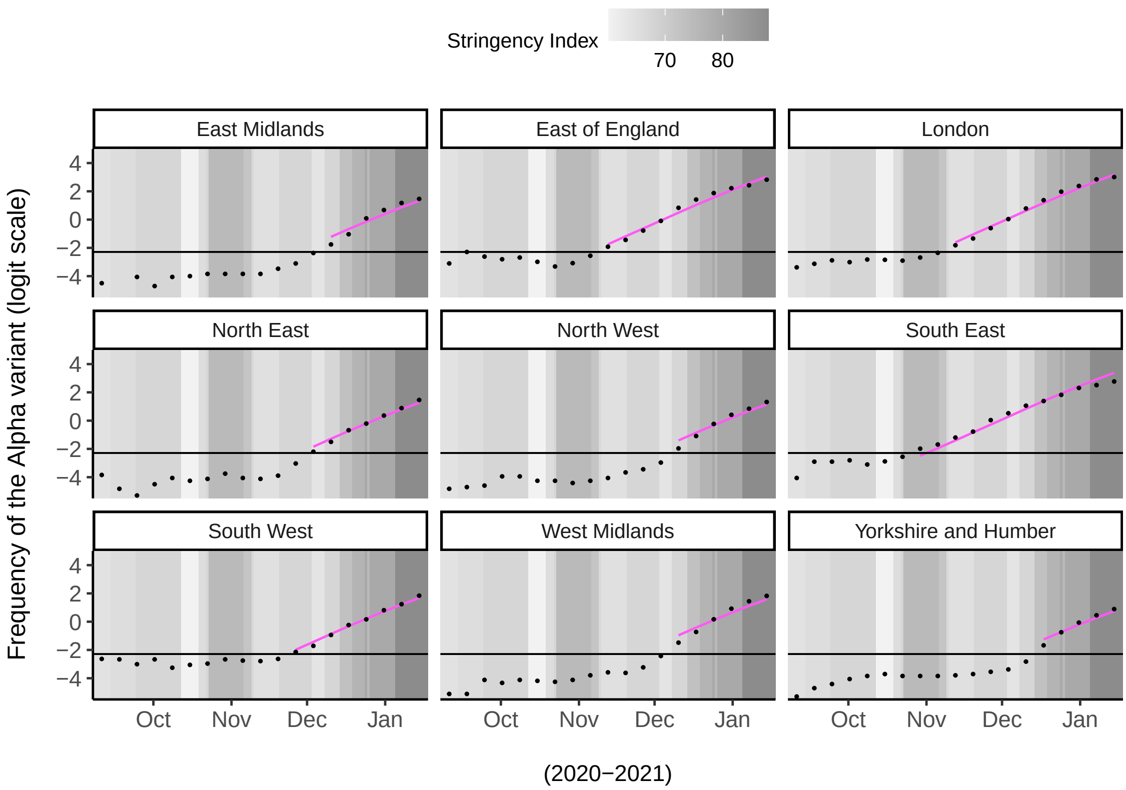


Figure S7: **Model fitting to logit-frequencies of the Alpha variant in the nine regions of England.** Frequencies of SGTF (black points, on logit scale) were used as a proxy for the Alpha variant. The magenta curbs show the fitted values based on a linear MEM with $\Delta\beta$ and $\Delta\gamma$ as fixed effects and the region as a random effect on the intercept of the model. We only fitted frequencies greater than or equal to the threshold frequency 10% (horizontal black lines) in order to get rid of the more stochastic part of these temporal dynamics (when the variant was not yet really well established in the host population).

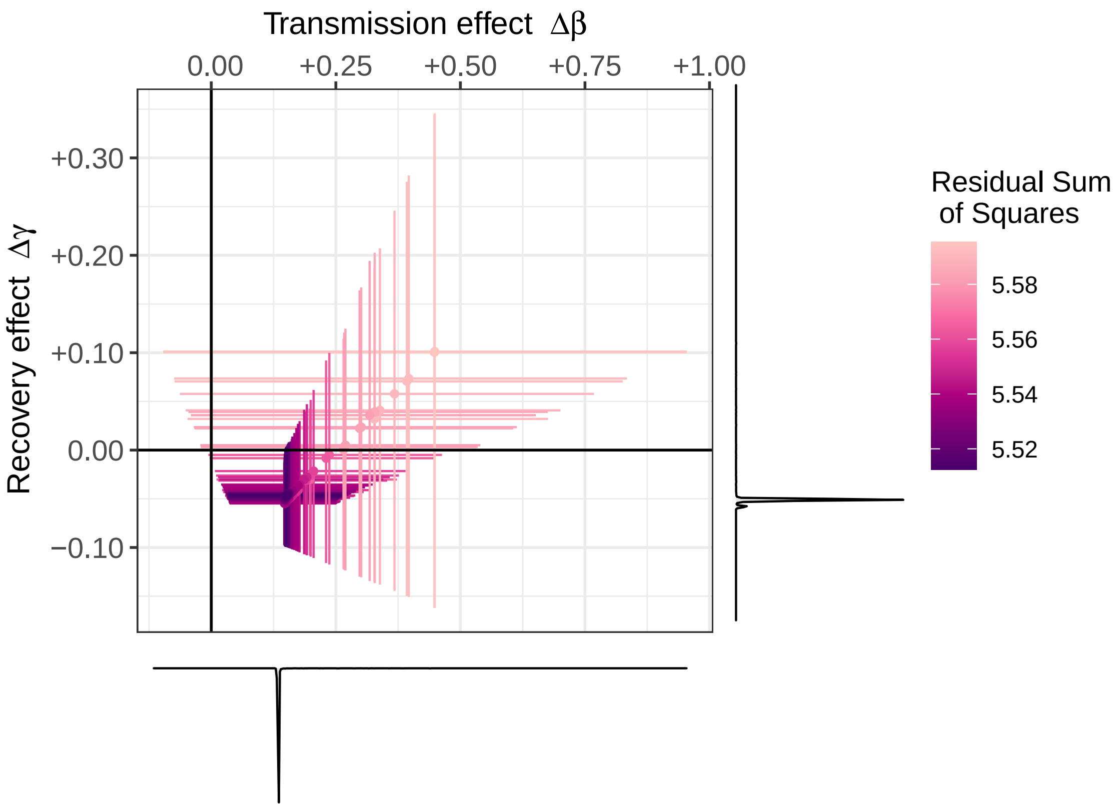


Figure S8: **Phenotypic profile of the Alpha variant (transmission and recovery rates) relative to the previous lineage with uncertainty on the control parameters propagated from the first step.** Estimates per day (points) and 95% confidence interval (crosses) are colored according to the residual sum of squares. We set $S/N = 0.75$, $\kappa = 0.2$, $\beta_w = 0.25$ and $\gamma_w = 0.1$ and each of the almost 2000 points corresponds to a pair $\{k; a\}$ that was previously estimated with wild bootstrap in the first step. Side curbs, representing the densities of the estimates of parameters $\Delta\beta$ (bottom) and $\Delta\gamma$ (right), show that the vast majority of these estimates are grouped around very similar values (dark purple). Indeed, for $\Delta\beta$, 95% of them are between 0.147 and 0.153, for which each corresponding 95% CI remain positive, while, for $\Delta\gamma$, 95% are between -0.054 and -0.046, among which 61% of the corresponding 95% CIs cross the zero axis. With these estimates of $\Delta\beta$ and $\Delta\gamma$, the selection coefficient $s(t)$ of the Alpha variant in the absence of NPI was computed, on average, around 0.11 per day (standard deviation: 0.003), that is 0.77 per week (standard deviation: 0.023).

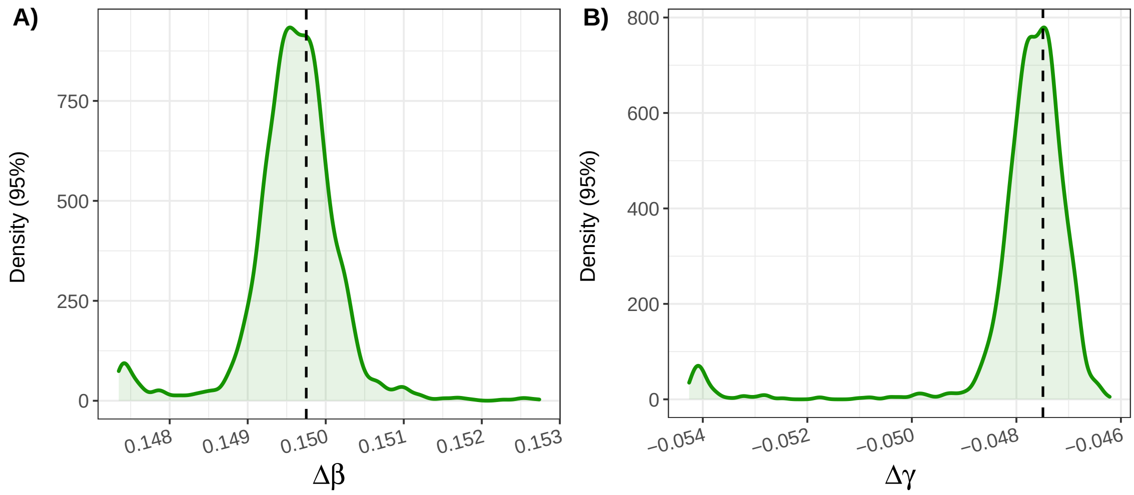


Figure S9: **95% distributions of the inferred phenotypic differences $\Delta\beta$ (transmission effect) and $\Delta\gamma$ (recovery effect) for the Alpha variant relative to the previous lineage.** Each value (per day) of $\Delta\beta$ and $\Delta\gamma$ was computed using a linear MEM and a particular pair $\{k; a\}$ that was previously estimated with wild bootstrap in the first step. Only values between the 2.5% quantile and the 97.5% quantile are represented. Vertical dashed lines indicate the estimates of $\Delta\beta$ and $\Delta\gamma$ using the best WLS estimates for parameters k and a.

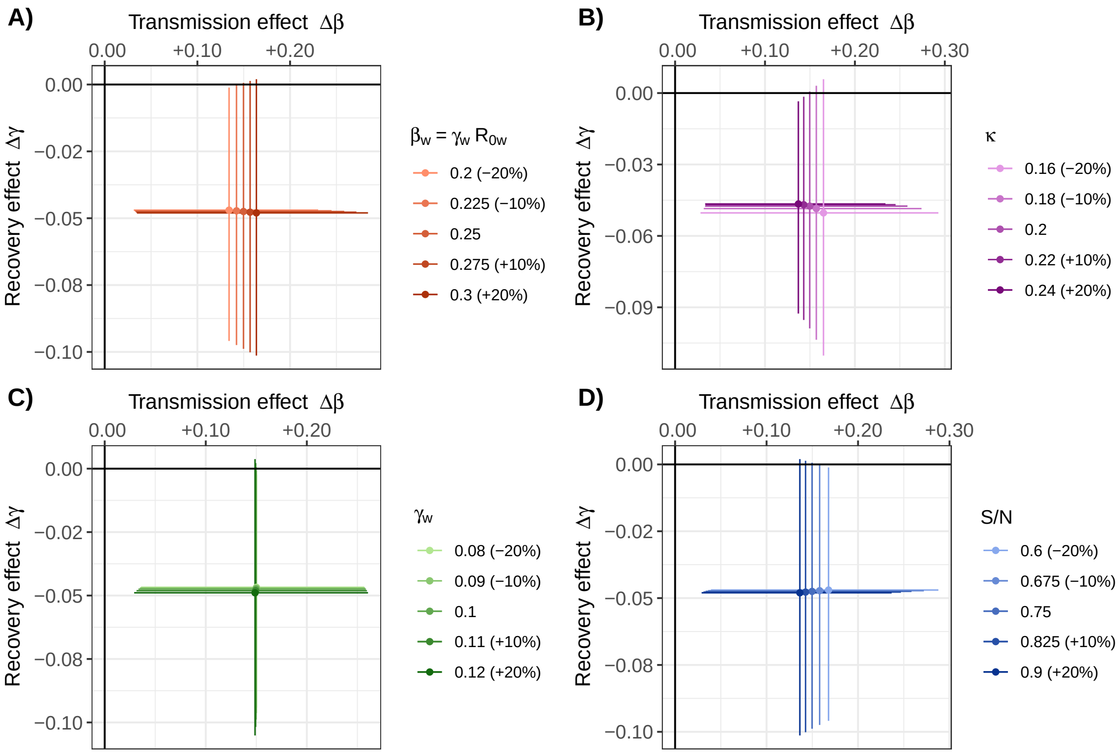


Figure S10: **Phenotypic differences between the Alpha variant and the resident strain with small variations in the fixed parameters.** $\pm 10\%$ and $\pm 20\%$ perturbations were applied separately to each fixed parameter to investigate robustness: $\beta_w = 0.25$ (A), $\kappa = 0.2$ (B), $\gamma_w = 0.1$ (C) and $S/N = 0.75$ (D). Keeping our best WLS estimates for control parameters k and a ($k = 1$ and $a = 3.78$), we reiterated MEMs to obtain new estimates (points) and 95% CIs (segments) of the phenotypic differences $\Delta\beta$ – transmission effect – and $\Delta\gamma$ – recovery effect – between the Alpha variant and the resident strain.

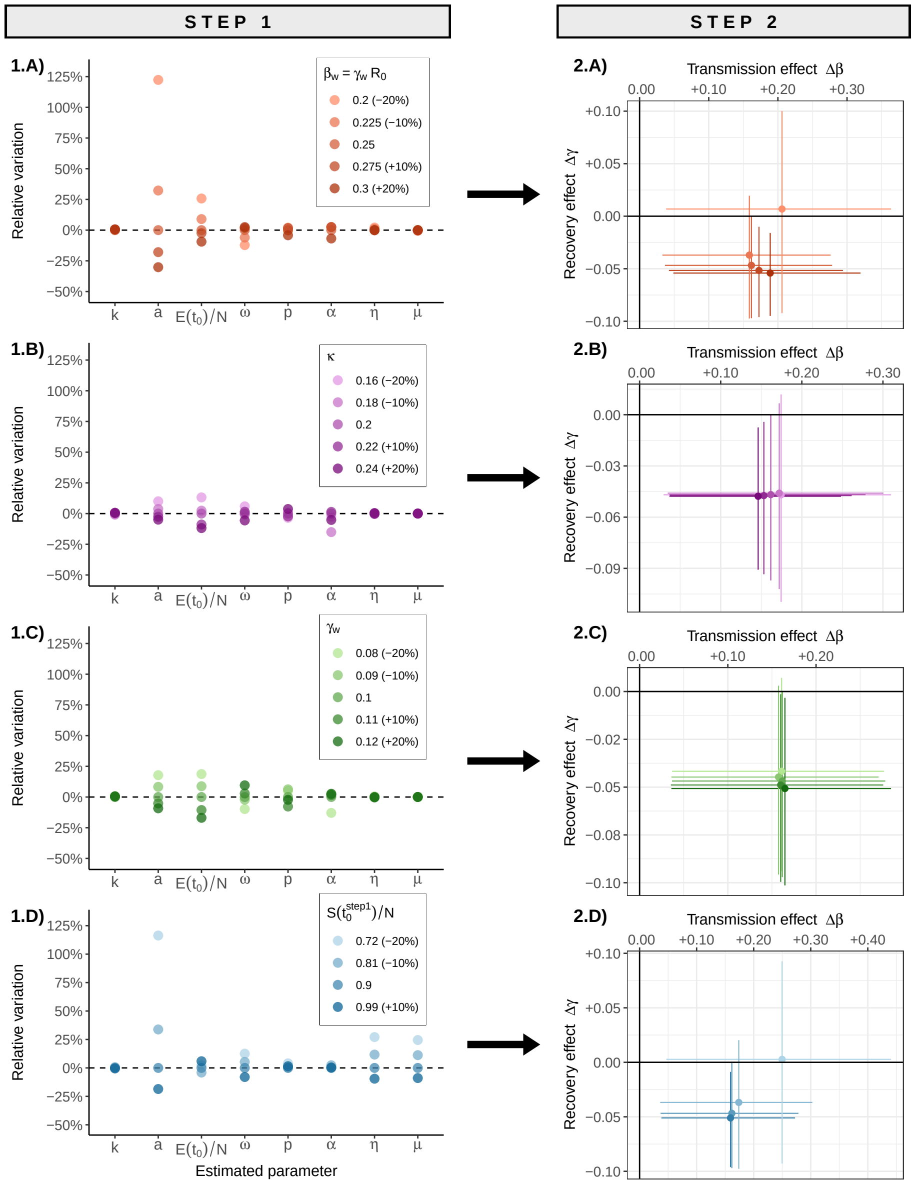


Figure S11: **Effects of small variations in the fixed parameters propagated through the two-step approach.** ±10% and ±20% perturbations were applied separately to each fixed parameter: (A) $\beta_w = 0.25$ (transmission rate of the WT), (B) $\kappa = 0.2$ (transition rate from E (exposed state) to I (infectious state); same for both strains), (C) $\gamma_w = 0.1$ (recovery rate of the WT) and (D) $S(t_0^{\text{step 1}})/N = 0.9$ (initial proportion of susceptible hosts where $t_0^{\text{step 1}}$ refers to the initial time point in step 1). (1) In the first step, each point corresponds to the best estimation (lowest WLS value) from 500 non-linear optimizations starting from uniformly drawn initial conditions (cf. **Table S2**); relative variations (y-axis) refer to the percentage of variation between the parameters estimated with perturbations and those without. (2) For the second step, we reiterated MEMs using estimates from (1) for the control parameters k and a along with the same corresponding ±10% and ±20% perturbations to obtain new estimates (points) and 95% CIs (segments) for the phenotypic differences $\Delta\beta$ and $\Delta\gamma$ (S/N was set in (2) consistently with the end of each simulation in (1)).

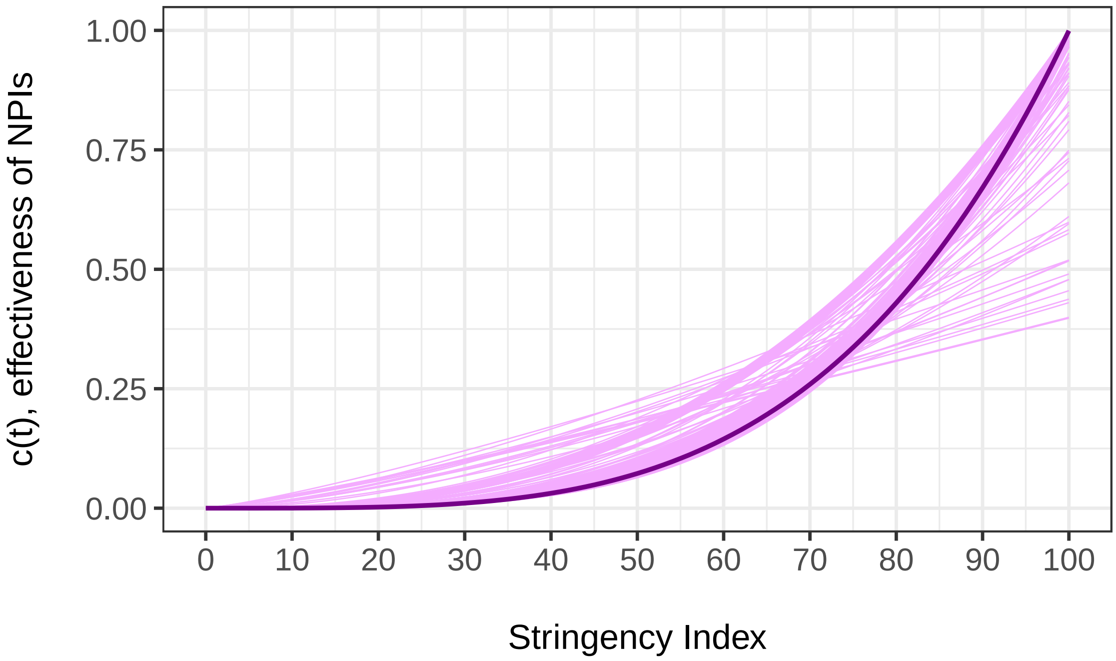


Figure S12: **Inferred relationship between the Stringency Index and the effectiveness of NPIs in the UK.** The link between the Stringency Index $\psi(t)$ and the effectiveness of NPIs $c(t)$ is modeled through the following function: $c(t) = k\,(\psi(t)/100)^a$. Depending on the value of a, this relationship may be concave ($0 < a < 1$), linear ($a = 1$) or convex ($a > 1$). The aim of the first step of our analysis is to infer the value of parameters k and a. The best WLS estimator yielded $k = 1$ and $a = 3.78$ (dark line); a joint distribution for these two parameters (light lines) was obtained using wild bootstrap computations. With all our estimates of a greater than 1, we always find a (more or less pronounced) convex relationship between the Stringency Index and the effectiveness of NPIs in the UK.

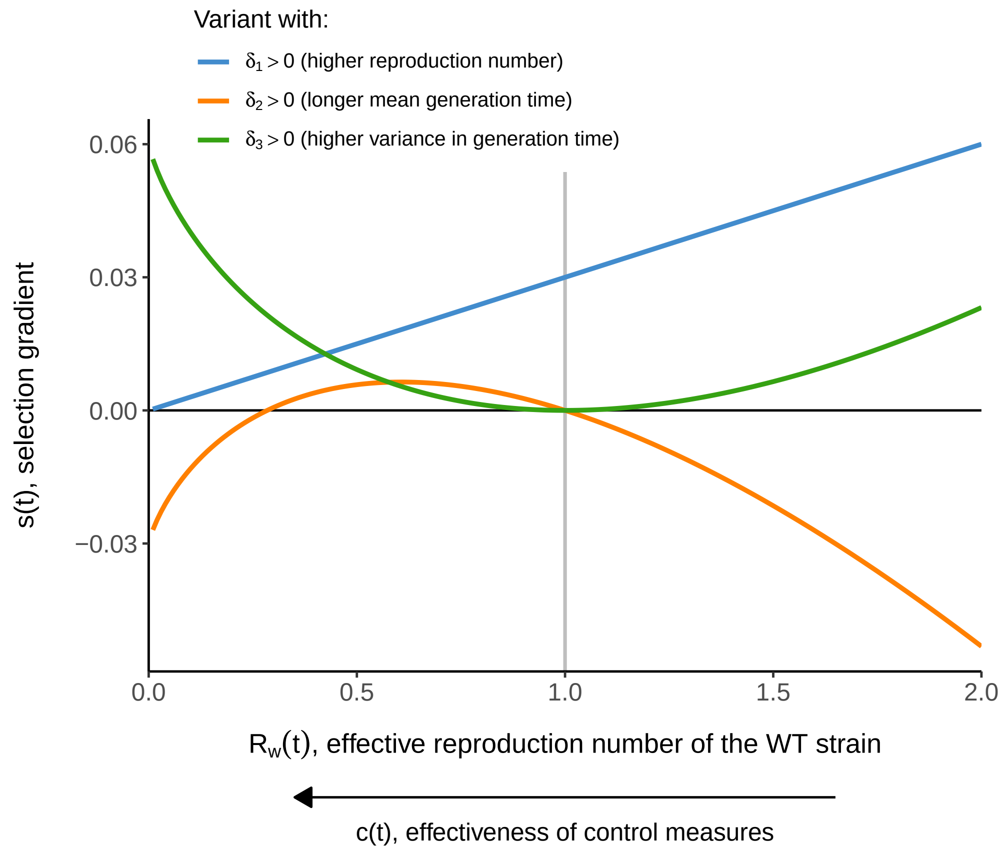


Figure S13: **Variation of the selection gradient of three types of variant as a function of the effective reproduction number of the resident strain (WT).** In (Blanquart et al., 2022), the authors propose that an emerging variant m may differ phenotypically from the WT strain w by its effective reproduction number $\mathcal{R}_m(t) = \mathcal{R}_w(t)\big(1+\delta_1\big)$ and/or by its mean generation time $\mu_m = \mu_w(t)\big(1+\delta_2\big)$ and/or by the variance of its generation time $\sigma_m^2 = \big(\sigma_w(t)\big(1+\delta_3\big)\big)^2$ (we keep the same notations as in the original article (Blanquart et al., 2022) for the phenotypic differences δ_1, δ_2 and δ_3). Generation times are assumed to be gamma-distributed and, in particular, exponentially distributed for the resident strain (special case of gamma distribution where $\sigma_w = \mu_w$) with $\mu_w = 10$. The selection gradient is computed under the assumption of weak selection and, for each curb, phenotypic differences are: $\big(\delta_1 = +30\%,\ \delta_2 = 0,\ \delta_3 = 0\big)$, $\big(\delta_1 = 0,\ \delta_2 = +30\%,\ \delta_3 = 0\big)$ and $\big(\delta_1 = 0,\ \delta_2 = 0,\ \delta_3 = +30\%\big)$, respectively. The variant is selected when its selection gradient is positive – conversely, counter-selected when it is negative. The case $\mathcal{R}_w(t) = 1$ (vertical grey line) corresponds to a stable epidemic. At the bottom of the figure, the horizontal arrow pointing to the left symbolizes more explicitly the impact of NPIs (control measures) that reduces $\mathcal{R}_w(t)$, therefore altering the selection gradient.

Table S1: **Summary of the parameters involved in the model of the first phase.** The first phase is the period that took place just before the emergence of the Alpha variant. $t_0^{\text{step 1}}$ refers to the time point at which the model is initialized.

Symbol	Description	Value
k	Maximum achievable efficacy of NPIs	estimated
a	'*Shape*' parameter for the relationship between the efficacy of NPIs and the Stringency Index	estimated
$E(t_0^{\text{step 1}})$	Initial number of exposed hosts	estimated
α	Virulence (*per capita* disease-induced mortality rate)	estimated
ω	Probability of symptom development	estimated
p	Probability to die for symptomatic hosts	estimated
η	Slope of the increase in screening effort over time	estimated
μ	Intercept of the increase in screening effort over time	estimated
$\mathcal{R}_0$	Basic reproduction number	2.5 (Ferguson et al., 2020; Kucharski et al., 2020; Li et al., 2020)
γ	*Per capita* recovery rate	0.1 day^{-1} (Byrne et al., 2020)
β	*Per capita* transmission rate	$\gamma\mathcal{R}_0 = 0.25$ day^{-1}
κ	Transition rate from exposed to infectious state	0.2 day^{-1} (Ding et al., 2021)
N	2020 UK population size	$\approx$ 67.9 million
$S(t_0^{\text{step 1}})/N$	Initial proportion of susceptible indivuals	0.9
$I_{Sd}(t_0^{\text{step 1}})$	Initial number of symptomatic hosts that will eventually die	$\frac{D(t_0^{\text{step 1}}+1)-D(t_0^{\text{step 1}})}{\alpha}$
$I_{Sr}(t_0^{\text{step 1}})$	Initial number of symptomatic hosts that will eventually recover	$\left(\frac{1-p}{p}\right) I_{Sd}(t_0^{\text{step 1}})$
$I_A(t_0^{\text{step 1}})$	Initial number of asymptomatic individual	$\left(\frac{1-\omega}{\omega}\right) I_S(t_0^{\text{step 1}})$
$R(t_0^{\text{step 1}})$	Initial number of recovered (immune) hosts	$N - S(t_0^{\text{step 1}}) - E(t_0^{\text{step 1}}) - I(t_0^{\text{step 1}})$

Table S2: **Summary of the parameters involved in the model of the second phase.** The second phase is the period that took place just after the emergence of the Alpha variant.

Symbol	**Description**	**Value**
$\Delta\beta$	Phenotypic difference between the Alpha variant and the resident strain in terms of transmission rate	estimated
$\Delta\gamma$	Phenotypic difference between the Alpha variant and the resident strain in terms of recovery rate	estimated
k	Maximum achievable efficacy of NPIs	estimates from the first step (best WLS estimate: 1)
a	'*Shape*' parameter for the relationship between the efficacy of NPIs and the Stringency Index	estimates from the first step (best WLS estimate: 3.78)
γ_w	*Per capita* recovery rate of the resident strain	0.1 day^{-1} (Byrne et al., 2020)
β_w	*Per capita* transmission rate of the resident strain	0.25 day^{-1}
κ	Transition rate from exposed to infectious state	0.2 day^{-1} (Ding et al., 2021)
S/N	Proportion of susceptible host (assumed constant for short enough periods of time during a controlled epidemic)	0.75

Table S3: **Summary of the initialization and optimization sets for the parameters estimated in the model of the first phase.** Because of the presence of local minima, optimization procedure was repeated for 1500 sets of initial values by drawing randomly in each initialization interval below according to a continuous uniform distribution. Parameter transformations enabled then to restrict optimization searches in more relevant ranges of values (referred to '*optimization intervals*' below). You may refer to the **Table S1** for the meaning of the symbols.

Symbol	**Initialization interval**	**Optimization interval**
k	$[0; 1]$	$[0; 1]$
a	$[0; 10]$	$\mathbb{R}_+^*$
$E(t_0^{\text{step } 1})$	$[0; 10^{-2}]$	$[0; 10^{-2}]$
α	$[0; 1]$	$[0; 1]$
ω	$[0.2; 0.8]$	$[0; 1]$
p	$[0; 0.1]$	$[0; 0.2]$
η	$[0; 10^{-4}]$	$[0; 1]$
μ	$[0; 10^{-2}]$	$[0; 1]$

Table S4: **Likelihood-based comparisons of nested linear MEM.** The tilde operator $\sim$ refers to the linear relationship between the response variable $\mathrm{logit}(\widetilde{f}_m(t))$, the logit-frequency of the variant, and the explanatory variables. The phenotypic differences between the variant and the resident strain $\Delta\beta$ (transmission effect) and $\Delta\gamma$ (recovery effect) are considered as fixed effects while $(1|Region)$ refers to a random effect of the region on the intercept of the model. With a significance level of 5%, results show a significant effect (*) for $\Delta\beta$ but not (.) for $\Delta\gamma$ (although the *p-value* associated with the latter is very close to the significance threshold). AIC is the Akaike Information Criterion and BIC is the Bayesian Information Criterion (models with lower values are preferred).

			$\mathrm{logit}(\widetilde{f}_m(t)) \sim \Delta\beta + \Delta\gamma + (1 \mid Region)$
	AIC	BIC	AIC 61.122 BIC 72.969
$\mathrm{logit}(\widetilde{f}_m(t)) \sim \Delta\beta + (1 \mid Region)$	62.867	72.345	*p-value* $= 0.053$ (.)
$\mathrm{logit}(\widetilde{f}_m(t)) \sim \Delta\gamma + (1 \mid Region)$	65.386	74.864	*p-value* $= 0.012$ $(*)$

References for supplementary figures and tables

Blanquart, F., Hozé, N., Cowling, B. J., Débarre, F., & Cauchemez, S. (2022). Selection for infectivity profiles in slow and fast epidemics, and the rise of sars-cov-2 variants. *eLife*. https://doi.org/10.7554/eLife.75791

Byrne, A. W., McEvoy, D., Collins, A. B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E. A., McAloon, C., O'Brien, K., Wall, P., Walsh, K. A., & More, S. J. (2020). Inferred duration of infectious period of sars-cov-2: Rapid scoping review and analysis of available evidence for asymptomatic and symptomatic covid-19 cases. *BMJ Open*, *10*(8). https://doi.org/10.1136/bmjopen-2020-039856

Ding, Z., Wang, K., Shen, M., Wang, K., Zhao, S., Song, W., Li, R., Li, Z., Wang, L., Feng, G., Hu, Z., Wei, H., Xiao, Y., Bao, C., Hu, J., Zhu, L., Li, Y., Chen, X., Yin, Y., ... Shen, H. (2021). Estimating the time interval between transmission generations and the presymptomatic period by contact tracing surveillance data from 31 provinces in the mainland of china. *Fundamental Research*, *1*(2), 104–110. https://doi.org/10.1016/j.fmre.2021.02.002

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., van Elsland, S., ... Ghani, A. C. (2020). *Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand*. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf

Kline, P. M., & Santos, A. (2012). A score based approach to wild bootstrap inference. *J. Econom. Methods*, *1*(1), 23–41. https://doi.org/10.1515/2156-6674.1006

Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., & Eggo, R. M. (2020). Early dynamics of transmission and control of covid-19: A mathematical modelling study. *Lancet Infect. Dis.*, *20*, 553–558. https://doi.org/10.1016/S1473-3099(20)30144-4

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., ... Feng, Z. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *N. Engl. J. Med.*, *382*(13), 1199–1207. https://doi.org/10.1056/NEJMoa2001316

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Stat.*, *16*(4), 1696–1708. https://doi.org/10.1214/aos/1176351062

Raue, A., Kreutz, C., Maiwald, T., Bachman, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, *25*(15), 1923–1929. https://doi.org/10.1093/bioinformatics/btp358

— Supplementary Information (SI Appendix) —

Phenotypic evolution of SARS-CoV-2: a statistical inference approach

Wakinyan Benhamou iD, Sébastien Lion iD, Rémi Choquet† iD and Sylvain Gandon† iD

CEFE, CNRS, Univ Montpellier, EPHE, IRD, Montpellier, France
†: equal contribution

July 5, 2023

Preamble

In this appendix, we show how phenotypic traits of the new variant affect its temporal dynamics. We derive the selection coefficient (or selection gradient) of a variant – a measure of how much it is favoured or disfavoured through natural selection – in a susceptible-exposed-infectious-recovered ($SEIR$) model. In sections S2-S5 we show how we can use a weak selection argument to obtain a useful approximation of the selection coefficient of the new variant. In section S6 we use a weak selection argument to derive an approximation of the differentiation of variant frequency between the exposed and the infectious states. Finally, in section S7, we detail how our analysis relates to the approach used in (Blanquart et al., 2022).

S1 $SEIR$ model

Let us consider a directly and horizontally transmitted disease and a host population of size N for which individuals are either susceptible (S), exposed (E, infected but not yet infectious), infectious (I) or recovered (R). For a given state, for instance S, we denote $S(t)$, where t is the current time, its density and $\dot{S}(t)$, its differentiation with respect to time. Demographic parameters (newborns, migration balance, natural mortality, ...) are neglected.

Let also consider a polymorphic pathogen population: the WT strain (resident), which will be represented with the subscript w, and the mutant strain (or variant), which will be represented with the subscript m (we then neglect any occurrence of new mutations). Therefore, $E(t)$ and $I(t)$ can be respectively decomposed into: $E(t) = E_w(t) + E_m(t)$ and $I(t) = I_w(t) + I_m(t)$. The variant may differ phenotypically from the WT in its effective transmission rate $\beta_m = \beta_w + \Delta\beta$ and/or in its recovery rate $\gamma_m = \gamma_w + \Delta\gamma$ and/or in its disease-induced mortality rate (virulence) $\alpha_m = \alpha_w + \Delta\alpha$ and/or in its transition rate from state E to state I $\kappa_m = \kappa_w + \Delta\kappa$ (note that $1/\kappa$ is thus the mean sojourn time in the exposed state, i.e. the latent period).

Besides, the transmission of both strains is more or less affected depending on $c(t)$, the effectiveness of governmental control measures – i.e. Non-Pharmaceutical Interventions (NPIs) – implemented at time t to mitigate the spread of the epidemic.

We also assume that superinfections do not occur – including co-infections with both strains – and that (persistent) immunity acquired with either strain protects effectively against both. We model the temporal dynamics of this $SEIR$ system with the following set of ordinary differential equations (ODEs):

$$\begin{cases} \dot{S}(t) = -(1-c(t))\overline{\beta}(t)S(t)\dfrac{I(t)}{N} \\ \dot{E}(t) = (1-c(t))\overline{\beta}(t)S(t)\dfrac{I(t)}{N} - \overline{\kappa}(t)E(t) \\ \dot{I}(t) = \overline{\kappa}(t)E(t) - \Big(\overline{\alpha}(t) + \overline{\gamma}(t)\Big)I(t) \\ \dot{R}(t) = \overline{\gamma}(t)I(t) \end{cases} \tag{S1}$$

The overlines refer to mean values of the phenotypic traits after averaging over the distribution of strain frequencies in the relevant compartments of the model:

$$\begin{cases} \overline{\kappa}(t) = (1-p_m(t))\kappa_w + p_m(t)\kappa_m \\ \overline{\beta}(t) = (1-q_m(t))\beta_w + q_m(t)\beta_m \\ \overline{\gamma}(t) = (1-q_m(t))\gamma_w + q_m(t)\gamma_m \\ \overline{\alpha}(t) = (1-q_m(t))\alpha_w + q_m(t)\alpha_m \end{cases}$$

where the frequency of the variant m in the compartment E is noted $p_m(t) = E_m(t)/E(t)$ and the frequency of the variant m in the compartment I is noted $q_m(t) = I_m(t)/I(t)$. For each strain i ($i \in \{w; m\}$):

$$\begin{cases} \dot{E}_i(t) = (1-c(t))\beta_i S(t)\dfrac{I_i(t)}{N} - \kappa_i E_i(t) \\ \dot{I}_i(t) = \kappa_i E_i(t) - \Big(\alpha_i + \gamma_i\Big)I_i(t) \end{cases}$$

By noting $\mathbf{X}(t) = \begin{pmatrix} E(t) & I(t) \end{pmatrix}^\top$, we have:

$$\dot{\mathbf{X}}(t) = \overline{\mathbf{R}}(t)\mathbf{X}(t)$$

with $\overline{\mathbf{R}}(t) = \begin{pmatrix} -\overline{\kappa}(t) & (1-c(t))\overline{\beta}(t)\frac{S(t)}{N} \\ \overline{\kappa}(t) & -\overline{\alpha}(t) - \overline{\gamma}(t) \end{pmatrix}$, the matrix of (average) transitions rates.

S2 Overall frequency of the variant

The overall frequency of the variant in the system at time t, $f_m(t)$, is:

$$f_m(t) = \frac{E_m(t) + I_m(t)}{E(t) + I(t)} = \frac{p_m(t)E(t) + q_m(t)I(t)}{E(t) + I(t)} = p_m(t)p(t) + q_m(t)q(t)$$

where $p(t) = E(t)/(E(t) + I(t))$ and $q(t) = I(t)/(E(t) + I(t))$ are the class frequencies of infected individuals in the exposed state E and in the infectious state I, respectively. Note therefore that $p(t) + q(t) = 1$.

The temporal dynamics of the variant can be tracked more conveniently using the following quantity (Lion, 2018; Lion & Gandon, 2022):

$$\widetilde{f}_m(t) = p_m(t)v^E(t)p(t) + q_m(t)v^I(t)q(t)$$

where, $v^E(t)$ and $v^I(t)$ are the reproductive values in state E and in state I, respectively.
Let $\mathbf{v}(t) = \begin{pmatrix} v^E(t) & v^I(t) \end{pmatrix}^\top$ be the vector of reproductive values and $\mathbf{f}(t) = \begin{pmatrix} p(t) & q(t) \end{pmatrix}^\top$ be the vector of frequencies within infected states; these two vectors are co-normalized such that $\mathbf{v}^\top\mathbf{f} = 1$. Under the assumption of weak selection, $\widetilde{f}_m(t)$ yields indeed a really useful expression:

$$\frac{\mathrm{d}\widetilde{f}_m(t)}{\mathrm{d}t} = \underbrace{\widetilde{f}_m(t)(1 - \widetilde{f}_m(t))}_{\text{Genetic variance}} \underbrace{\mathbf{v}(t)^\top \Delta\mathbf{R}(t)\mathbf{f}(t)}_{\text{s(t), selection coefficient}}, \tag{S2}$$

or more simply, using the logit function, that is ln(frequency of the variant / frequency of the WT strain):

$$s(t) = \frac{\mathrm{d}\ \mathrm{logit}(\widetilde{f}_m(t))}{\mathrm{d}t} = \mathbf{v}(t)^\top \Delta\mathbf{R}(t)\mathbf{f}(t). \tag{S3}$$

$\Delta\mathbf{R}(t)$ is the matrix of differences in transition rates between the mutant and the WT strains such that:

$$\Delta\mathbf{R}(t) = \mathbf{R}_m(t) - \mathbf{R}_w(t) = \begin{pmatrix} -\Delta\kappa & (1 - c(t))\Delta\beta\frac{S(t)}{N} \\ \Delta\kappa & -\Delta\alpha - \Delta\gamma \end{pmatrix}.$$

The selection coefficient $s(t)$ is thus given by:

$$\begin{aligned} s(t) &= \mathbf{v}(t)^\top \Delta\mathbf{R}(t)\mathbf{f}(t) \\ &= p(t)\Delta\kappa\Big(v^I(t) - v^E(t)\Big) + q(t)\left[(1 - c(t))\Delta\beta\frac{S(t)}{N}v^E(t) - \Big(\Delta\alpha + \Delta\gamma\Big)v^I(t)\right]. \end{aligned} \tag{S4}$$

The main problem with this theoretical expression is that class frequencies ($p(t)$ and $q(t)$) and reproductive values ($v^E(t)$ and $v^I(t)$) are generally not available from public health data. In the following section we show how we can derive useful approximations for these quantities.

S3 Class frequencies within infected states & growth rate

Following (Lion, 2018; Lion & Gandon, 2022), the temporal dynamics of the vector of class frequencies is given by:

$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \overline{\mathbf{R}}(t)\mathbf{f}(t) - \overline{r}(t)\mathbf{f}(t)$$

with $\overline{r}(t)$, the growth rate of the epidemic:

$$\overline{r}(t) \;=\; \mathbf{1}^\top\overline{\mathbf{R}}(t)\mathbf{f}(t) \;=\; q(t)\Big((1-c(t))\overline{\beta}(t)\frac{S(t)}{N} - \overline{\alpha}(t) - \overline{\gamma}(t)\Big). \tag{S5}$$

Therefore:

$$\begin{cases} \dot{p}(t) &= q(t)(1-c(t))\overline{\beta}(t)\dfrac{S(t)}{N} - p(t)\Big(\overline{\kappa}(t) + \overline{r}(t)\Big) \\ \dot{q}(t) &= p(t)\overline{\kappa}(t) - q(t)\Big(\overline{\alpha}(t) + \overline{\gamma}(t) + \overline{r}(t)\Big) \end{cases} \tag{S6}$$

When the difference between the vital rates of the mutant vs. resident is small and $\mathcal{O}(\varepsilon)$ (i.e. weak selection), the dynamics of the frequency $\tilde{f}_m(t)$ is also $\mathcal{O}(\varepsilon)$ (equation (S2)) while equation (S6) is $\mathcal{O}(1)$. This implies that $p(t)$ and $q(t)$ can be treated as fast variables while $\widetilde{f}_m(t)$ is a slow variable. Using a quasi-equilibrium approximation, i.e. setting the right-hand sides of (S6) to 0, we then have:

$$\begin{cases} q(t)(1-c(t))\overline{\beta}(t)\dfrac{S(t)}{N} &= p(t)\Big(\overline{\kappa}(t) + \overline{r}(t)\Big) \\ p(t)\overline{\kappa}(t) &= q(t)\Big(\overline{\alpha}(t) + \overline{\gamma}(t) + \overline{r}(t)\Big) \end{cases}$$

which yields:

$$\frac{p(t)}{q(t)} = \frac{(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}}{\overline{\kappa}(t) + \overline{r}(t)} = \frac{\overline{\alpha}(t) + \overline{\gamma}(t) + \overline{r}(t)}{\overline{\kappa}(t)}. \tag{S7}$$

In addition,

$$\frac{p(t)}{q(t)} = \frac{\overline{\alpha}(t) + \overline{\gamma}(t) + \overline{r}(t)}{\overline{\kappa}(t)} \qquad \Longleftrightarrow \qquad \frac{1-q(t)}{q(t)} = \frac{\overline{\alpha}(t) + \overline{\gamma}(t) + q(t)\Big((1-c(t))\overline{\beta}(t)\frac{S(t)}{N} - \overline{\alpha}(t) - \overline{\gamma}(t)\Big)}{\overline{\kappa}(t)}$$

$$\Longleftrightarrow \qquad q(t)^2\Big((1-c(t))\overline{\beta}(t)\frac{S(t)}{N} - \overline{\alpha}(t) - \overline{\gamma}(t)\Big) + q(t)\Big(\overline{\kappa}(t) + \overline{\alpha}(t) + \overline{\gamma}(t)\Big) - \overline{\kappa}(t) = 0$$

Only the following solution satisfies $q(t) \in [0,1]$:

$$q(t) = \frac{-\Big(\overline{\kappa}(t) + \overline{\alpha}(t) + \overline{\gamma}(t)\Big) + \sqrt{\Big(\overline{\kappa}(t) - \overline{\alpha}(t) - \overline{\gamma}(t)\Big)^2 + 4\overline{\kappa}(t)(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}}}{2\Big((1-c(t))\overline{\beta}(t)\frac{S(t)}{N} - \overline{\alpha}(t) - \overline{\gamma}(t)\Big)}$$

Under weak selection and when $q(t)$ is at equilibrium, the growth rate of the epidemic can thus be approximated by:

$$\overline{r}(t) \approx \frac{1}{2}\left[-\overline{\kappa}(t) - \overline{\alpha}(t) - \overline{\gamma}(t) + \sqrt{\Big(\overline{\kappa}(t) - \overline{\alpha}(t) - \overline{\gamma}(t)\Big)^2 + 4\overline{\kappa}(t)(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}}\;\right] \tag{S8}$$

Note that, when $(1-c(t))\overline{\beta}(t)S(t)/N - \overline{\alpha}(t) - \overline{\gamma}(t) = 0$, then $\overline{r}(t) = 0$ which is also consistent for its approximation (S8).

Besides, starting from either the real expression of the growth rate (S5) or its approximation (S8), we have:

$$\lim_{\kappa \to +\infty} \overline{r}(t) = (1-c(t))\overline{\beta}(t)\frac{S(t)}{N} - \overline{\alpha}(t) - \overline{\gamma}(t).$$

In other words, we recover the growth rate of the corresponding nested SIR model as a limit of this model.

S4 Reproductive values within infected states

Following (Lion, 2018; Lion & Gandon, 2022), the temporal dynamics of reproductive values are given by:

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{v}(t)^\top}{\mathrm{d}t} &= -\mathbf{v}(t)^\top\overline{\mathbf{R}}(t) + \overline{r}(t)\mathbf{v}(t)^\top \\
&= -\begin{pmatrix} v^E(t) & v^I(t)\end{pmatrix}\begin{pmatrix} -\overline{\kappa}(t) & (1-c(t))\overline{\beta}(t)\frac{S(t)}{N} \\ \overline{\kappa}(t) & -\overline{\alpha}(t) - \overline{\gamma}(t)\end{pmatrix} + \overline{r}(t)\begin{pmatrix} v^E(t) & v^I(t)\end{pmatrix}.
\end{aligned}$$

Therefore:

$$\begin{cases}
\dfrac{\mathrm{d}v^E(t)}{\mathrm{d}t} &= \overline{\kappa}(t)\Big(v^E(t) - v^I(t)\Big) + \overline{r}(t)v^E(t) \\
\dfrac{\mathrm{d}v^I(t)}{\mathrm{d}t} &= -(1-c(t))\overline{\beta}(t)\dfrac{S(t)}{N}v^E(t) + \Big(\overline{\alpha}(t) + \overline{\gamma}(t) + \overline{r}(t)\Big)v^I(t)
\end{cases} \tag{S9}$$

As previously, we see that the reproductive values are fast variables, so that we can use a quasi-equilibrium approximation. Setting the right-hand sides of (S9) become equal to 0, we obtain:

$$\begin{cases}
\overline{\kappa}(t)v^I(t) &= \Big(\overline{r}(t) + \overline{\kappa}(t)\Big)v^E(t) \\
(1-c(t))\overline{\beta}(t)\dfrac{S(t)}{N}v^E(t) &= \Big(\overline{\alpha}(t) + \overline{\gamma}(t) + \overline{r}(t)\Big)v^I(t)
\end{cases}$$

Which yields:

$$\frac{v^E(t)}{v^I(t)} = \frac{\overline{\kappa}(t)}{\overline{r}(t) + \overline{\kappa}(t)} = \frac{\overline{\alpha}(t) + \overline{\gamma}(t) + \overline{r}(t)}{(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}} \tag{S10}$$

S5 Approximation of the selection coefficient of the variant

Using the quasi-equilibrium approximation for vectors $\mathbf{f}(t)$ and $\mathbf{v}(t)$ (*cf.* (S7) and (S10), respectively), the expression of the selection coefficient in equation (S4) becomes:

$$s(t) = p(t)\Delta\kappa\left(v^E(t)\frac{\overline{r}(t)+\overline{\kappa}(t)}{\overline{\kappa}(t)} - v^E(t)\right) + q(t)\left[(1-c(t))\Delta\beta\frac{S(t)}{N}v^I(t)\frac{\overline{r}(t)+\overline{\alpha}(t)+\overline{\gamma}(t)}{(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}} - \left(\Delta\alpha+\Delta\gamma\right)v^I(t)\right]$$

$$= p(t)v^E(t)\overline{r}(t)\frac{\Delta\kappa}{\overline{\kappa}(t)} + q(t)v^I(t)\left[\frac{\Delta\beta}{\overline{\beta}(t)}\left(\overline{r}(t)+\overline{\alpha}(t)+\overline{\gamma}(t)\right) - \Delta\alpha - \Delta\gamma\right].$$

Since combining quasi-equilibrium approximations (S7) and (S10) yields

$$\frac{p(t)v^E(t)}{q(t)v^I(t)} = \frac{\overline{\alpha}(t)+\overline{\gamma}(t)+\overline{r}(t)}{\overline{\kappa}(t)+\overline{r}(t)}$$

and, using the co-normalization $\mathbf{v}^\top\mathbf{f} = p(t)v^E(t) + q(t)v^I(t) = 1$, then:

$$q(t)v^I(t) = \frac{\overline{\kappa}(t)+\overline{r}(t)}{\overline{\kappa}(t)+\overline{\alpha}(t)+\overline{\gamma}(t)+2\overline{r}(t)} \qquad \text{and} \qquad p(t)v^E(t) = \frac{\overline{\alpha}(t)+\overline{\gamma}(t)+\overline{r}(t)}{\overline{\kappa}(t)+\overline{\alpha}(t)+\overline{\gamma}(t)+2\overline{r}(t)}.$$

Thus:

$$s(t) = \frac{\left(\overline{\alpha}(t)+\overline{\gamma}(t)+\overline{r}(t)\right)\overline{r}(t)\frac{\Delta\kappa}{\overline{\kappa}(t)} + \left(\overline{\kappa}(t)+\overline{r}(t)\right)\left[\frac{\Delta\beta}{\overline{\beta}(t)}\left(\overline{r}(t)+\overline{\alpha}(t)+\overline{\gamma}(t)\right) - \Delta\alpha - \Delta\gamma\right]}{\overline{\kappa}(t)+\overline{\alpha}(t)+\overline{\gamma}(t)+2\overline{r}(t)} \tag{S11}$$

Using (S8) to approximate the growth rate of the epidemic, the selection coefficient of the variant becomes after some rearrangements:

$$s(t) \approx \frac{2(1-c(t))\frac{S(t)}{N}\left(\Delta\kappa\overline{\beta}(t)+\overline{\kappa}(t)\Delta\beta\right) + \Delta\kappa\left(\overline{\kappa}(t)-\overline{\alpha}(t)-\overline{\gamma}(t)-Z(t)\right) - \left(\Delta\alpha+\Delta\gamma\right)\left(\overline{\kappa}(t)-\overline{\alpha}(t)-\overline{\gamma}(t)+Z(t)\right)}{2Z(t)} \tag{S12}$$

with

$$Z(t) = \sqrt{\left(\overline{\kappa}(t)-\overline{\alpha}(t)-\overline{\gamma}(t)\right)^2 + 4\overline{\kappa}(t)(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}}.$$

Here again, we can recover the expression of $s(t)$ from the nested SIR model (Day & Gandon, 2006, 2007) by taking the limit:

$$\lim_{\kappa \to +\infty} s(t) \quad = \quad (1-c(t))\Delta\beta\frac{S(t)}{N} - \Delta\alpha - \Delta\gamma.$$

As in the main text, we now assume that the virulence may be neglected ($\alpha_m = \alpha_w = 0$) and that there is no difference between the variant and the WT strains in terms of latent period, i.e. $\Delta\kappa = 0$.

The approximation (S11) of the selection coefficient reduces then to:

$$s(t) = \frac{\kappa + \overline{r}(t)}{\kappa + \overline{\gamma}(t) + 2\overline{r}(t)} \left[\frac{\Delta\beta}{\overline{\beta}(t)} \left(\overline{r}(t) + \overline{\gamma}(t) \right) - \Delta\gamma \right] \tag{S13}$$

Or, using (S12), i.e. based on an approximation of the growth rate:

$$s(t) \approx \frac{2\kappa(1-c(t))\Delta\beta\frac{S(t)}{N} - \left(\kappa - \overline{\gamma}(t)\right)\Delta\gamma - \sqrt{\left(\kappa - \overline{\gamma}(t)\right)^2 + 4\kappa(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}} \ \ \Delta\gamma}{2\ \sqrt{\left(\kappa - \overline{\gamma}(t)\right)^2 + 4\kappa(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}}} \tag{S14}$$

Since d $\text{logit}(\widetilde{f}_m(t))/\text{d}t = s(t)$, we get the following approximation of $\text{logit}(\widetilde{f}_m(t))$ by integrating the last approximation of $s(t)$ between the time points t_0 and $t = t_0 + \Delta t$:

$$\begin{aligned}\text{logit}(\widetilde{f}_m(t)) \approx \text{logit}(\widetilde{f}_m(t_0)) + \kappa \int_{t_0}^{t} \left(\frac{(1-c(t))S(t)/N}{\sqrt{(\kappa - \overline{\gamma}(t))^2 + 4\kappa(1-c(t))\overline{\beta}(t)S(t)/N}} \right) dt \ \ \Delta\beta \\ - \frac{1}{2}\left[\int_{t_0}^{t} \left(\frac{\kappa - \overline{\gamma}(t)}{\sqrt{(\kappa - \overline{\gamma}(t))^2 + 4\kappa(1-c(t))\overline{\beta}(t)S(t)/N}} \right) dt + \Delta t \right] \Delta\gamma\end{aligned} \tag{S15}$$

Assuming also that $\overline{\gamma}(t) \approx \gamma_w$, $\overline{\beta}(t) \approx \beta_w$ (weak selection) and $S(t)/N \approx S/N$ – i.e. the proportion of susceptible individuals varies sufficiently little throughout the considered time period –, we eventually obtain the expression we used in the main text:

$$\begin{aligned}\text{logit}(\widetilde{f}_m(t)) \approx \text{logit}(\widetilde{f}_m(t_0)) + \kappa \int_{t_0}^{t} \left(\frac{(1-c(t))}{\sqrt{(\kappa - \gamma_w)^2 + 4\kappa(1-c(t))\beta_w S/N}} \right) dt \ \ \Delta\beta\frac{S}{N} \\ - \frac{1}{2}\left[\left(\kappa - \gamma_w\right) \int_{t_0}^{t} \left(\frac{1}{\sqrt{(\kappa - \gamma_w)^2 + 4\kappa(1-c(t))\beta_w S/N}} \right) dt + \Delta t \right] \Delta\gamma\end{aligned} \tag{S16}$$

S6 Differentiation between the exposed and the infectious compartments

In this section, we start with a *SEIR* model in a very general form. The particular transition rates used previously and in the main text will be specified after the study of this general case.

By taking up the matrix form of the temporal dynamics of $\mathbf{X}(t) = \begin{pmatrix} E(t) & I(t) \end{pmatrix}^\top$:

$$\dot{\mathbf{X}}(t) = \overline{\mathbf{R}}(t)\mathbf{X}(t)$$

with $\overline{\mathbf{R}}(t) = \begin{pmatrix} \overline{r}^{\overleftarrow{EE}} & \overline{r}^{\overleftarrow{EI}} \\ \overline{r}^{\overleftarrow{IE}} & \overline{r}^{\overleftarrow{II}} \end{pmatrix}$, the matrix of average transitions rates (the arrows indicate the sense of the transitions). We recall that the overlines refer to mean values of the phenotypic traits after averaging over the distribution of strain frequencies, such that:

$$\begin{cases} \overline{r}^{\overleftarrow{EE}} = r_m^{\overleftarrow{EE}} - (1 - p_m(t))\Delta r^{\overleftarrow{EE}} \\ \overline{r}^{\overleftarrow{EI}} = r_m^{\overleftarrow{EI}} - (1 - q_m(t))\Delta r^{\overleftarrow{EI}} \\ \overline{r}^{\overleftarrow{IE}} = r_m^{\overleftarrow{IE}} - (1 - p_m(t))\Delta r^{\overleftarrow{IE}} \\ \overline{r}^{\overleftarrow{II}} = r_m^{\overleftarrow{II}} - (1 - q_m(t))\Delta r^{\overleftarrow{II}} \end{cases}$$

In which, with $(i, j) \in \{E, I\}^2$, we denote the phenotypic differences: $\Delta r^{\overleftarrow{ji}} = r_m^{\overleftarrow{ji}} - r_w^{\overleftarrow{ji}}$, where the subscript m refer to the variant and the subscript w to the WT.

Note that these transition rates may or may not be time-dependent, depending on the model. For the sake of simplicity, we do not make here this (potential) time dependency explicit.

The temporal dynamics of $p_m(t)$ and $q_m(t)$ are given by:

$$\begin{cases} \dot{p}_m(t) = & p_m(t)(1 - p_m(t))\Delta r^{\overleftarrow{EE}} + q_m(t)(1 - q_m(t))\Delta r^{\overleftarrow{EI}} \left(\dfrac{q(t)}{p(t)}\right) + (q_m(t) - p_m(t))\overline{r}^{\overleftarrow{EI}} \left(\dfrac{q(t)}{p(t)}\right) \\ \dot{q}_m(t) = & q_m(t)(1 - q_m(t))\Delta r^{\overleftarrow{II}} + p_m(t)(1 - p_m(t))\Delta r^{\overleftarrow{IE}} \left(\dfrac{p(t)}{q(t)}\right) + (p_m(t) - q_m(t))\overline{r}^{\overleftarrow{IE}} \left(\dfrac{p(t)}{q(t)}\right) \end{cases} \tag{S17}$$

To focus on the differentiation between the exposed and the infectious compartments, we study here the variable $Q(t)$ such that:

$$Q(t) = \frac{p_m(t)}{(1 - p_m(t))} \frac{(1 - q_m(t))}{q_m(t)}. \tag{S18}$$

Thus:

$$\ln(Q(t)) = \mathrm{logit}(p_m(t)) - \mathrm{logit}(q_m(t)).$$

The temporal dynamics of $Q(t)$ is therefore given by:

$$\dot{Q}(t) = \frac{q_m(t)(1 - q_m(t))\dot{p}_m(t) - p_m(t)(1 - p_m(t))\dot{q}_m(t)}{\Big((1 - p_m(t))q_m(t)\Big)^2}.$$

By expanding the expressions for $\dot{p}_m(t)$ and $\dot{q}_m(t)$ and after numerous rearrangements, we obtain:

$$\begin{aligned} \frac{\mathrm{d}\ \ln(Q(t))}{\mathrm{d}t} = & \underbrace{\frac{q(t)}{p(t)}\left(\frac{1 - q_m(t)}{1 - p_m(t)}\right)\Delta r^{\overleftarrow{EI}} - \frac{p(t)}{q(t)}\left(\frac{1 - p_m(t)}{1 - q_m(t)}\right)\Delta r^{\overleftarrow{IE}} + \Delta r^{\overleftarrow{EE}} - \Delta r^{\overleftarrow{II}}}_{\text{Effect of selection } (\mathcal{O}(\varepsilon))} \\ & \underbrace{-\Big(Q(t) - 1\Big)\left(\frac{q(t)}{p(t)}\frac{q_m(t)}{p_m(t)} r_m^{\overleftarrow{EI}} + \frac{p(t)}{q(t)}\frac{1 - p_m(t)}{1 - q_m(t)} r_m^{\overleftarrow{IE}}\right)}_{\text{Effect of "migration" } (\mathcal{O}(1))}. \end{aligned} \tag{S19}$$

In the neutral case ($\varepsilon = 0$), the mutant strain m and the WT strain w have the same phenotype, that is: $\forall (i,j) \in \{E,I\}^2$, $\Delta r^{\overleftarrow{ji}} = 0$, and we rapidly have $p_m(t)/q_m(t) = (1-q_m(t))/(1-p_m(t)) = Q(t) = 1$ because "migration" – i.e. transitions between compartments E and I, including transmissions – spatially homogenises the frequencies of the variant. Selection will disrupt these quantities to $\mathcal{O}(\varepsilon)$.

Solving (S19) for $Q(t)$ based on a quasi-equilibrium approach, i.e. setting the right-hand sides of (S19) to 0, and using a Taylor expansion for the solution about the neutral case to order ε yields:

$$Q(t) \approx 1 + \frac{\Delta r^{\overleftarrow{EE}} + \left(\frac{q(t)}{p(t)}\right) \Delta r^{\overleftarrow{EI}} - \left(\frac{p(t)}{q(t)}\right) \Delta r^{\overleftarrow{IE}} - \Delta r^{\overleftarrow{II}}}{\left(\frac{q(t)}{p(t)}\right) r_m^{\overleftarrow{EI}} + \left(\frac{p(t)}{q(t)}\right) r_m^{\overleftarrow{IE}}} + \mathcal{O}(\varepsilon^2). \tag{S20}$$

By replacing the general form of the transition rates with the particular parameters of the model (S1), we get after some rearrangements:

$$Q(t) \approx 1 + \frac{\left(\frac{q(t)}{p(t)}\right)(1-c(t))\Delta\beta S(t)/N - \left(\frac{1}{q(t)}\right)\Delta\kappa + \Delta\alpha + \Delta\gamma}{\left(\frac{q(t)}{p(t)}\right)(1-c(t))\beta_m S(t)/N + \left(\frac{p(t)}{q(t)}\right)\kappa_m} + \mathcal{O}(\varepsilon^2). \tag{S21}$$

Note that $q(t)$ and $p(t)/q(t)$ may then be approximated by their quasi-equilibrium values.

Assuming that the virulence may be neglected ($\alpha_m = \alpha_w = 0$) and that there is no difference between the variant and the WT strains in terms of latent period, i.e. $\Delta\kappa = 0$, the previous equation then reduces to:

$$Q(t) \approx 1 + \frac{\left(\frac{q(t)}{p(t)}\right)(1-c(t))\Delta\beta S(t)/N + \Delta\gamma}{\left(\frac{q(t)}{p(t)}\right)(1-c(t))\beta_m S(t)/N + \left(\frac{p(t)}{q(t)}\right)\kappa} + \mathcal{O}(\varepsilon^2). \tag{S22}$$

It is interesting to note that the quasi-equilibrium of Q depends on $\Delta\beta$ and $\Delta\gamma$. More specifically, this expression predicts that the value of Q will be greater than 1 in the case of a variant with a higher transmission rate ($\Delta\beta$ >0 and $\Delta\gamma$ =0) while the value of Q will be less than 1 in the case of a variant with a longer duration of infectiousness, i.e. with a lower recovery rate, ($\Delta\gamma$ <0 and $\Delta\beta = 0$). Hence, some data on the differentiation between different host compartments (here between E and I) could potentially yield another way to estimate these two quantities.

S7 Relation with Blanquart *et al.* (2022), eLife

In (Blanquart et al., 2022), the growth rate of the epidemic $r(t)$ and the effective reproduction number $\mathcal{R}(t)$ – i.e. the average number of secondary infections – are linked through the framework popularized by Wallinga and Lipsitch in (Wallinga & Lipsitch, 2007). Let us consider an epidemic that grows exponentially at a rate $r(t)$ and a probability density function g for the generation time – i.e. timing of secondary infections. Assuming that the distribution of the age of infections stabilises very rapidly, the relationship between $r(t)$ and $\mathcal{R}(t)$ are given by (Wallinga & Lipsitch, 2007):

$$\frac{1}{\mathcal{R}(t)} = \int_0^{+\infty} e^{-r(t)a} g(a)\,\mathrm{d}a. \tag{S23}$$

Let us consider the scenario where a new variant m emerges and spreads in a host population previously dominated by a wild type strain w. The selection coefficient associated with the new variant can be computed from the difference in the *per capita* growth rate of the two variants: $s(t) = r_m(t) - r_w(t)$. The higher growth rate of the new variant can be due to different phenotypic effects acting on the transmission and/or the duration of infection and/or the shape of the whole distribution g. In our analysis we assume that the mean and the variance of the generation time distribution are linked due to the assumption of exponentially distributed sojourn times. In contrast, other studies have allowed the mean and the variance of the distribution to be independently modified by the mutations of the new variant (Blanquart et al., 2022; Park et al., 2022). More specifically we follow (Blanquart et al., 2022) and assume that the variant is characterized by its effective reproduction number $\mathcal{R}_m(t)$ and by its generation time distribution with mean μ_m and standard deviation σ_m (likewise, $\mathcal{R}_w(t)$, μ_w and σ_w, respectively, for the resident stain) such that:

$$\begin{cases} \mathcal{R}_m(t) & = \mathcal{R}_w(t)\big(1 + \delta_1\big) \\ \mu_m(t) & = \mu_w(t)\big(1 + \delta_2\big) \\ \sigma_m(t) & = \sigma_w(t)\big(1 + \delta_3\big) \end{cases} \tag{S24}$$

where δ_1, δ_2 and δ_3 refer to the effects of the mutation of the new variant on the three phenotypic traits as in (Blanquart et al., 2022). To characterize these phenotypic differences between the variant and the resident strain, one must then look at δ_1, δ_2 and δ_3.

In the well-known $S(E)IR$ models formalised by a system of ODEs, susceptible hosts S are infected with a constant transmission rate β and infectious individuals I recover at a constant rate γ. In (Blanquart et al., 2022), the authors assume that temporal variations in behavior and NPIs would affect the transmission, only captured by variability in the parameter $\mathcal{R}_w(t)$, without affecting the generation time distribution. We use the same assumption in our analysis through $c(t) \in [0;\ 1]$, the effectiveness of NPIs. Accounting for these control measures, the effective reproduction number in classical $S(E)IR$ models is given by: $\mathcal{R}(t) = \frac{(1-c(t))\beta}{\gamma} \frac{S(t)}{N}$, with $S(t)/N$ the proportion of susceptible hosts at time t in the population (of size N). We recall the notations made in the main text for the resident strain and the variant, respectively: β_w and $\beta_m = \beta_w + \Delta\beta$ referred to transmission rates; γ_w and $\gamma_m = \gamma_w + \Delta\gamma$ referred to recovery rates. Thus, assuming $c(t)$ to be the same for both strains, we have:

$$\begin{cases} \mathcal{R}_w(t) & = (1 - c(t))\dfrac{\beta_w}{\gamma_w}\dfrac{S(t)}{N} \\ \mathcal{R}_m(t) & = (1 - c(t))\left(\dfrac{\beta_w + \Delta\beta}{\gamma_w + \Delta\gamma}\right)\dfrac{S(t)}{N} \end{cases} \tag{S25}$$

As discussed in (Park et al., 2022), this is indeed valid for interventions that reduce transmission – e.g. social distancing, face covering – but no longer holds for interventions that lead to isolation of infected individuals – e.g. contact tracing.

In the following, our aim is to show the links between the framework developed in (Blanquart et al., 2022) – with phenotypic differences δ_1, δ_2 and δ_3 – and the framework we developed in this study –

with phenotypic differences $\Delta\beta$ and $\Delta\gamma$ – through the SIR and $SEIR$ models.

S7.1 SIR model

In the classical SIR model formalised by ODEs, the generation time is exponentially distributed (and thus memoryless). Let's start, however, with a gamma-distributed generation time as the exponential distribution is merely a special case of the gamma distribution family. Under this assumption, by substituting g in (S23) for the probability density function of the gamma distribution with mean μ_m and standard deviation σ_m, growth rates $r_w(t)$ and $r_m(t)$ thus become (Blanquart et al., 2022):

$$\begin{cases} r_w(t) &= \left(\mathcal{R}_w(t)^{\left(\frac{\sigma_w}{\mu_w}\right)^2} - 1\right)\dfrac{\mu_w}{\sigma_w^2} \\ r_m(t) &= \left(\Big(\mathcal{R}_w(t)\big(1+\delta_1\big)\Big)^{\left(\frac{\sigma_w\left(1+\delta_3\right)}{\mu_w\left(1+\delta_2\right)}\right)^2} - 1\right)\dfrac{\mu_w\big(1+\delta_2\big)}{\Big(\sigma_w\big(1+\delta_3\big)\Big)^2} \end{cases} \tag{S26}$$

Under the assumption of weak selection – i.e. δ_1, δ_2 and δ_3 are small and $\mathcal{O}(\varepsilon)$ – the selection gradient $s(t) = r_m(t) - r_w(t)$ is:

$$\begin{aligned} s(t) = \; & \left(\frac{\mathcal{R}_w(t)^{\left(\frac{\sigma_w}{\mu_w}\right)^2}}{\mu_w}\right)\delta_1 + \left(\left(\mathcal{R}_w(t)^{\left(\frac{\sigma_w}{\mu_w}\right)^2} - 1\right)\frac{\mu_w}{\sigma_w^2} - \frac{2\mathcal{R}_w(t)^{\left(\frac{\sigma_w}{\mu_w}\right)^2}\ln\big(\mathcal{R}_w(t)\big)}{\mu_w}\right)\delta_2 + \\ & 2\left(\frac{\mathcal{R}_w(t)^{\left(\frac{\sigma_w}{\mu_w}\right)^2}\ln\big(\mathcal{R}_w(t)\big)}{\mu_w} - \left(\mathcal{R}_w(t)^{\left(\frac{\sigma_w}{\mu_w}\right)^2} - 1\right)\frac{\mu_w}{\sigma_w^2}\right)\delta_3 + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{S27}$$

At equilibrium (i.e. $\mathcal{R}_w(t) = 1$), $s(t)$ reduces to: $s(t) = \delta_1/\mu_w + \mathcal{O}(\varepsilon^2)$. In **Fig. S13**, we plot an example of relation between $s(t)$ and $\mathcal{R}_w(t)$ according to (S27) for three types of variant. In accordance with (Blanquart et al., 2022), we can see that:

- Higher δ_1 are always selected (whatever the value of $\mathcal{R}_w(t)$);
- Lower δ_2 are selected when $\mathcal{R}_w(t) > 1$ (conversely, higher δ_2 are selected when $\mathcal{R}_w(t) < 1$), except in some cases when $\mathcal{R}_w(t)$ becomes too small;
- Higher δ_3 are always selected as soon as $\mathcal{R}_w(t) \neq 1$.

When $\mathcal{R}_w(t)$ is not too far from 1, $\ln\big(\mathcal{R}_w(t)\big) \approx \mathcal{R}_w(t) - 1$, and, eventually assuming that the generation time of the resident strain is exponentially distributed (special case of gamma distribution where $\sigma_w = \mu_w$), (S27) becomes:

$$s(t) \approx \left(\frac{\mathcal{R}_w(t)}{\mu_w}\right)\delta_1 + \left(\frac{\big(\mathcal{R}_w(t)-1\big)\big(1-2\mathcal{R}_w(t)\big)}{\mu_w}\right)\delta_2 + 2\left(\frac{\big(\mathcal{R}_w(t)-1\big)^2}{\mu_w}\right)\delta_3 + \mathcal{O}(\varepsilon^2). \tag{S28}$$

This expression, easier to understand than the previous one, leads to the same interpretations.

We now use the expressions in (S25) for the effective reproduction number of the resident strain and of the variant. Under the assumption of weak selection – i.e. $\Delta\beta$ and $\Delta\gamma$ are small and $\mathcal{O}(\varepsilon)$ – a Taylor expansion for the effective reproduction number of the variant $\mathcal{R}_m(t)$ about the neutral case ($\varepsilon = 0$) to order ε yields:

$$\mathcal{R}_m(t) = \mathcal{R}_w(t)\left(1 + \underbrace{\frac{\Delta\beta}{\beta_w} - \mu_w\Delta\gamma}_{\delta_1}\right) + \mathcal{O}(\varepsilon^2), \tag{S29}$$

with $\mu_w = 1/\gamma_w$, the mean generation time for the resident strain.

Likewise, for $\mu_m = 1/(\gamma_w + \Delta\gamma)$, the mean generation time of the variant:

$$\mu_m(t) = \mu_w(t)\left(1 \underbrace{-\mu_w\Delta\gamma}_{\delta_2}\right) + \mathcal{O}(\varepsilon^2), \tag{S30}$$

This result is the same for σ_m, the standard deviation of the generation time of the variant, as $\sigma_m = \mu_m$ for the exponential distribution. Hence, using the notations of (Blanquart et al., 2022):

$$\begin{cases} \delta_1 & = \dfrac{\Delta\beta}{\beta_w} - \mu_w\Delta\gamma \\ \delta_2 & = -\mu_w\Delta\gamma \\ \delta_3 & = -\mu_w\Delta\gamma \end{cases} \tag{S31}$$

Substituting these expressions for δ_1, δ_2 and δ_3 in (S27) for the exponential case ($\mu_w = \sigma_w = 1/\gamma_w$) along with the expression of $\mathcal{R}_w(t)$ in (S25), $s(t)$ reduces to:

$$s(t) \approx (1 - c(t))\Delta\beta\frac{S(t)}{N} - \Delta\gamma + \mathcal{O}(\varepsilon^2), \tag{S32}$$

which is indeed known to be the expression of the selection gradient in the simplest SIR model (Day & Gandon, 2006, 2007).

S7.2 *SEIR* model

We now add an exposed state – i.e. infected but not yet infectious –, that individuals leave at a constant rate κ, altering the generation time distribution. Let us assume that the infectious period is gamma-distributed with mean μ^I and standard deviation σ^I and that the exposed period is exponentially distributed with mean $1/\kappa$. Therefore, the convolution:

$$g(a) = \int_0^a \kappa e^{-\kappa x} \frac{(a-x)^{\left(\frac{\mu^I}{\sigma^I}\right)^2 - 1} e^{-\frac{\mu^I(a-x)}{(\sigma^I)^2}}}{\Gamma\left[\left(\frac{\mu^I}{\sigma^I}\right)^2\right]\left(\frac{(\sigma^I)^2}{\mu^I}\right)^{\left(\frac{\mu^I}{\sigma^I}\right)^2}} \mathrm{d}x$$

– where Γ is the Gamma function –, yields a probability density function for the generation time with mean $\mu = 1/\kappa + \mu^I$ and standard deviation $\sigma = \sqrt{1/\kappa^2 + \left(\sigma^I\right)^2}$. Substituting the probability density function in (S23) for this convolution gives:

$$\mathcal{R}(t) = \left(1 + \frac{r(t)}{\kappa}\right)\left(1 + \frac{\left(\sigma^I\right)^2}{\mu^I} r(t)\right)^{\left(\frac{\mu^I}{\sigma^I}\right)^2}. \tag{S33}$$

The issue with this expression for the $SEIR$ model is that, although it is easy to express $\mathcal{R}(t)$ as a function of $r(t)$, the reverse (expressing $r(t)$ as a function of $\mathcal{R}(t)$) does not seem to be true. Nevertheless, we may look at some special cases.

First, when $1/\kappa \to 0^+$ – i.e. the $SEIR$ model tends to the SIR model since the exposed individuals tend, on average, to leave their compartment instantaneously –, we find indeed the result for the SIR model (S26).

Besides, when the infectious period is now exponentially distributed (with $\sigma^I = \mu^I = 1/\gamma$), the generation time is hypoexponentially distributed (generalized Erlang distribution) and the previous expression becomes:

$$\mathcal{R}(t) = \left(1 + \frac{r(t)}{\kappa}\right)\left(1 + \mu^I r(t)\right), \tag{S34}$$

as already shown in (Wallinga & Lipsitch, 2007), which yields:

$$r(t) = \frac{-\kappa\mu^I - 1 + \sqrt{\left(\kappa\mu^I - 1\right)^2 + 4\kappa\mu^I\mathcal{R}(t)}}{2\mu^I} \tag{S35}$$

It corresponds to the expression we used for the growth rate of the epidemic in this study.

Assuming no change in κ between the resident strain w and the variant m – i.e. same latent period, on average, for both strains –, we still have $\mathcal{R}_m(t) = \mathcal{R}_w(t)\left(1 + \delta_1\right)$ and we obtain from the expression of the mean generation time μ_m in (S24) an expression for the mean duration of infectiousness μ_m^I:

$$\begin{aligned} \mu_m(t) = \mu_w(t)\left(1 + \delta_2\right) \quad &\iff \quad \frac{1}{\kappa} + \mu_m^I = \left(\frac{1}{\kappa} + \mu_w^I\right)\left(1 + \delta_2\right) \\ &\iff \quad \mu_m^I = \mu_w^I + \delta_2\left(\frac{1}{\kappa} + \mu_w^I\right) \end{aligned} \tag{S36}$$

Substituting $\mathcal{R}(t)$ and μ^I in (S35) using (S24) and (S36), respectively, we can calculate the selection gradient $s(t) = r_m(t) - r_w(t)$. Again, a Taylor expansion about the neutral case (weak selection) gives:

$$\begin{aligned} s(t) = & \left(\frac{\kappa \mathcal{R}_w(t)}{\sqrt{\left(\kappa\mu_w^I - 1\right)^2 + 4\kappa\mu_w^I \mathcal{R}_w(t)}}\right)\delta_1 + \\ & \left(\frac{\left(\kappa\mu_w^I + 1\right)\left(\sqrt{\left(\kappa\mu_w^I - 1\right)^2 + 4\kappa\mu_w^I \mathcal{R}_w(t)} + \kappa\mu_w^I\left(1 - 2\mathcal{R}_w(t)\right) - 1\right)}{2\kappa\left(\mu_w^I\right)^2\sqrt{\left(\kappa\mu_w^I - 1\right)^2 + 4\kappa\mu_w^I \mathcal{R}_w(t)}}\right)\delta_2 + \mathcal{O}(\varepsilon^2). \end{aligned} \quad \text{(S37)}$$

At equilibrium (i.e. $\mathcal{R}_w(t) = 1$), $s(t)$ is simply: $s(t) = \delta_1 / \left(1/\kappa + \mu_w^I\right) + \mathcal{O}(\varepsilon^2)$. Furthermore, in any case, we also have:

- Higher δ_1 are always selected (whatever the value of $\mathcal{R}_w(t)$);
- Lower δ_2 are selected when $\mathcal{R}_w(t) > 1$ (conversely, higher δ_2 are selected when $\mathcal{R}_w(t) < 1$).

As in the previous subsection with the SIR model, a weak selection approximation of $\mathcal{R}_m(t)$ from (S25) yields:

$$\mathcal{R}_m(t) = \mathcal{R}_w(t)\Bigg(1 + \underbrace{\frac{\Delta\beta}{\beta_w} - \mu_w^I\Delta\gamma}_{\delta_1}\Bigg) + \mathcal{O}(\varepsilon^2), \quad \text{(S38)}$$

and, for the mean generation time of the variant $\mu_m = 1/\kappa + 1/(\gamma_w + \Delta\gamma)$:

$$\mu_m(t) = \mu_w(t)\Bigg(1 \underbrace{- \frac{\kappa\left(\mu_w^I\right)^2}{\kappa\mu_w^I + 1}\Delta\gamma}_{\delta_2}\Bigg) + \mathcal{O}(\varepsilon^2). \quad \text{(S39)}$$

Hence, with the notations of (Blanquart et al., 2022):

$$\begin{cases} \delta_1 & = \dfrac{\Delta\beta}{\beta_w} - \mu_w^I\Delta\gamma \\ \delta_2 & = -\dfrac{\kappa\left(\mu_w^I\right)^2}{\kappa\mu_w^I + 1}\Delta\gamma \end{cases} \quad \text{(S40)}$$

Substituting these expressions for δ_1 and δ_2 in (S37) along with the expression of $\mathcal{R}_w(t)$ in (S25) and $\mu_w^I = 1/\gamma_w$, the selection gradient becomes after some rearrangements:

$$\begin{aligned} s(t) \quad = \quad & \left(\frac{\kappa}{\sqrt{\left(\kappa - \gamma_w\right)^2 + 4\kappa(1 - c(t))\beta_w\frac{S(t)}{N}}}\right)(1 - c(t))\Delta\beta\frac{S(t)}{N} \\ & - \frac{1}{2}\left(\frac{\kappa - \gamma_w}{\sqrt{\left(\kappa - \gamma_w\right)^2 + 4\kappa(1 - c(t))\beta_w\frac{S(t)}{N}}} + 1\right)\Delta\gamma + \mathcal{O}(\varepsilon^2). \end{aligned} \quad \text{(S41)}$$

This is the theoretical derivation of the selection gradient we used in this study (cf. equation (S14)).

References

Blanquart, F., Hozé, N., Cowling, B. J., Débarre, F., & Cauchemez, S. (2022). Selection for infectivity profiles in slow and fast epidemics, and the rise of sars-cov-2 variants. *eLife*. https://doi.org/10.7554/eLife.75791

Day, T., & Gandon, S. (2006). Insights from price's equation into evolutionary epidemiology. *Disease Evolution: Models, Concepts, and Data Analysis*, 23–44.

Day, T., & Gandon, S. (2007). Applying population-genetic models in theoretical evolutionary epidemiology. *Ecology Letters*, *10*, 876–888. https://doi.org/10.1111/j.1461-0248.2007.01091.x

Lion, S. (2018). Class structure, demography and selection: Reproductive-value weighting in nonequilibrium, polymorphic populations. *Am. Nat.*, *191*(5). https://doi.org/10.1086/696976

Lion, S., & Gandon, S. (2022). Evolution of class-structured populations in periodic environments. *Evolution*, *76*(8), 1674–1688. https://doi.org/10.1101/2021.03.12.435065

Park, S. W., Bolker, B. M., Funk, S., Metcalf, C. J. E., Weitz, J. S., Grenfell, B. T., & Dushoff, J. (2022). The importance of the generation interval in investigating dynamics and control of new sars-cov-2 variants. *J.R. Soc. Interface*, *19*(191). https://doi.org/10.1098/rsif.2022.0173

Wallinga, J., & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B*, *274*, 599–604. https://doi.org/10.1098/rspb.2006.3754

Chapter three

Virus Evolution, 2024, **10(1)**, veae069
DOI: https://doi.org/10.1093/ve/veae069
Advance access publication date: 15 October 2024
Research Article

Evolution of virulence in emerging epidemics: from theory to experimental evolution and back

Wakinyan Benhamou[1,†,*], François Blanquart[2,†], Marc Choisy[3,4], Thomas W. Bengruber[5], Rémi Choquet[1,‡], Sylvain Gandon[1,‡]

[1]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France
[2]Centre for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, PSL Research University, Paris, France
[3]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, United Kingdom
[4]Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
[5]Reinhard-zu-Rhynernstrasse 7, Hamm 59069, Germany

[†]These authors contributed equally to this work.
[‡]These authors also contributed equally to this work.
*Corresponding author. CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France. E-mail: wakinyan.benhamou@cefe.cnrs.fr

Abstract

The experimental validation of theoretical predictions is a crucial step in demonstrating the predictive power of a model. While quantitative validations are common in infectious diseases epidemiology, experimental microbiology primarily focuses on the evaluation of a qualitative match between model predictions and experiments. In this study, we develop a method to deepen the quantitative validation process with a polymorphic viral population. We analyse the data from an experiment carried out to monitor the evolution of the temperate bacteriophage λ spreading in continuous cultures of *Escherichia coli*. This experimental work confirmed the influence of the epidemiological dynamics on the evolution of transmission and virulence of the virus. A variant with larger propensity to lyse bacterial cells was favoured in emerging epidemics (when the density of susceptible cells was large), but counter-selected when most cells were infected. Although this approach qualitatively validated an important theoretical prediction, no attempt was made to fit the model to the data nor to further develop the model to improve the goodness of fit. Here, we show how theoretical analysis—including calculations of the selection gradients—and model fitting can be used to estimate key parameters of the phage life cycle and yield new insights on the evolutionary epidemiology of the phage λ. First, we show that modelling explicitly the infected bacterial cells which will eventually be lysed improves the fit of the transient dynamics of the model to the data. Second, we carry out a theoretical analysis that yields useful approximations that capture at the onset and at the end of an epidemic the effects of epidemiological dynamics on selection and differentiation across distinct life stages of the virus. Finally, we estimate key phenotypic traits characterizing the two strains of the virus used in our experiment such as the rates of prophage reactivation or the probabilities of lysogenization. This study illustrates the synergy between experimental, theoretical, and statistical approaches; and especially how interpreting the temporal variation in the selection gradient and the differentiation across distinct life stages of a novel variant is a powerful tool to elucidate the evolutionary epidemiology of emerging infectious diseases.

Keywords: evolutionary epidemiology; life-history evolution; competition; natural selection; statistical inference; bacteriophage λ

1. Introduction

Evolutionary epidemiology theory predicts that the evolution of pathogen transmission is driven by the availability of susceptible hosts. At the onset of an epidemic, when the density of susceptible hosts is high, more transmission is favoured by natural selection. When a positive covariance exists between transmission and virulence (Anderson and May, 1982; Alizon et al., 2009; Alizon and Michalakis, 2015), this selection for higher transmission can indirectly select for higher virulence (Bull, 1994; Day, 2002; Lenski and May, 1994; Frank, 1996; Day and Proulx, 2004; Gandon and Day, 2007). Yet, an experimental validation of this prediction was needed to demonstrate the relevance of these predictions on the evolution of pathogens in emerging epidemics.

This prediction was put to the test in a previous study using experimental evolution of the temperate bacteriophage (or phage) λ (Berngruber et al., 2013). Phages are viruses that infect bacteria and phage λ is the archetypal temperate phage, which can switch between a lytic and a lysogenic life style. Upon infection, the virus may commit to the lytic pathway by hijacking the host's replication machinery to produce new virions (viral particles) and eventually release them in the environment after the lysis of the host cell. Alternatively, the virus may commit to the lysogenic pathway by integrating its genome into the bacterial chromosome where it will lie in a dormant state as a prophage and be hereditarily transmitted to daughter cells at the pace of lysogen divisions. The prophage may also regain virulence by excising itself from the

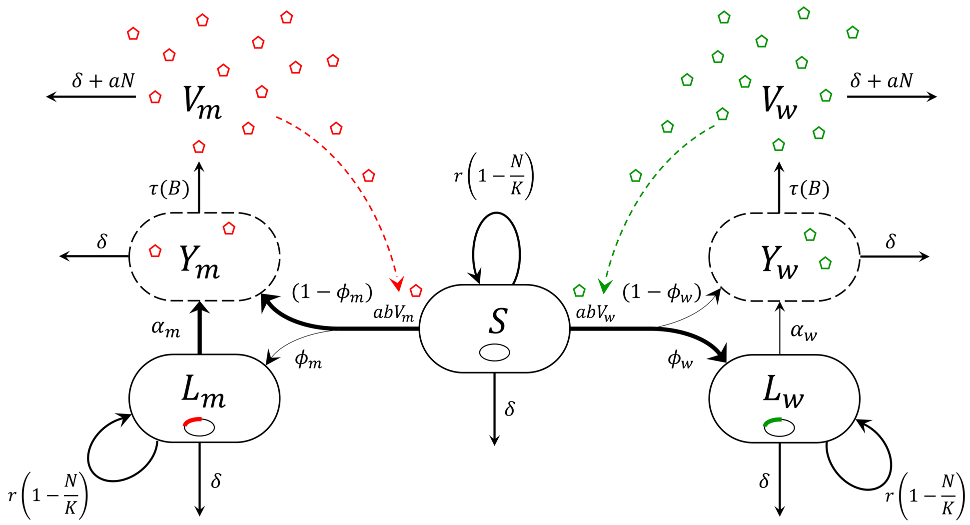


Figure 1. Flow chart of the phage-bacteria system. The subscripts w and m refer to the wildtype (in green) and mutant (or virulent, in red) strain of phage λ, respectively. In the bacterial population (*E. coli*) of size N, bacteria are either susceptible (S), lysogenic (L)—i.e. carrying a prophage in their chromosome (small ellipses)—or carrying phages (small pentagons) replicating in their cytoplasm prior to lysis (Y); V is the free virus stage (culture medium). Width of the arrows between S, L, and Y reflects the relative rates of different events: the wildtype strain is transmitted mostly vertically (high probability of lysogenization and low rate of prophage reactivation) while the virulent strain is transmitted mostly horizontally (low probability of lysogenization and high rate of prophage reactivation). Dashed arrows symbolize the role of the free viral particles in the force of infection (epidemiological feedback). See notations in Table 1.

host genome, switching to a lytic cycle (reactivation, also called induction) and thus shifting from vertical to horizontal transmission (Lwoff, 1953; Echols, 1972; Ptashne, 1992; Gandon, 2016). The evolution experiment designed in Berngruber et al. (2013) monitored the competition between two strains of phage λ with distinct life-history strategies in continuous cultures of *Escherichia coli*. The first strain is the wildtype, which is known to have a relatively large lysogenization rate and low reactivation rate. The second strain is the λcI857 variant, which carries a point mutation in the transcriptional repressor protein cI (St-Pierre and Endy, 2008; Sussman and Jacob, 1962); the cI mutant is known to be more virulent and transmitted mostly horizontally through lytic cycles. Berngruber et al. (2013) developed a mathematical model tailored to the life cycle of phage λ. Numerical simulations of this model using parameter values from previous experimental studies led to three theoretical predictions: (i) the virulent strain outcompetes the wildtype when susceptible hosts are abundant, but the direction of selection is reversed as soon as the epidemic reaches high prevalence, (ii) the lower the initial prevalence, the higher the increase in virulence during the epidemic and (iii) the virulent strain is always more frequent among viral particles than among prophages. Tracking both the epidemiology (prevalence) and the evolution of the virus (frequency of the virulent strain among viral particles and among infected bacteria), all three predictions were confirmed experimentally (Berngruber et al., 2013). Yet, the data were only used as a qualitative validation of the theory and no attempt was made to explore the quantitative match between the predicted and the observed dynamics of the virus.

In the present work, we show how the quantitative analysis of the experimental results from Berngruber et al. (2013) improves our understanding of the evolutionary epidemiology of phage λ. First, we modified the structure of the epidemiological model to better capture the transient evolutionary dynamics of the virus among infected bacteria. Second, we carry out a theoretical analysis of this model to provide useful approximations to predict the evolutionary dynamics of the virus at different stages of the epidemic. In particular, we compute the selection gradients and the differentiation across distinct life stages of the virus. Finally, we develop a statistical inference approach to obtain quantitative estimates of the parameters of the model and, especially, the life-history traits of the different strains of phage λ.

2. Materials and methods

Following Berngruber et al. (2013), we first model the competition between the wildtype strain—hereafter denoted by the subscript w—and the mutant strain λcI857—hereafter denoted by the subscript m—of phage λ in a chemostat with a well-mixed continuous culture of its bacterial host *E. coli* (Fig. 1). We summarize the notations in Table 1. We then recall the experimental data generated in the original study and describe how we generate simulated data to validate our ability to estimate model parameters. In the last section, we detail the statistical inference approach we developed.

2.1 A model coupling epidemiology and evolution of phage virulence

2.1.1 *Epidemiology*

Bacteria are either susceptible (S) or infected with phage λ (we do not consider resistant bacteria); infected bacteria may be lysogenic (L), following phage integration (lysogenization), or carrying phages replicating in their cytoplasm prior to lysis (Y), following either lytic infection or prophage reactivation. Unlike the original model in Berngruber et al. (2013), adding an extra stage Y allows us to take the lysis time into account; we show below that this new model improves the goodness of fit to the data (see Results section '3.2.2 Inference from experimental data'). The free virus stage (V) corresponds to the free viral particles (virions) that are in the culture medium—'free' meaning 'extracellular' here. Throughout, for each state variable (aka compartment), for instance S, we denote by $S(t)$ its density at the current time t and $\dot{S}(t)$ its derivative with respect to time. Hence, $L(t) = L_w(t) + L_m(t)$ is the total density at time t of L cells, $Y(t) = Y_w(t) + Y_m(t)$, of Y cells, $V(t) = V_w(t) + V_m(t)$, of free viral particles, and $N(t) = S(t) + L(t) + Y(t)$, of bacteria.

Table 1. Notations. The subscripts w and m refer to the wild-type strain and the mutant (or virulent) strain λcI857 of phage λ, respectively. Overlines refer to mean values of life-history traits across all genotypes. SD stands for 'standard deviation'.

Term	Definition
N	Bacterial population
S	Susceptible bacteria
L, L_w, L_m	Lysogens (bacteria carrying a prophage)
Y, Y_w, Y_m	Bacteria with phages replicating in their cytoplasm prior to lysis
V, V_w, V_m	Free viral particles (virions)
P	Prevalence $(1-S/N)$
p	Frequency of L cells infected by the mutant strain m
q	Frequency of the mutant strain m in the free virus stage (V)
f	Frequency of Y cells infected by the mutant strain m
g	Frequency of cells infected (either Y or L) by the mutant strain m
$\overline{\alpha}, \alpha_w, \alpha_m$	Rates of prophage reactivation; $\alpha_w < \alpha_m$
$\overline{\phi}, \phi_w, \phi_m$	Probabilities of lysogenization (phage integration into the host genome upon infection); $\phi_w > \phi_m$
$\Delta\alpha, \Delta\phi$	Phenotypic differences between the mutant and wildtype strain; $\Delta\alpha = \alpha_m - \alpha_w, \Delta\phi = \phi_m - \phi_w$
a	Adsorption rate of free viral particles onto the surface of bacteria
b	Probability of fusion (injection of the phage genome) upon adsorption
τ	Lysis rate; $1/\tau$ corresponds to the mean lysis time
B	Burst size (number of released viral particles upon lysis)
r	Bacterial intrinsic growth rate
K	Carrying capacity for the bacterial population: $K = 10^9$ cells
δ	Dilution rate of the continuous culture: $\delta = 0.8\ \mathrm{h}^{-1}$
$\mathcal{R}_{0,w}, \mathcal{R}_{0,m}$	Basic reproduction numbers of the phage
$\mathcal{S}$	Selection gradient of the virulent phage (rate at which it grows or declines in frequency on the logit scale)
Q^{VL}	Differentiation of the virulent phage between free phages and prophages
$\sigma_P, \sigma_g, \sigma_q$	SD of measurement errors for $\text{logit}(P)$, $\text{logit}(g)$ and $\text{logit}(q)$, respectively

Susceptible and lysogenic bacteria grow at a *per capita* logistic rate $r(1-N(t)/K)$, where r is the intrinsic growth rate and K the carrying capacity. We assume, as in Berngruber et al. (2013), that the prophage does not affect the intrinsic growth rate of its host and that vertical transmission is perfect. Bacteria and virions are removed from the chemostat at a dilution rate δ. Free viral particles adsorb onto bacterial cells at a rate a; adsorption is non-reversible, that is, the fate of adsorbed viruses is only to infect or die. Infection of a susceptible host also requires the injection of the phage's genetic material into the bacterial cytoplasm (fusion, with probability b). The force of infection—i.e. the *per capita* infection rate—is therefore given by $abV(t)$. We assume that superinfection (including coinfection with both strains) does not occur. In particular, prophage establishment of phage λ is known to provide cellular immunity, or superinfection inhibition (Lwoff, 1953; Ptashne, 1992; Berngruber et al., 2010; Gandon, 2016). Upon infection, the wildtype and virulent strains of phage λ may either be integrated as prophages into the host genome (lysogenic cycle) with probabilities ϕ_w and ϕ_m, respectively, such that $\Delta\phi = \phi_m - \phi_w < 0$, or start the biosynthesis and assembly of viral copies (lytic cycle) with complementary probabilities. Once integrated, reactivations of the wildtype and virulent prophages occur at a rate α_w and α_m, respectively, such that $\Delta\alpha = \alpha_m - \alpha_w > 0$. Following either lytic infections or prophage reactivations, host cells are lysed at a rate τ and eventually release B viral particles upon lysis (burst size). We assume that parameters a, b, τ, and B are the same for the wild-type and the mutant which yields the following system of ordinary differential equations (ODEs):

$$\begin{cases} \dot{S}(t) = \underbrace{rS(t)\left(1-\frac{N(t)}{K}\right)}_{\text{Growth}} - \underbrace{abV(t)S(t)}_{\text{Infection}} - \underbrace{\delta S(t)}_{\text{Removal}} \\ \dot{L}(t) = \underbrace{rL(t)\left(1-\frac{N(t)}{K}\right)}_{\substack{\text{Growth \&}\\ \text{vertical transmission}}} + \underbrace{\overline{\phi}(t)abV(t)S(t)}_{\text{Lysogenization}} - \underbrace{\overline{\alpha}(t)L(t)}_{\substack{\text{Prophage}\\ \text{reactivation}}} - \underbrace{\delta L(t)}_{\text{Removal}} \\ \dot{Y}(t) = \underbrace{(1-\overline{\phi}(t))abV(t)S(t)}_{\text{Lytic infection}} + \underbrace{\overline{\alpha}(t)L(t)}_{\substack{\text{Prophage}\\ \text{reactivation}}} - \underbrace{\tau Y(t)}_{\text{Lysis}} - \underbrace{\delta Y(t)}_{\text{Removal}} \\ \dot{V}(t) = \underbrace{\tau Y(t)B}_{\text{Virion release}} - \underbrace{aN(t)V(t)}_{\text{Adsorption}} - \underbrace{\delta V(t)}_{\text{Removal}} \end{cases}, \quad (1)$$

where $\overline{\phi}(t)$ is the mean probability of lysogenization upon infection and $\overline{\alpha}(t)$ the mean rate of reactivation among prophages:

$$\begin{cases} \overline{\phi}(t) = q(t)\phi_m + (1-q(t))\phi_w \\ \overline{\alpha}(t) = p(t)\alpha_m + (1-p(t))\alpha_w \end{cases}, \quad (2)$$

with $q(t) = V_m(t)/V(t)$ and $p(t) = L_m(t)/L(t)$, the frequencies of the virulent strain at time t among compartments V and L, respectively.

2.1.2 *Evolution*

Along with these epidemiological dynamics, we also track the evolutionary dynamics of phage λ. We recall that $p(t)$ and $q(t)$ refer to frequencies of the virulent strain in compartment L and V, respectively. In addition, we also denote $f(t) = Y_m(t)/Y(t)$, the frequency of the virulent strain at time t among Y cells, and $g(t)$, the frequency of the virulent strain in infected cells (either Y or L) such that:

$$g(t) = \frac{Y_m(t)+L_m(t)}{Y(t)+L(t)} = f(t)\left(\frac{Y(t)}{Y(t)+L(t)}\right) + p(t)\left(\frac{L(t)}{Y(t)+L(t)}\right). \quad (3)$$

Using (1) and (2), one may calculate the ODE for each of these frequencies (see Supplementary Appendix S1.2). More conveniently, we will then focus on logit-frequencies instead, that is, the log odds ln(frequency of the mutant/frequency of the wild-type). Taken together, the equations of the temporal dynamics of the frequencies and the model (1) yield the coupled evolutionary-epidemiological dynamics of this phage-bacteria system. The analysis of this model can provide key insight on the evolutionary forces acting on the virus; it may also provide useful approximations for the change in mutant frequency at different stages of the epidemic.

2.2 Time series datasets

2.2.1 *Experimental data*

We use experimental time series obtained in the first evolution experiment of Berngruber et al. (2013). Briefly, this experiment started with eight independent chemostats (5 mL chamber volume, maintained at 35°C) of well-mixed bacterial cultures of *E. coli* MG1655 (RecA+) at carrying capacity. Bacteria were initially infected by both the wildtype and virulent strain λcI857 of phage λ (prophage stage with initial ratio 1:1). Two treatments were considered using four chemostats each: (i) an epidemic treatment—low initial prevalence, around 1%—and (ii) an endemic treatment—high initial prevalence, around 99%. Throughout the course of the experiment, several quantities were monitored: the prevalence, the frequency infected by each strain and the strain frequency in the culture medium (free virus stage). Samplings in each chemostat were performed hourly, from 1 to 60 h maximum. The prevalence and the frequency of infected hosts were tracked using flow cytometry (FACS) with fluorescent protein marker colours (CFP and YFP) while the strain frequency in the free virus stage was tracked by qPCR.

2.2.2 Simulated data

Alongside experimental data, we also carry out an analysis based on simulated data in order to validate our ability to infer parameters from experimental data. For this purpose, we take: $\alpha_w = 7\times10^{-3}$, $\alpha_m = 2\times10^{-2}$, $\phi_w = 0.2$, $\phi_m = 2\times10^{-2}$, $a = 3\times10^{-9}$, $b = 0.1$, $B = 80$, $r = 1.4$, $\tau = 1.5$, $K = 10^9$ and $\delta = 0.8$. At $t = 0$, bacteria are at carrying capacity K with initial prevalence 1% (epidemic treatment) or 99% (endemic treatment) and the initial prophage ratio for the two strains is 1:1. We simulate the deterministic model (1) for the epidemic and endemic treatment (Supplementary Fig. S1). We then add i.i.d. Gaussian noise at each time point to mimic measurement errors on the logit-prevalence $\text{logit}(P(t))$, the logit-frequency of hosts infected by the virulent phage $\text{logit}(g(t))$ and the logit-frequency of the virulent phage in the free virus stage $\text{logit}(q(t))$—this is independently repeated four times for each simulation to obtain four replicates (chemostats) per treatment. We modulate data quantity through two sampling frequencies: 0.1 h^{-1} vs. 1 h^{-1}, and we modulate data quality through two standard deviations (SD) of measurement errors: 0.01 vs. 0.5. We thus end up with four combinations of data quantity and quality (see example in Supplementary Fig. S2) from which we then try to recover parameter values.

2.3 Maximum likelihood estimation

For the estimation process, we used a two-step approach. Using theoretical results, we first compute point estimates of the rates of prophage reactivation α_w and α_m. Then, we fix the latter to estimate the remaining parameters of the model using non-linear optimizations. Note that it is possible to run non-linear optimizations to estimate all parameters but this two-step approach makes the optimization easier by reducing the dimensionality of the problem. The Sieve bootstrap method (Bühlmann, 1997; Ulloa et al., 2013) is used to compute the joint distributions of all these estimated parameters.

2.3.1 Estimation of the rates of prophage reactivation

From the analysis of the model (see Results section '3.1 Theoretical analysis'), we show that when the system reaches high prevalence the selection gradient $\mathcal{S}$ of the mutant—i.e. the rate at which it grows or declines in frequency on the logit scale—is simply given by: $\mathcal{S} = \alpha_w - \alpha_m = -\Delta\alpha$ (see Results section '3.1.2 Evolution at the end of the epidemic'). Furthermore, the differentiation $\mathcal{Q}^{VL}$ of the virulent strain between free phages and prophages converges towards approximately: $\mathcal{Q}^{VL} = \alpha_m/\alpha_w = 1 + \Delta\alpha/\alpha_w$ (see Results section '3.1.3 Differentiation across compartments'). Combining these two expressions enables us to estimate separately both reactivation rates:

$$\begin{cases} \alpha_w &= \dfrac{\mathcal{S}}{1-\mathcal{Q}^{VL}} \\ \alpha_m &= \dfrac{\mathcal{S}\times\mathcal{Q}^{VL}}{1-\mathcal{Q}^{VL}} \end{cases}.$$

For each chemostat, we therefore only keep the data from the time point the prevalence has reached 95%. We fit a linear model on the logit-frequency infected by the virulent phage $\text{logit}(g(t))$ to estimate the slope $\mathcal{S}$, and we estimate $\mathcal{Q}^{VL}$ by calculating the geometric mean of $\frac{q(t)(1-g(t))}{(1-q(t))g(t)}$ (see details in Supplementary Appendix S3.1). Note that we substitute $p(t)$ by $g(t)$ for the calculation of the differentiation because we only have access to $g(t)$ in the experiment, and $g(t)$ is almost identical to $p(t)$ towards the end of the epidemic (see Results section '3.1.2 Evolution at the end of the epidemic' and Supplementary Fig. S1).

2.3.2 Estimation of the remaining parameters

We have three response variables: (i) $\text{logit}(P(t))$, the logit-prevalence, (ii) $\text{logit}(g(t))$, the logit-frequency of hosts infected by the virulent phage, and (iii) $\text{logit}(q(t))$, the logit-frequency of the virulent phage in the free virus stage. Let $P_0^{epidemic}$ and $P_0^{endemic}$ be the initial conditions (at $t = 0$) of the prevalence in the epidemic and endemic treatments, respectively, and p_0, the initial condition of the frequency infected by the virulent phage (identical for both treatments). Let $\theta = \left(\phi_w,\ \phi_m,\ a,\ b,\ r,\ \tau,\ P_0^{epidemic},\ P_0^{endemic},\ p_0\right)$ be the vector of model parameters to estimate. Several parameter values are fixed: $K = 10^9$ cells and $\delta = 0.8$ h^{-1} (Berngruber et al., 2013); α_w and α_m are fixed to their previous point estimates (cf. Material and methods section '2.3.1 Estimation of the rates of prophage reactivation'); and $B = 80$ virus.cell^{-1} (Wang, 2006) as we show in Results section '3.2.1 Inference from simulated data' that the burst size B is not separately identifiable from parameter b. We estimate as well σ_P, σ_g, and σ_q, the SD of measurement errors for each of the three response variables, respectively. We assume that variation around the deterministic dynamics stems solely from measurement errors and we thereby neglect any additional process stochasticity. This is justified in particular by the very controlled conditions of the experiment and by the large population sizes of both the bacteria and the phage in the chemostats. For each response variable, measurement errors are assumed to be normally distributed and i.i.d. across all treatments, replicates, and time points. The corresponding likelihoods ($\mathcal{L}$) are thence respectively given by:

$$\mathcal{L}_P(\theta,\sigma_P) = \prod_{i,j,t} \varphi\left(\text{logit}(P_{i,j}(t))^{data} \,\middle|\, \text{logit}(P_{i,j}(\theta,t))^{sim},\ \sigma_P^2\right),$$

$$\mathcal{L}_g(\theta,\sigma_g) = \prod_{i,j,t} \varphi\left(\text{logit}(g_{i,j}(t))^{data} \,\middle|\, \text{logit}(g_{i,j}(\theta,t))^{sim},\ \sigma_g^2\right),$$

$$\mathcal{L}_q(\theta,\sigma_q) = \prod_{i,j,t} \varphi\left(\text{logit}(q_{i,j}(t))^{data} \,\middle|\, \text{logit}(q_{i,j}(\theta,t))^{sim},\ \sigma_q^2\right),$$

where i refers to treatments (epidemic vs. endemic), j to the jth replicate (chemostat), t to time points and $\varphi(.\,|\,\mu,\sigma^2)$ to the probability density function of the Normal distribution with mean μ and variance σ^2; *sim* indicates model outputs. We then denote $\hat{\theta}$, the maximum likelihood estimation (MLE) estimator of θ, such that:

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\left(\ln(\mathcal{L}_P(\theta,\sigma_P)) + \ln(\mathcal{L}_g(\theta,\sigma_g)) + \ln(\mathcal{L}_q(\theta,\sigma_q))\right).$$

In practice, we minimize the negative overall log-likelihood using the Nelder-Mead (aka downhill simplex) algorithm (Nelder and Mead, 1965). Parameter bounds (reported in Supplementary Table S2) are enforced through parameter transformations. Due to the presence of local minima, optimizations are repeated for 2000 sets of uniformly drawn starting points to ensure convergence to a global minimum. The best MLE set of estimates $\hat{\theta}$ corresponds to the fit associated with the lowest negative overall log-likelihood with successful completion.

2.3.3 Confidence intervals

We generate bootstrapped data to compute 95% CIs of our parameters using Sieve bootstrap (Bühlmann, 1997; Ulloa et al., 2013) on the residuals between experimental data and the best fit of our model. For this purpose, autoregressive moving-average (ARMA) models are fitted to the time series of centred residuals of each chemostat independently. We then use ARMA models to simulate new residuals from which we reconstruct new datasets. We eventually reiterate the above estimation procedure (cf. Materials and methods section '2.3.1 Estimation of the rates of prophage reactivation' to '2.3.2 Estimation of the remaining parameters'), but starting non-linear optimizations only from the best MLE estimates we obtained with the original data. By repeating this for 10 000 bootstrapped datasets, we compute the joint distributions of estimated parameters.

2.4 Details of the implementation

Numerical simulations and data analyses were carried out using R (R Core Team, 2022) version 4.2.0 (2022-04-22). ODEs were solved numerically by the function `ode`—with method `lsoda`—from the package `deSolve` (Soetaert et al., 2010). Non-linear optimizations for MLE were tackled with the function `nmk` (Kelley, 1999), from the package `dfoptim`, which gave here more stable results than `optim` from base R. Fit and selection of ARMA models for Sieve bootstrap were carried out using the function `auto.arima`, from the package `forecast` (Hyndman and Khandakar, 2008).

3. Results

3.1 Theoretical analysis

When $r > \delta$, the virus-free system converges to an equilibrium where $S(\infty) = K(1 - \delta/r)$. When a single strain $k \in \{w, m\}$ of the virus (with phenotypes α_k and ϕ_k) is introduced in the bacterial population (fully susceptible, with density S_0), the fate of the phage-bacteria system depends on the basic reproduction number of the pathogen $\mathcal{R}_{0,k}$ which is given by:

$$\mathcal{R}_{0,k} = \frac{A + \sqrt{A^2 - 4r\left(1 - \frac{S_0}{K}\right)(1-\phi_k) a b S_0 \tau B (aS_0 + \delta)(\alpha_k + \delta)(\tau + \delta)}}{2(aS_0 + \delta)(\alpha_k + \delta)(\tau + \delta)}, \quad (4)$$

with $A = r\left(1 - \frac{S_0}{K}\right)(aS_0 + \delta)(\tau + \delta) + abS_0\tau B(\alpha_k + (1 - \phi_k)\delta)$ [see Supplementary Appendix S2.1 for the construction of the next-generation matrix, following Diekmann et al. (2010)]. When $\mathcal{R}_{0,k} < 1$, the virus goes extinct and the bacterial population converges to the previous virus-free equilibrium. Alternatively, when $\mathcal{R}_{0,k} > 1$, an epidemic breaks out and eventually stabilizes to an endemic equilibrium where all the cells are infected by the virus (Supplementary Appendix S2.1).

In the following, we analyse the evolutionary dynamics (i) at the beginning of an epidemic where $S(t) = S_0$ and (ii) when the system stabilizes towards the endemic equilibrium where $S(t) = 0$.

3.1.1 *Evolution in an emerging epidemic*

At the onset of the epidemic, susceptible cells are highly abundant. For the sake of simplicity, we analyse the dynamics of the epidemic when the host density is assumed to be constant over time ($S(t) = S_0$). As in the experimental design, we start with a bacterial population at carrying capacity ($S_0 \approx K$). Density dependence reduces cell reproduction and vertical transmission of the virus. Consequently, the epidemic is mainly driven by the lytic pathway. At $t = 0$ though, only lysogens are introduced at very low density. After a very short time, the phage-bacteria system reaches its new dynamical regime, following prophage reactivations, lyses and releases of virions. From there, we assume that $L(t)/Y(t) \approx 0$ and we focus on the dynamics of the mutant in the compartment Y. We derive an approximation of the selection gradient $\mathcal{S}$ of the virulent phage, which corresponds to the rate at which it increases in frequency on the logit scale, and we show in Supplementary Appendix S2.2 that:

$$\mathcal{S} \propto -\Delta\phi, \quad (5)$$

meaning that it is selected for ($\Delta\phi < 0$, cf. Supplementary Fig. S3A). This is consistent with the experimental data where the virulent phage transiently outcompetes the wildtype at the early stage of the epidemic treatment (Berngruber et al., 2013). Our prediction is accurate when the density of susceptible hosts remains effectively constant over time; otherwise, however, our prediction deviates from the simulation whose rate slows down because the density of susceptible cells is rapidly decreased by the spread of the epidemic (Fig. 2A and Supplementary Fig. S3C).

3.1.2 *Evolution at the end of the epidemic*

At the end of the epidemic, we expect that all the cells will be infected by a prophage and there will no longer be any susceptible cells. Consequently, no horizontal transmission takes place and, in contrast with the previous scenario, we can neglect the density of Y cells relative to the density of lysogens. Indeed, since $1/\alpha_w \gg 1/\tau$ (time elapsed between phage integration and reactivation is much longer than lysis time), then $Y(t)/L(t) \approx 0$ (see details in Supplementary Appendix S2.3). Note that this also means that the frequency infected by the virulent strain is mainly driven by the corresponding frequency in L, that is, $g(t) \approx p(t)$. We show in Supplementary Appendix S2.3 that the selection gradient of the virulent phage is given by:

$$\mathcal{S} = -\Delta\alpha. \quad (6)$$

As a result, the virulent phage is counter-selected in the long-term ($\Delta\alpha > 0$) and, in each compartment, linearly declines in frequency at a rate $\mathcal{S}$ (negative slope) on the logit scale (Fig. 2B and Supplementary Fig. S1B).

3.1.3 *Differentiation across compartments*

The two previous subsections deal with the dynamics of evolution over time; let us now look at the dynamics of selection between different compartments. We define here the differentiation $\mathcal{Q}^{VL}$ between the free virus stage and lysogens as:

$$\mathcal{Q}^{VL}(t) = \frac{q(t)}{1 - q(t)} \frac{1 - p(t)}{p(t)}, \quad (7)$$

such that:

$$\ln\left(\mathcal{Q}^{VL}(t)\right) = \mathrm{logit}(q(t)) - \mathrm{logit}(p(t)).$$

We show in Supplementary Appendix S2.4 that, at the end of the epidemic, the differentiation $\mathcal{Q}^{VL}(t)$ converges approximately towards:

$$\mathcal{Q}^{VL} \approx \frac{\alpha_m}{\alpha_w} = 1 + \frac{\Delta\alpha}{\alpha_w}, \quad (8)$$

(Fig. 2C), a value of 1 corresponding to no differentiation. We obtain the same expression for the convergence of the differentiation between Y and L cells and around 1 between the free virus stage V and Y cells (cf. Supplementary Appendix 2.4). Interestingly, these results are also valid at $t = 0^+$ (Supplementary Fig. S4) as there are no free phage particles yet and thereby no horizontal transmission which, within a very short space of time, resembles the end of the epidemic. When the system stabilizes at the end of

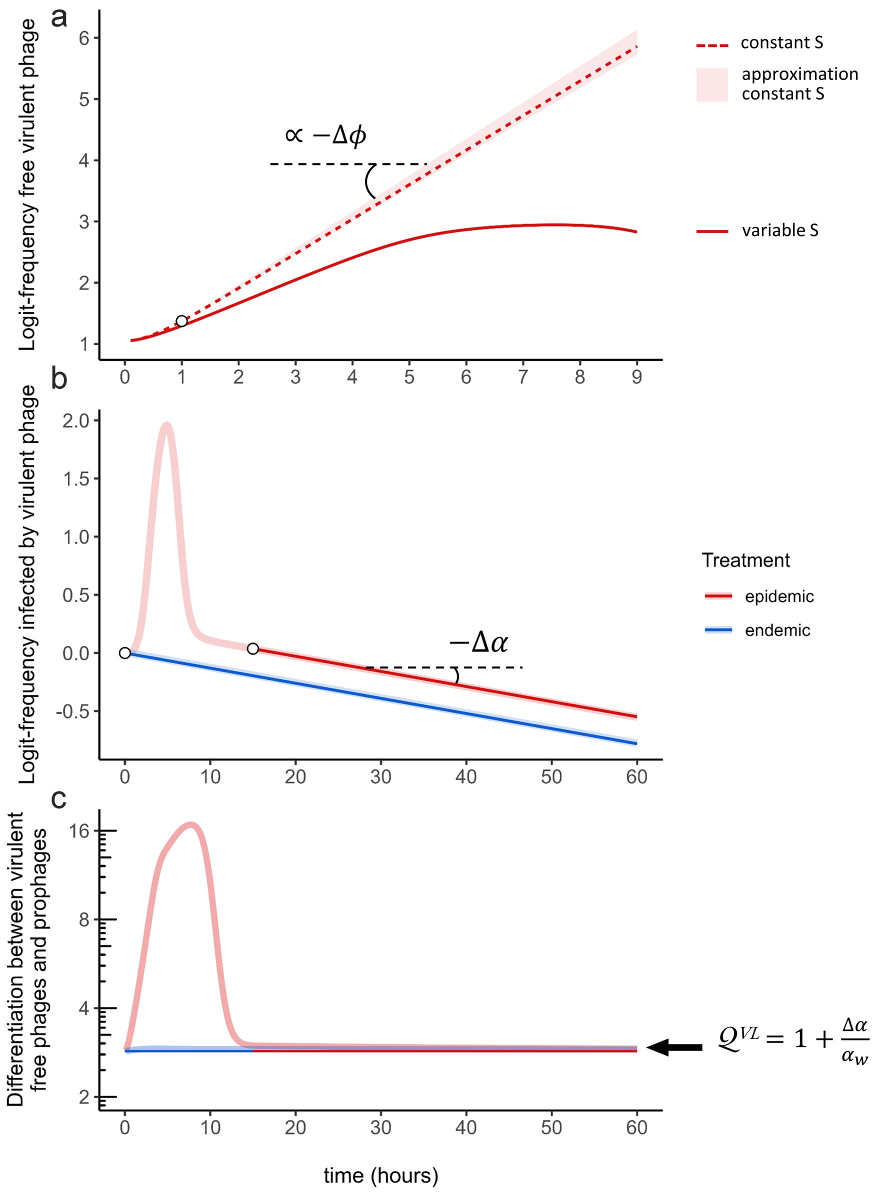


Figure 2. Theoretical predictions. At $t = 0$, bacteria are at carrying capacity K with initial prevalence 1% (epidemic treatment) or 99% (endemic treatment). The initial prophage ratio for the two strains is 1:1. See Supplementary Table S1 for parameter values. (a) The virulent phage is selected for at the early stage of the epidemic. Our approximation for the trajectory of the logit-frequency of the virulent phage (light red shading, see Supplementary Appendix S2.2) predicts the dynamics of the virulent phage when the density of susceptible cells is kept constant (dashed red line). The slope of the upper bound of our approximation is proportional to $-\Delta\phi = \phi_w - \phi_m$ (equation (5)). As only lysogens are introduced in the susceptible bacterial population at $t = 0$, we let the phage-bacteria system reach its new dynamical regime and only start our approximation at $t = 1$ (white dot). However, the rapid drop in the availability of susceptible hosts due to the viral epidemic reduces the increase of the virulent phage relative to our approximation (compare the full red line and the dashed red line). (b) The virulent phage is counter-selected when the epidemic reaches high prevalence (indicated with the white dots in the epidemic and endemic treatments). At the end of the epidemic, the logit-frequency of the virulent phages decreases linearly with negative slope $-\Delta\alpha = \alpha_w - \alpha_m$ (equation (6)). (c) The virulent strain is more frequent among free viruses (V) than among prophages (L). When the system reaches high prevalence, we predict the differentiation between these two compartments to converge towards an equilibrium that is approximately equal to $1 + \Delta\alpha/\alpha_w = \alpha_m/\alpha_w$ (equation (8)).

the epidemic, the virulent strain is therefore more frequent among free viruses and Y cells than among L cells (prophages) but we also expect almost no differentiation between free viruses and Y cells (Supplementary Fig. S4).

3.2 Statistical inference

3.2.1 *Inference from simulated data*

We first conduct a simulation study. We start by validating our estimation method of the rates of prophage reactivation α_w and α_m. Combining equations (6) and (8), we compute point estimates of both parameters (cf. Materials and methods section '2.3.1 Estimation of the rates of prophage reactivation') for a large number of simulated datasets. Overall, estimated values show a very good match with those used in the original simulations (Supplementary Fig. S5), especially when the SD of measurement errors is small.

We then show that parameters b and B may not be separately identifiable using the simulated dataset closest to the original deterministic simulation (sampling effort = 10 h^{-1} and SD of measurement error = 0.01, see Supplementary Fig. S2A). We fix K

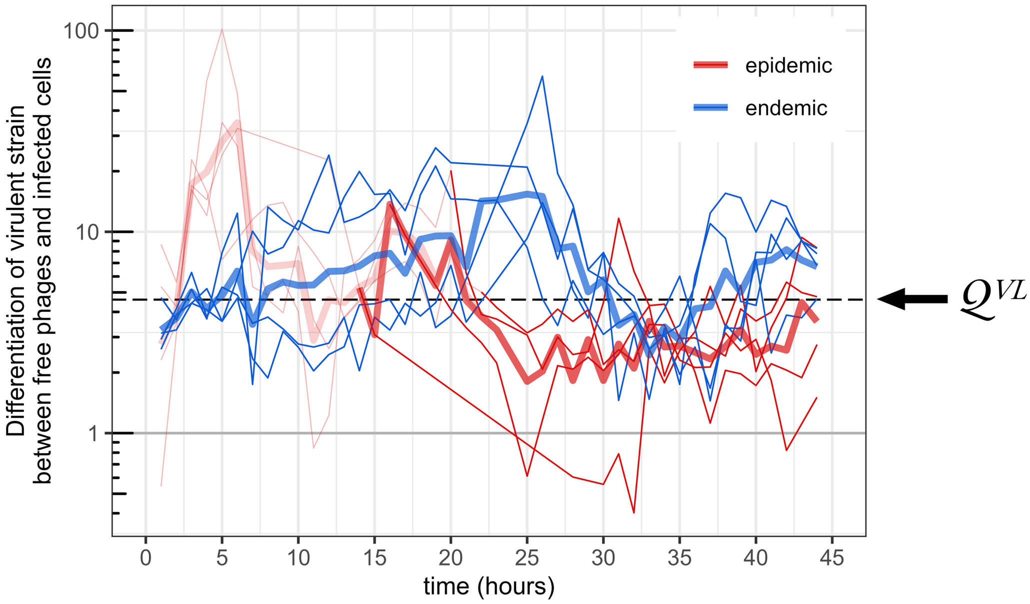


Figure 3. Differentiation of the virulent strain between free phages and infected cells. We compute $q(t)(1-g(t))/((1-q(t))g(t))$ for each chemostat (thin lines) and the corresponding geometric mean for each treatment (thick lines). The value 1 (horizontal grey line) corresponds to no differentiation. For each chemostat, values before the time point the system has reached a prevalence of 95% are shown in transparent color. With high prevalence, we have $g(t) \approx p(t)$ and we thereby obtain the differentiation of the virulent strain between free phage particles (V) and prophages (L). In these conditions, such differentiation is expected to reach an equilibrium—the horizontal dashed line indicates the corresponding geometric mean (4.60) across all chemostats and time points—with theoretical value $1+\Delta\alpha/\alpha_w = \alpha_m/\alpha_w$ (cf. Fig. 2C).

and δ to their known values, as well as α_w and α_m, as though the rates of prophage reactivation have been correctly estimated beforehand. Setting (b, B) to many combinations of values ($(b, B) \in [0,\ 0.2] \times [0,\ 100]$) and maximizing over the remaining parameters, the resulting landscape of the overall log-likelihood suggests that only the product $b \times B$ is identifiable (Supplementary Fig. S6). Such compensatory effect may be understood as the impossibility here of discriminating whether phage λ infect fewer bacteria (lower probability b of fusion) but release more viral copies upon lysis (larger burst size B) or vice-versa. This is the reason why we fix B in the following and only estimate b.

Fixing B to its true value used in the original simulations, point estimates of parameters are quite close to their true values (Supplementary Fig. S7). However, some parameters—like r, τ or b—seem more difficult to estimate with a lower sampling effort.

3.2.2 *Inference from experimental data*

The original model used in Berngruber et al. (2013) failed to properly capture the evolutionary dynamics among infected hosts at the early stage of the epidemic. Such discrepancy was an opportunity to go back from experimental data to theory and to challenge the structure of the model. We noticed that adding an extra stage Y in the infected compartment allowed us to better fit the experimental data than the original model (ΔAIC= −101). This significant improvement strongly supported the new version of the model developed in this study and for which we present the estimation results below.

First, we estimate: $\Delta\alpha = 9.31 \times 10^{-3}\ \mathrm{h}^{-1}$ and $\mathcal{Q}^{VL} = 4.60$ (see Fig. 3 for the latter). Point estimates of the rates of prophage reactivation (expressed in h^{-1}) are thus: $\hat{\alpha}_w = 2.58 \times 10^{-3}$ and $\hat{\alpha}_m = 1.19 \times 10^{-2}$.

Second, we perform non-linear optimizations to estimate the remaining parameters. Overall, computed dynamics of the logit-prevalence and logit-frequencies of the virulent phage closely fit experimental data (Fig. 4). Best MLE estimates are listed in Table 2, along with their 95% bootstrap-based CIs (see Supplementary Fig. S8 for the density distributions of estimated parameters). We show pairwise correlation coefficients in Supplementary Fig. S9. Some parameters are correlated positively, in particular: α_w with α_m, p_0 with both reactivation rates, ϕ_m with $P_0^{epidemic}$ and r with b; or negatively, in particular: τ with b and r. In Supplementary Fig. S10, we plot the 95% distributions of fitted values. To test the impact of the fixed value used for the burst size B, we study how perturbations of the original value (80 virus cell^{-1}) affect the estimation of the other parameter values (new point estimates from non-linear optimizations). In Supplementary Fig. S11, we show that all parameters are extremely robust to these perturbations, except for parameter b as expected from Supplementary Fig. S6, (but the product $b \times B$ is always around 3.95). It is also worth noting that the basic reproduction number is not affected by the choice of the value of B, as it always appears as a product with b in (4). We estimate the basic reproduction number of the wildtype strain $\mathcal{R}_{0,w}$ to be 1.48 (95% CI [1.04, 2.15]) and of the virulent strain $\mathcal{R}_{0,m}$ to be 2.20 (95% CI [1.58, 3.05]). Interestingly, the basic reproduction number of the virulent mutant $\mathcal{R}_{0,m}$ is higher than that of the wildtype $\mathcal{R}_{0,w}$, but the virulent mutant always loses the competition in the long term. The basic reproduction number gives the expected number of secondary infections in an otherwise fully susceptible host population (Anderson and May, 1991). The generation interval distribution of the two strains being similar for the lytic cycle, the basic reproduction number provides a good predictor of the relative fitness of the two strains at the early stage of the epidemic (Wallinga and Lipsitch, 2007; Park et al., 2019): the mutant outcompetes the wildtype as $\mathcal{R}_{0,m} > \mathcal{R}_{0,w}$. When the density of susceptible host cells drops, however, the basic reproduction number becomes a poor predictor of fitness. Eventually, only the strain that tolerates the worst environment in terms of resource density—in this case, the lowest density of susceptible host cells—survives (pessimization principle) (Diekmann, 2004). The wildtype replicates better at lower S cells densities than

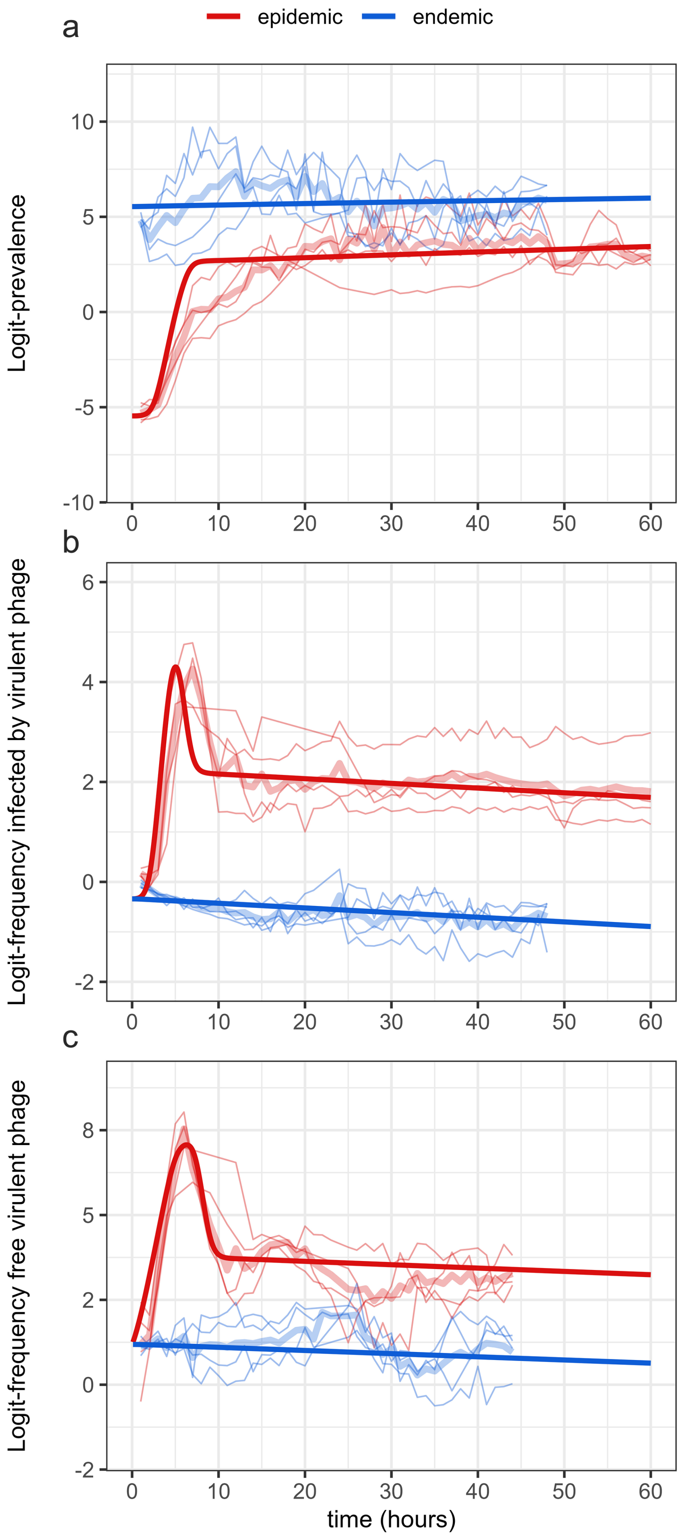


Figure 4. Fitted values on experimental data. Fitted values (thick dark lines) were obtained from the simulation based on the best MLE estimations (see Table 2). We estimate model parameters using experimental data (light-colored thin lines) from the evolution experiment designed in Berngruber et al. (2013); initial prevalence is either low (epidemic condition, in red) or high (endemic condition, in blue). Light-colored thick lines correspond to mean logit-values across all replicates per treatment. (a) Logit-prevalence of infection; (b) logit-frequency of cells infected (involved either in a lysogenic or a lytic cycle) by the the mutant (virulent) phage; (c) logit-frequency of the virulent phage in the culture medium (free virus stage).

Table 2. MLE parameter estimations. In a first step, parameters α_w and α_m are estimated directly using analytic approximations and linear models. In a second step, we fix α_w and α_m to their previous point estimates and, starting from 2000 sets of initial values, non-linear optimizations maximizing the overall log-likelihood with the experimental data are run to estimate the remaining parameters. 95% CIs are based on Sieve bootstrap (6686 sets of estimates). Throughout, fixed parameters are: $K = 10^9$ cells, $\delta = 0.8\,h^{-1}$ and $B = 80$ virus.cell^{-1}.

Parameter	Estimated value	95% CI	Unit
Parameters of the phage			
α_w	2.58×10^{-3}	$[1.94 \times 10^{-4},\ 7.87 \times 10^{-3}]$	h^{-1}
α_m	1.19×10^{-2}	$[1.03 \times 10^{-3},\ 2.85 \times 10^{-2}]$	h^{-1}
ϕ_w	0.347	[0.198, 0.501]	–
ϕ_m	2.91×10^{-2}	$[7.70 \times 10^{-3},\ 6.94 \times 10^{-2}]$	–
a	1.00×10^{-6}	$[4.34 \times 10^{-7},\ 1.00 \times 10^{-6}]$	h^{-1}.cell^{-1}
b	4.94×10^{-2}	$[2.44 \times 10^{-2},\ 7.51 \times 10^{-2}]$	–
τ	1.08	[0.83, 3.95]	h^{-1}
Parameter of the bacteria			
r	1.60	[0.77, 5.00]	h^{-1}
Initial conditions			
$P_0^{epidemic}$	4.24×10^{-3}	$[1.00 \times 10^{-3},\ 1.13 \times 10^{-2}]$	–
$P_0^{endemic}$	99.61×10^{-2}	[0.988, 0.998]	–
p_0	0.416	[0.400, 0.517]	–
SD of measurement errors			
σ_P	1.51	[1.13, 2.00]	–
σ_g	0.52	[0.52, 0.75]	–
σ_q	0.83	[0.74, 1.16]	–

the virulent strain because it relies more on vertical transmission and less to transmission to new susceptible cells. $\mathcal{R}_0$ is not maximized in the long-term because the 'niche' of the virus is multidimensional, again because of the two modes of replication (Lion and Metz, 2018).

We compared our estimates of model parameters with previous studies (see reported values in Supplementary Table S3), focusing in particular on the parameters of the phage. Note that previous estimations of these life-history traits were obtained from experimental assays using monomorphic phage cultures. Our estimate of the rate of prophage reactivation of the wildtype strain α_w falls in the range of previous *in vitro* estimates, which is very large [between 10^{-7} and $10^{-2}\,h^{-1}$ (De Paepe et al., 2016; Little et al., 1999; Zong et al., 2010)], probably due to the variability in experimental methods. For the cI variant, the orders of magnitude of α_m are more consistent (Zong et al., 2010; De Paepe et al., 2016). Other studies, using single-cell monitoring of infected cells to estimate the probability of lysogenization of the wildtype strain ϕ_w, obtained very similar results when the multiplicity of infection (MOI) is low, around 0.35 on average (Zeng et al., 2010)—but see the section 'Discussion' for the effect of MOI. We estimate a 12-fold decrease for the probability of lysogenization of the virulent strain ϕ_m and, to our knowledge, this parameter has not been estimated elsewhere. The lysis time $1/\tau$, estimated close to 1 h (though with larger uncertainty), is also consistent with the biology of phage λ (Zong et al., 2010; De Paepe and Taddei, 2006; De Paepe et al., 2016; Lindberg et al., 2014; Shao and Wang, 2008). On the other hand, the adsorption rate a is poorly estimated. Although this issue did not arise with simulated data, our estimates based on experimental data are mostly stuck at the upper bound (10^{-6}), several orders of magnitude higher than expected (De Paepe and Taddei, 2006; De Paepe et al., 2016; Lindberg et al., 2014; Shao and Wang, 2008). To investigate the sensitivity of the other estimates to the estimated value of a, we reiterate non-linear optimizations fixing a to different values ranging from 10^{-9} to 1 to compute new point estimates. The inference of most parameters was robust to the value of a, with no relative variation higher than 50% found for values of a between 10^{-8} to 1 (Supplementary Fig. S12A). We carry out the same sensitivity analysis for the intrinsic growth rate of the host r which was also poorly estimated (its distribution being mainly flat): with values of r ranging between 1 and 5, the probabilities of lysogenization are little affected by the perturbations in r (relative variation contained between -7% and +15%) while parameters b and τ—the most correlated with r—are the most sensitive (Supplementary Fig. S12B).

4. Discussion

In a previous study, Berngruber et al. (2013) carried out an evolution experiment with the phage λ to validate several theoretical predictions on the evolution of virulence and transmission. This study used a two-step approach: (i) a mathematical model tailored to the biology of phage λ was used to show how epidemiological dynamics feedback on the evolution of the pathogen and (ii) tracking the variation of the frequency of a viral mutant across time and across compartments (infected host and free virus) confirmed the impact of the density of susceptible cells on the transient evolution of the virus at different stages of its life cycle. This experimental validation of the theory focused on the qualitative match of the experimental results with the predictions of the model and provided empirical support for evolutionary epidemiology theory. In the present study, we adopt a reverse approach where we start from the data and try to improve the theoretical model developed to describe the evolutionary epidemiology of phage λ. We calculate the selection gradients and the differentiation across the life stages of the virus at the onset and at the end of the epidemic. This analysis allows us to better quantify the processes driving the evolutionary dynamics of the virus population.

First, we noticed that the previous model failed to properly capture the transient evolutionary dynamics in the infected compartment. We improved the goodness of fit with the experimental data by distinguishing two types of infected cells: lysogens (L) and cells committed to the lytic pathway (Y). Indeed, adding extra stages allowed us to observe a transitory peak in the frequency of hosts infected by the virulent phage (Supplementary Fig. S13A, middle panel), similar to what we observed in the experimental data. In this model, the peak is due to the frequency of the virulent phage among Y bacteria that, unlike L bacteria, transiently overshoots during the acute phase of the epidemic. Crucially, lysis was assumed to be instantaneous in the original model (Berngruber et al., 2013) and including this additional stage in the phage life cycle allowed us to take the lag between infection and lysis into account (Yin and McCaskill, 1992; You and Yin, 1999; Mitarai et al., 2016; Brown et al., 2022; Geng et al., 2023; De Paepe et al., 2016). For the sake of simplicity and parsimony, we only kept a single extra stage Y, yielding exponentially distributed lysis time (Mitarai et al., 2016).

Second, we performed a theoretical analysis of this new model to go beyond the numerical approach used in Berngruber et al. (2013). In particular, we obtained analytic approximations for the spread of the virulent phage at the beginning and at the end of the epidemic: (i) the virulent mutant increases in frequency at the onset of the epidemic (when the density of susceptible hosts is high) with a speed approximately proportional to $-\Delta\phi$; (ii) the virulent mutant decreases in frequency at the end of the epidemic (when the density of susceptible hosts is low) with a negative selection gradient $-\Delta\alpha$, which is consistent with the prediction that in the long-term, and in the absence of an influx of susceptible hosts, the temperate phage should evolve a fully lysogenic strategy where $\alpha=0$ and $\phi=1$ (Bruce et al., 2021; Wahl et al., 2019); and (iii) the virulent phage is always more frequent in the free virus stage than among prophages and this differentiation is approximated by $1+\Delta\alpha/\alpha_w$ at the end of the epidemic. The first approximation only held for a very limited period of time because the density of susceptible cells drops very rapidly and this epidemiological feedback affects the selection on transmission and virulence. In contrast, the final two predictions were valid as soon as the epidemic reaches high prevalence and, interestingly, yielded a novel way to estimate the rates of prophage reactivation of both strains.

Third, we developed a statistical inference approach to estimate the parameters of the model. We implemented a two-step MLE procedure: (i) we first obtained point estimates of the reactivation rates α_w and α_m of both viral strains and (ii) we then fixed α_w and α_m to their point estimates and ran non-linear optimizations to infer the remaining parameters of the model (except the burst size B which had to be fixed because of an identifiability issue). The Sieve bootstrap technique was used to generate joint distributions for all the estimated parameters (Table 2). We showed that our estimates of the key phenotypic traits α_w, α_m, and ϕ_w were consistent with existing literature (Supplementary Table S3).

Crucially, our inference approach is quite different from previous studies. First, we analyse the dynamics of a polymorphic viral population. Second, we use three types of data to infer the parameter values of the model: (i) epidemiological data (i.e. the prevalence of the infection), (ii) temporal variations in variant frequencies, and (iii) differentiation of the variant frequency across compartments. Each data type carries complementary information; together, they allow us to jointly estimate the life-history traits of both strains of the phage. This novel method is particularly well suited to estimate the rates of prophage reactivation, for which only the endemic treatment is needed—see Supplementary Appendix S3.2 where we propose a Bayesian counterpart to easily obtain posterior distributions and credible intervals for these two parameters (applied to our experimental data, we show results in Supplementary Fig. S14A). Estimating the probabilities of lysogenization is however more difficult as it requires to monitor the transient dynamics of the epidemic treatment. The rapid epidemiological feedback that occurred in emerging epidemics makes it necessary to consider both the epidemiology and the evolution of the infection. It might thus be relevant to carry out shorter experiments for the epidemic treatment—i.e. focusing on the early state of the epidemic—but with a more frequent sampling effort to monitor the change in frequency more precisely. Alternatively, the use of two-stage chemostats, where the influx of susceptible hosts could be maintained experimentally (Bonachela and Levin, 2014; Husimi et al., 1982), might also be an option worth investigating.

Even though we have improved the original model, many features of the present model could be challenged. For example, we assume all phenotypic traits to be constant across time and across chemostats. Yet, key life-history traits of phage λ are known to vary with the environment (Wegrzyn and Wegrzyn 2005) (i.e. phenotypic plasticity). In particular, the probability of lysogenization and the reactivation rate of the λ mutant used in the evolution experiment is temperature-sensitive (Berngruber et al., 2013; St-Pierre and Endy, 2008; Zong et al., 2010). Interestingly, the estimated rates of reactivation varied among chemostats (Supplementary Fig. S14B, using a Bayesian approach). This variation may have been driven by small variations in temperature among the chemostats that could have affected the reactivation rate of the mutant (higher rates of reactivation would be expected with higher temperature). The probability of lysogenization of phage λ is also known to be more likely to occur at high MOI (Kourilsky, 1973; Kourilsky and Knapp, 1975; Zeng et al., 2010). MOI-dependent phenotypic plasticity may also affect the rate of reactivation of temperate phages (Bruce et al., 2021), or the lysis time of virulent phages (Rutberg and Rutberg, 1965). Yet, our model has the advantage of the parsimony, while already convincingly reproducing the qualitative patterns observed in the data. More sophisticated models would improve the fit to the data compared to simpler, nested models. Additional details may be rooted in experimental knowledge on the system, but the data may not be rich enough to infer the extra-parameters precisely (and accordingly, these more complex models may not be selected over simpler ones in statistical model comparisons). Moreover, simpler models are easier to analyse mathematically and lend themselves more easily to interpretation. It is thus delicate to know where to draw the line between the parsimony and the goodness of fit of a model. Navigating in this trade-off, a first model was developed capturing the transient increase in virulent strain (Berngruber et al., 2013), and we extended it with the known delay between phage infection and lysis to better capture the similarity in variant frequency in cells and free viruses. Our final model fits the data well, but not all parameters can be inferred precisely, in ways that depend on the details of the model and would have been difficult to anticipate before formally fitting the model.

Our study emphasizes the benefits of combining theoretical, statistical, and experimental approaches. Combining these perspectives effectively requires an iterative process. Each step is an opportunity to challenge and improve our understanding of the biological model, to elucidate microbial life cycles and to provide support and guidance during experimentation. While most experimental life-history assays only focus on monomorphic population, we demonstrate the relevance of evolution experiments where different strains are put in competition and where tracking their relative frequencies over time may yield novel and efficient ways to estimate some dimensions of their phenotypes, especially when it is difficult to access absolute densities. This approach could be used under different conditions and in particular *in vivo* (De Paepe et al., 2016). The present study may thus help characterize the phenotypic traits of microbial organisms used in experimental evolution. More broadly, this overall quantitative process is similar in spirit with what can be done in epidemiological studies, particularly in public health—though generally with a much simpler description of the life cycle. During the COVID-19 pandemic, for instance, the successive emergence and sweeps of SARS-CoV-2 variants of concern has raised many questions about the short- and long-term evolution of the virus, especially in terms of transmission and virulence. Statistical inference based on demographic/epidemiological data (e.g. prevalence, deaths, hospital admissions) and genetic data (e.g. strain frequencies) quantified variants properties such as transmission, virulence, infectious period, immune escape, or generation interval distribution [e.g. Davies et al. (2021); Blanquart et al. (2022); Benhamou et al. (2023)]. In many scenarios, the frequency of a variant could be measured in different compartments, for example between naive and vaccinated hosts (e.g. for SARS-CoV-2 variants Omicron), or between infected hosts and an environmental compartment (e.g. enteropathogenic *E. coli* in environmental waters). As shown in this study, interpreting the temporal variation in the selection gradient and the differentiation of a novel variant between distinct compartments may be a powerful tool to clarify the relationship between its frequencies in distinct life stages and to better characterize its phenotype.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

Funding

We acknowledge the French Ministry of Higher Education and Research and the École Normale Supérieure Paris-Saclay for the PhD scholarship of WB.

Data availability

Data that were used in this study, along with the scripts for the analyses (in R), are available in this GitHub repository as well as at https://zenodo.org/records/13138349.

References

Alizon S, Hurford A, Mideo N, *et al.* Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. *J Evol Biol* 2009;**22**:245–259.

Alizon S, Michalakis Y. Adaptive virulence evolution: the good old fitness-based approach. *Trends Ecol Evol* 2015;**30**:248–254.

Anderson RM, May RM. Coevolution of hosts and parasites. *Parasitology* 1982;**85**:411–426.

Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1991.

Benhamou W, Lion S, Choquet R, *et al.* Phenotypic evolution of SARS-CoV-2: a statistical inference approach. *Evolution* 2023;**77**:2213–2223.

Berngruber TW, Froissart R, Choisy M, *et al.* Evolution of virulence in emerging epidemics. *PLoS Pathogens* 2013;**9**:e1003209.

Berngruber TW, Weissing FJ, Gandon S. Inhibition of superinfection and the evolution of viral latency. *J Virol* 2010;**84**:10200–10208.

Blanquart F, Hozé N, Cowling BJ *et al.* Selection for infectivity profiles in slow and fast epidemics, and the rise of SARS-CoV-2 variants. *Elife* 2022;**11**:e75791.

Bonachela JA, Levin SA. Evolutionary comparison between viral lysis rate and latent period. *J Theor Biol* 2014;**345**:32–42.

Brown S, Mitarai N, Sneppen K. Protection of bacteriophage-sensitive *Escherichia coli* by lysogens. *Proc Natl Acad Sci*, 2022, **119**:e2106005119.

Bruce JB, Lion S, Buckling A *et al.* Regulation of prophage induction and lysogenization by phage communication systems. *Curr Biol* 2021;**31**:5046–5051.

Bühlmann P. Sieve bootstrap for time series. *Bernoulli* 1997; 123–148.

Bull JJ. Virulence. *Evolution* 1994;**48**:1423–1437.

Davies NG, Abbott S, Barnard RC, *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. *Science* 2021;**372**:eabg3055.

Day T. On the evolution of virulence and the relationship between various measures of mortality. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **269**, 1317–1323, 2002.

Day T, Proulx SR. A general theory for the evolutionary dynamics of virulence. *Am Naturalist* 2004;**163**:E40–E63.

De Paepe M, Taddei F. Viruses' life history: towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS Biol* 2006;**4**:e193.

De Paepe M, Tournier L, Moncaut E *et al.* Carriage of λ latent virus is costly for its bacterial host due to frequent reactivation in monoxenic mouse intestine. *PLoS Genet* 2016;**12**:e1005861.

Diekmann O. *A Beginner's Guide to Adaptive Dynamics*, Vol. 63, Banach Center Publications, 2004, 47–86.

Diekmann O, Heesterbeek J, Roberts MG. The construction of next-generation matrices for compartmental epidemic models. *J R Soc Interface* 2010;**7**:873–885.

Echols H. Developmental pathways for the temperate phage: lysis vs lysogeny. *Annu Rev Genet* 1972;**6**:157–190.

Frank SA. Models of parasite virulence. *Q Rev Biol* 1996;**71**:37–78.

Gandon S. Why be temperate: lessons from bacteriophage λ. *Trends Microbiol* 2016;**24**:356–365.

Gandon S, Day T. The evolutionary epidemiology of vaccination. *J R Soc Interface* 2007;**4**:803–817.

Geng Y, Nguyen TVP, Homaee E, *et al.* Using population dynamics to count bacteriophages and their lysogens. *bioRxiv* 2023; 2023–10.

Husimi Y, Nishigaki K, Kinoshita Y, *et al.* Cellstat—a continuous culture system of a bacteriophage for the study of the mutation rate and the selection process at the DNA level. *Rev Sci Instrum* 1982;**53**:517–522.

Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for r. *J Stat Softw* 2008;**27**:1–22.

Kelley CT. *Iterative Methods for Optimization*. SIAM, 1999.

Kourilsky P. Lysogenization by bacteriophage lambda: I. multiple infection and the lysogenic response. *Mol Gen Genet* 1973;**122**:183–195.

Kourilsky P, Knapp A. Lysogenization by bacteriophage lambda: III. Multiplicity dependent phenomena occurring upon infection by lambda. *Biochimie* 1975;**56**:1517–1523.

Lenski RE, May RM. The evolution of virulence in parasites and pathogens: reconciliation between two competing hypotheses. *J Theor Biol* 1994;**169**:253–265.

Lindberg HM, McKean KA, Wang I-N. Phage fitness may help predict phage therapy efficacy. *Bacteriophage* 2014;**4**:e964081.

Lion S, Metz JA. Beyond R0 maximisation: on pathogen evolution and environmental dimensions. *Trends Ecol Evol* 2018;**33**:458–473.

Little JW, Shepley DP, Wert DW. Robustness of a gene regulatory circuit. *EMBO J* 1999;**18**:4299–4307.

Lwoff A. Lysogeny. *Bacteriol Rev* 1953;**17**:269–337.

Mitarai N, Brown S, Sneppen K. Population dynamics of phage and bacteria in spatially structured habitats using phage λ and *Escherichia coli*. *J Bacteriol* 2016;**198**:1783–1793.

Nelder JA, Mead R. A simplex method for function minimization. *Comput J* 1965;**7**:308–313.

Park SW, Champredon D, Weitz JS, *et al.* A practical generation-interval-based approach to inferring the strength of epidemics from their speed. *Epidemics* 2019;**27**:12–18.

Ptashne M. *A Genetic Switch: Phage λ and Higher Organisms*. Oxford: Blackwell Publishers, 1992.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022.

Rutberg B, Rutberg L. Role of superinfecting phage in lysis inhibition with phage t4 in *Escherichia coli*. *J Bacteriol* 1965;**90**: 891–894.

Shao Y, Wang I-N. Bacteriophage adsorption rate and optimal lysis time. *Genetics* 2008;**180**:471–482.

Soetaert K, Petzoldt T, Setzer RW. Solving differential equations in R: package deSolve. *J Stat Softw* 2010;**33**:1–25.

St-Pierre F, Endy D. Determination of cell fate selection during phage lambda infection. *Proc Natl Acad Sci*, 2008;**105**:20705–20710.

Sussman R, Jacob F. On a thermosensitive repression system in the *Escherichia coli* lambda bacteriophage. *Comptes Rendus* *héBdomadaires des séAnces de l'Académie des sciences* 1962;**254**: 1517–1519.

Ulloa G, Allende-Cid H, Allende H. Sieve bootstrap prediction intervals for contaminated non-linear processes. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Proceedings, Part I*, **18**, pp.84–91. Springer: Havana, Cuba, 2013.

Wahl LM, Betti MI., Dick DW *et al*. Evolutionary stability of the lysis-lysogeny decision: why be virulent?. *Evolution* 2019;**73**:92–98.

Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc R Soc B: Biol Sci*, 2007; **274**:599–604.

Wang I-N. Lysis timing and bacteriophage fitness. *Genetics* 2006;**172**:17–26.

Wegrzyn G, Wegrzyn A. Genetic switches during bacteriophage λ development. *Prog Nucleic Acid Res Mol Biol* 2005;**79**: 1–48.

Yin J, McCaskill J. Replication of viruses in a growing plaque: a reaction-diffusion model. *Biophys J* 1992;**61**: 1540–1549.

You L, Yin J. Amplification and spread of viruses in a growing plaque. *J Theor Biol* 1999;**200**:365–373.

Zeng L, Skinner SO, Zong C *et al*. Decision making at a subcellular level determines the outcome of bacteriophage infection. *Cell* 2010;**141**:682–691.

Zong C, So L, Sepúlveda LA *et al*. Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. *Mol Syst Biol* 2010;**6**:440.

Virus Evolution, 2024, 10(1), veae069 , DOI: https://doi.org/10.1093/ve/veae069, Advance access publication date: 15 October 2024, Research Article

— Supplementary Figures and Tables —

Evolution of virulence in emerging epidemics: from theory to experimental evolution and back

Wakinyan Benhamou, François Blanquart, Marc Choisy, Thomas W. Berngruber, Rémi Choquet and Sylvain Gandon

May 23, 2024

Supplementary figures (pp. 3-16)

Supplementary tables (pp. 17-18)

Supplementary references (p. 19)

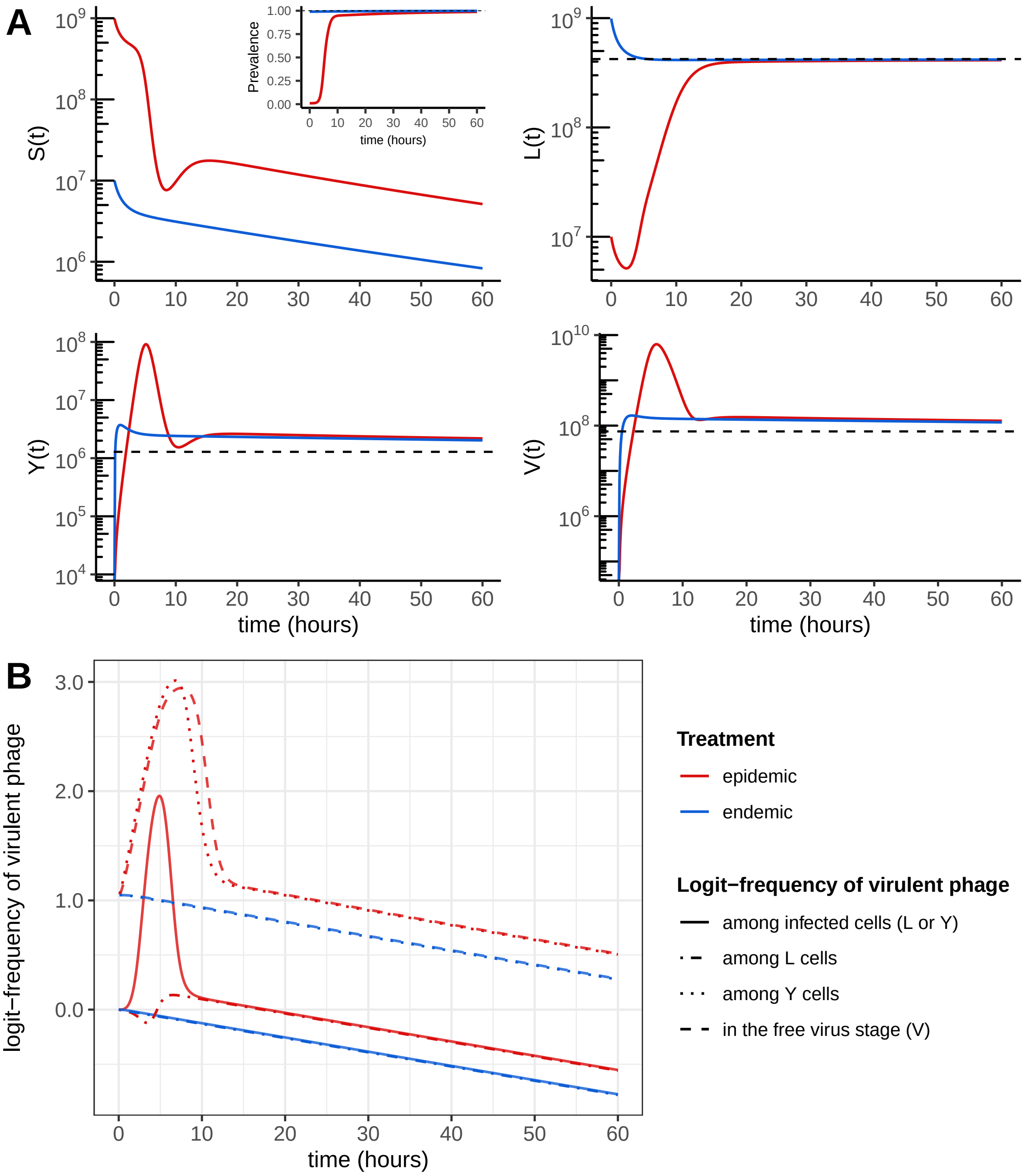


Figure S1: **Numerical simulations for the epidemic and endemic treatments.** (A) Total density of each compartment (horizontal dashed lines indicate endemic equilibrium values) and (B) logit-frequencies of the virulent strain m. See **Table S1** for parameter values. At $t = 0$, bacteria are at carrying capacity K with initial prevalence 1% (epidemic treatment) or 99% (endemic treatment). The initial prophage ratio for the two strains is 1:1.

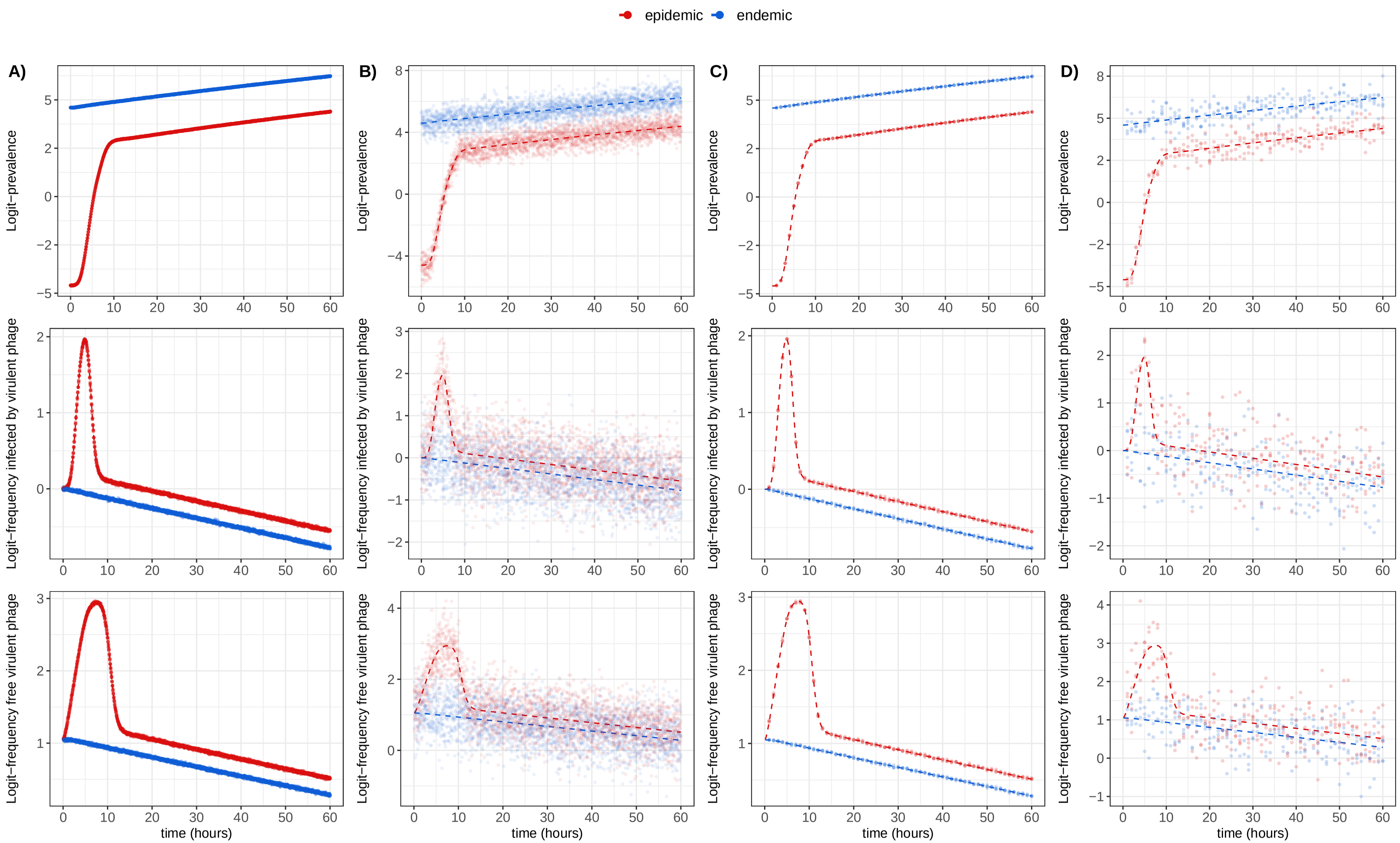


Figure S2: **Simulated datasets differing in quantity and/or quality.** These four sets of simulated data (points) stem from the same two deterministic simulations (epidemic vs. endemic treatment, in dashed lines, see **Table S1** for parameter values and initial conditions) to which we add white Gaussian noise to mimic measurement errors. This was performed four times for each simulation to generate four replicates (chemostats) per treatment. Simulated datasets differ in terms of sampling effort: 1 (C-D) vs. 10 h^{-1} (A-B), and/or accuracy: SD of measurement errors = 0.01 (A-C) vs. 0.5 (B-D).

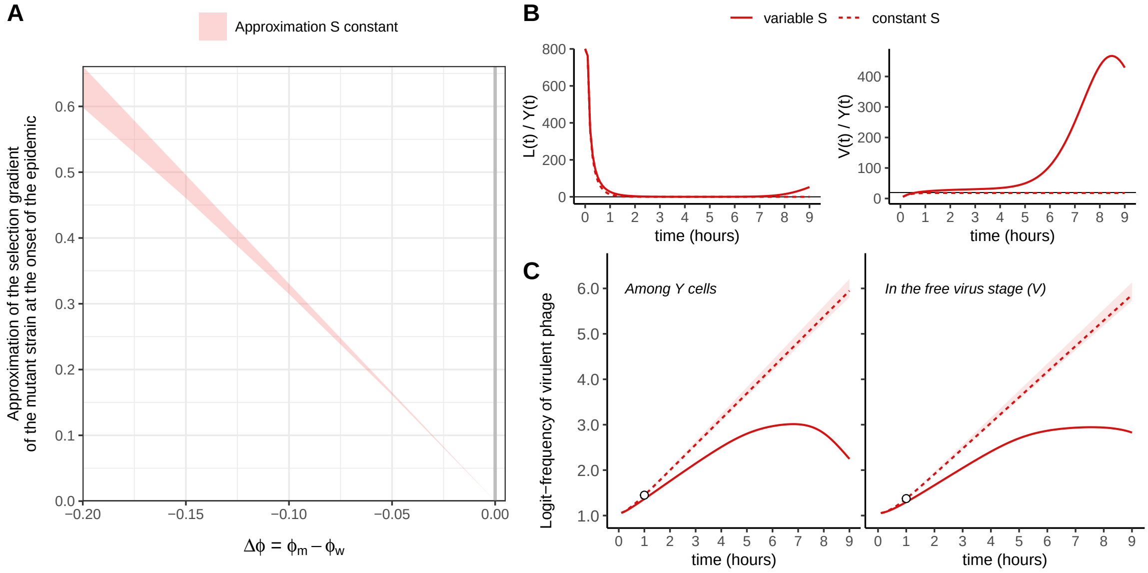


Figure S3: **Approximation of the selection gradient of the virulent phage at the early stage of the epidemic.** At $t = 0$, bacteria are at carrying capacity K with initial prevalence 1% (epidemic treatment) and initial prophage ratio for the two strains 1:1. We investigate the case where the pool of susceptible host is kept constant to its initial value $S(t = 0) = 0.99 \times K$ throughout the course of infection (dashed lines). See **Table S1** for parameter values. In panel A, we vary $\Delta\phi = \phi_m - \phi_w$ by varying ϕ_m; in panels B and C, $\Delta\phi = -0.18$. (A) Approximation interval for the selection gradient of the mutant strain (slope of $\text{logit}(f(t))$) against $\Delta\phi$, as given in **SI Appendix §S2.2** (equation (S17)); note that the selection gradient is positive when $\Delta\phi$ is negative, meaning that the virulent phage is predicted to be selected for at the early stage of the epidemic. (B) Temporal dynamics of the ratios $L(t)/Y(t)$ and $V(t)/Y(t)$ along with their approximations (horizontal black lines). (C) Predictions of the trajectories of the logit-frequencies of the virulent phage among lytic cells (Y) and in the free virus stage (V), starting from values at $t = 1$ (white points). Indeed, at $t = 0$, only lysogens are introduced in the susceptible population, so we first let the phage-bacteria system reaches its new dynamical regime – when epidemiological dynamics have reached (quasi-)equilibrium in panel B. The shaded area derives from the approximation interval illustrated in panel A

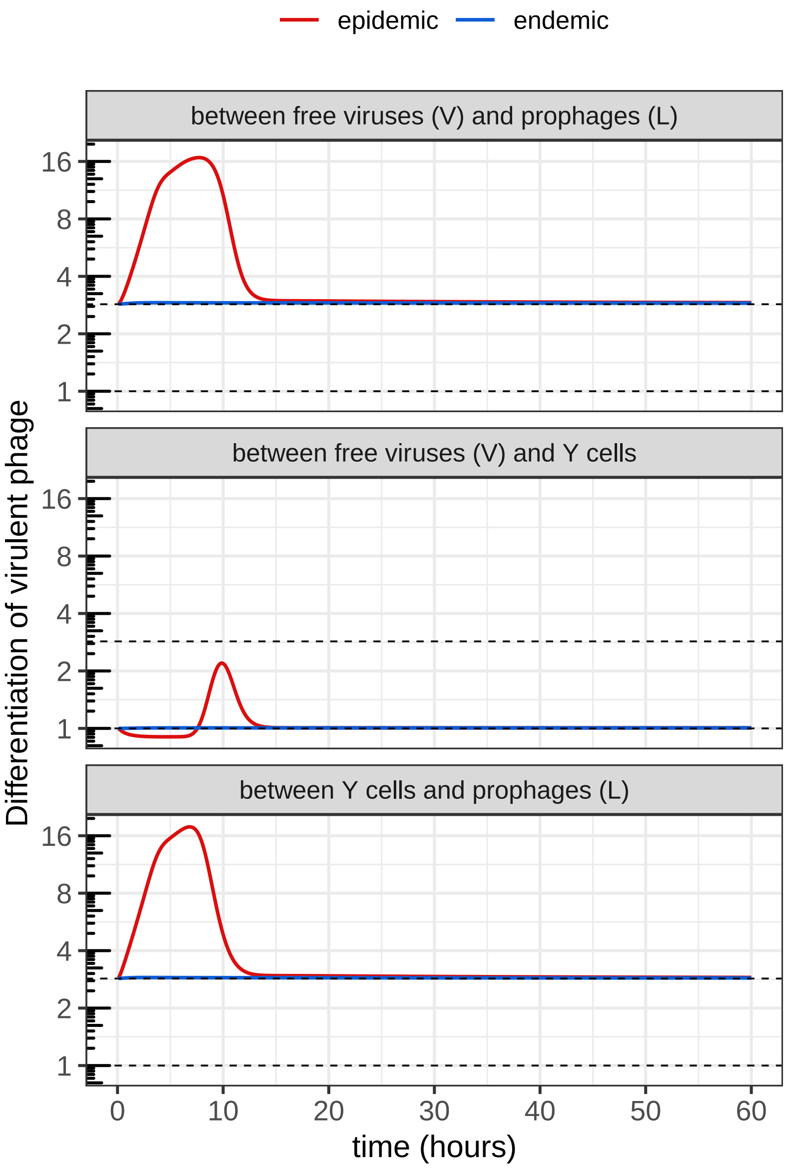


Figure S4: **Differentiation of the virulent phage.** In **SI Appendix §S2.4**, we show that the differentiation of the virulent phage between free viruses (V) or lytic cells (Y) and prophages (L) is predicted to converge towards approximately $1 + \Delta\alpha/\alpha_w > 1$ (uppermost dashed line), and that the differentiation between free viruses (V) and lytic cells (Y) towards approximately 1 (no differentiation, lowermost dashed line). In the endemic case – including the late stage of the epidemic –, the virulent strain is therefore more frequent among free viruses and lytic cells than among prophages but is as frequent among free viruses as among lytic cells. See **Table S1** for parameter values and initial conditions.

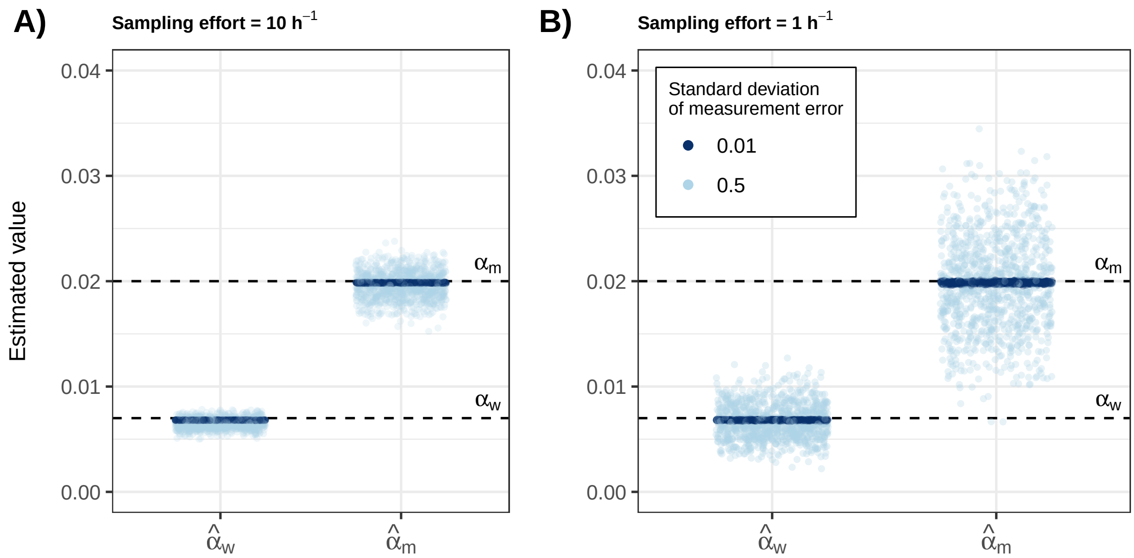


Figure S5: **Point estimates of the reactivation rates α_w and α_m from simulated datasets.** We consider data with different sampling efforts ((A) 10 vs. (B) 1 h^{-1}) and/or SD of measurement errors (0.01 (dark blue) vs. 0.5 (light blue)); for each combination, we computed 1,000 simulated datasets by adding Gaussian noise to the same two simulations (epidemic vs. endemic) shown in **Fig. S1**; repeating this four times, we independently generate four replicates (chemostats) per treatment. For each replicate, we only keep the data from the time point the system had reached a prevalence of 95%. From each simulated dataset, we eventually compute point estimates of both parameters α_w and α_m (see **Materials and methods §2.3.1** and **SI Appendix §S3.1**) which overall show a good match with the values used in the original simulation (horizontal dashed lines).

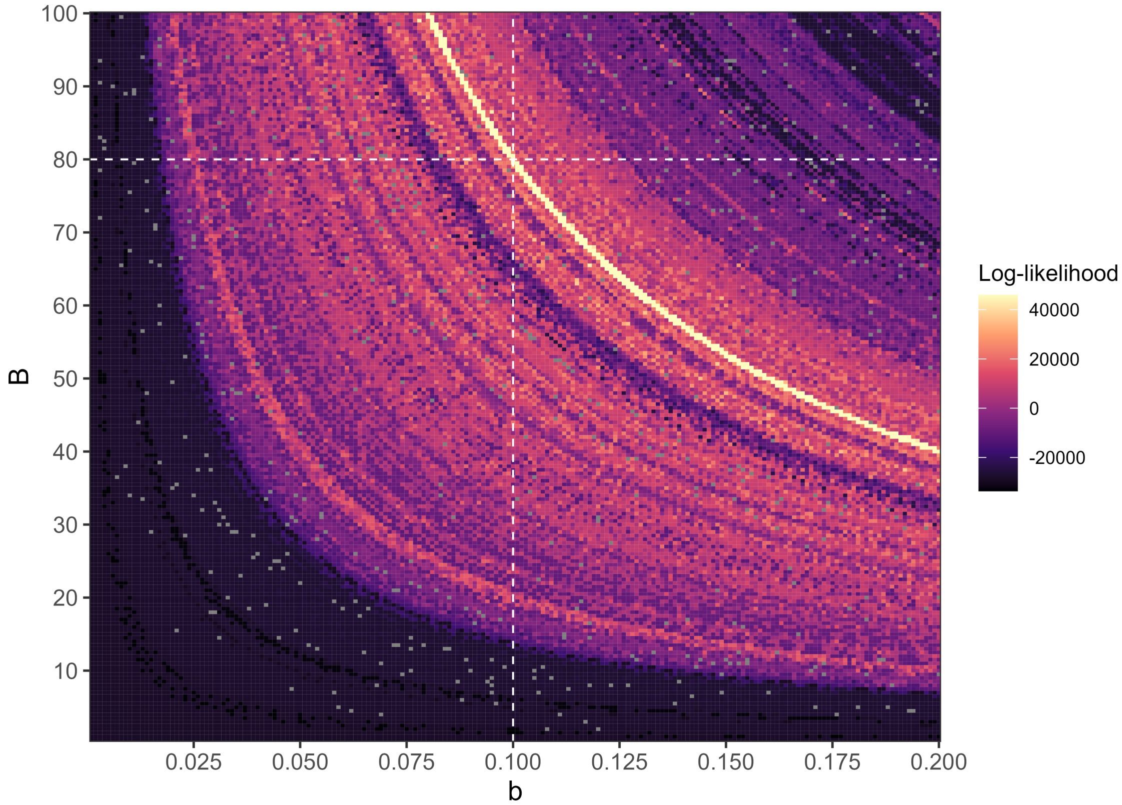


Figure S6: **Log-likelihood landscape according to the values of parameters b and B.** We used the simulated dataset closest to the original deterministic simulation (sampling effort = 10 h^{-1} and SD of measurement error = 0.01, see **Fig. S2-A**). For each pair $(b,\ B) \in [0,\ 0.2] \times [0,\ 100]$, we also fixed the rates of prophage reactivation α_w and α_m (as though correctly estimated beforehand) as well as K and δ while we maximized the overall log-likelihood over the remaining parameters. White dashed lines indicate the values used in the original simulations. Non successful completions are shown in grey. This landscape points out that parameters b and B are not separately identifiable; its shape suggests that only the product $b \times B$ is identifiable, in particular the log-likelihood always reaches the highest value (lightest curve) when $b \times B$ is around 8.

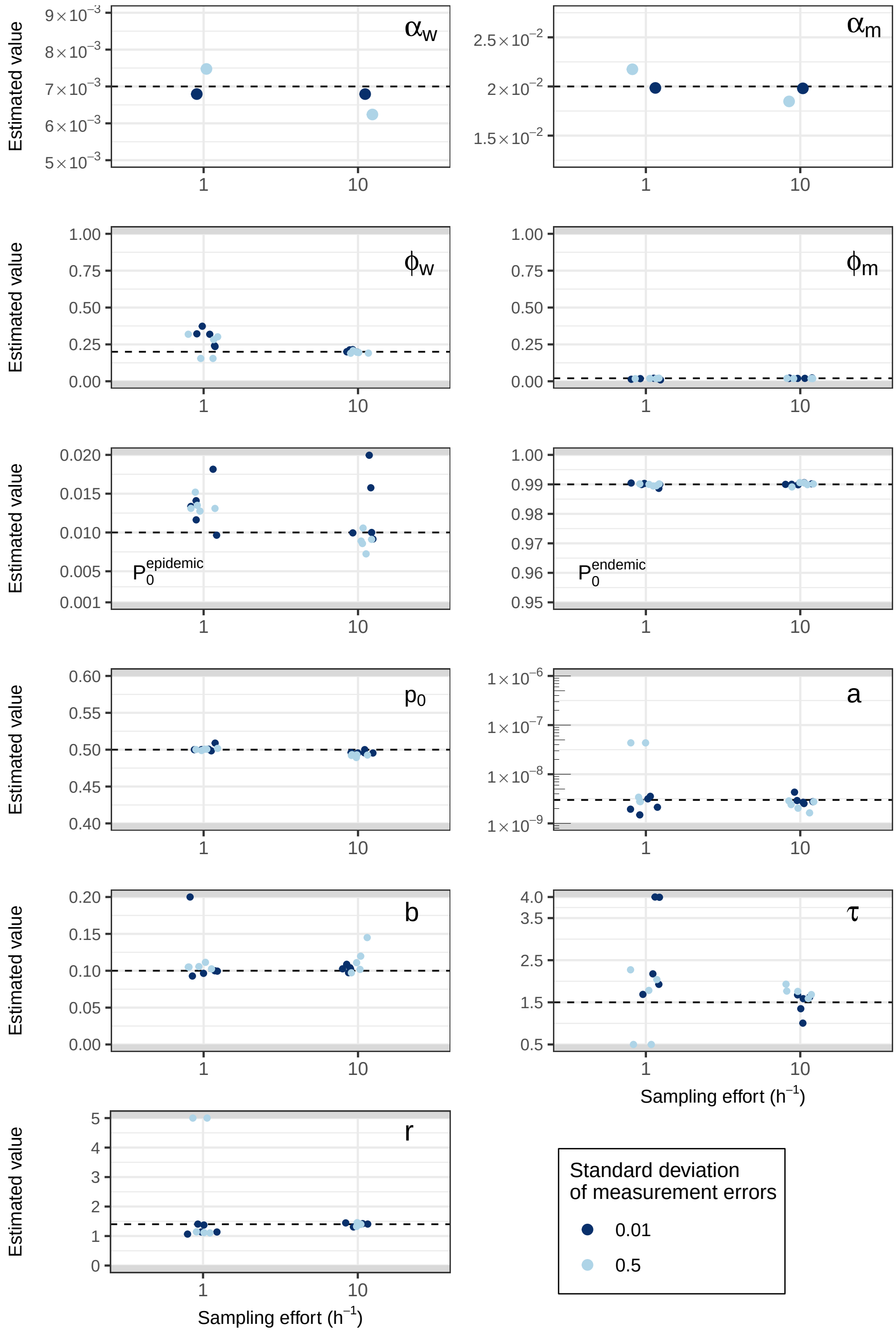


Figure S7: **Parameter point estimates from simulated data.** We use the four simulated datasets described in **Fig. S5**, which all stem from the same deterministic simulation but differ in sampling effort (1 vs. 10 h^{-1}) and/or SD of measurement errors (0.01 vs. 0.5). For each simulated dataset, we first estimate α_w and α_m; we then run non-linear optimizations to estimate the remaining parameters (K, δ and B are fixed), starting from a set of 2,000 initial values and keeping parameter values from the best fit. We repeat this procedure for five sets of initial values, yielding five point estimates for each remaining parameter. Dashed horizontal lines refers to the values we use to generate simulated data and grey areas indicate out-of-bounds ranges of values.

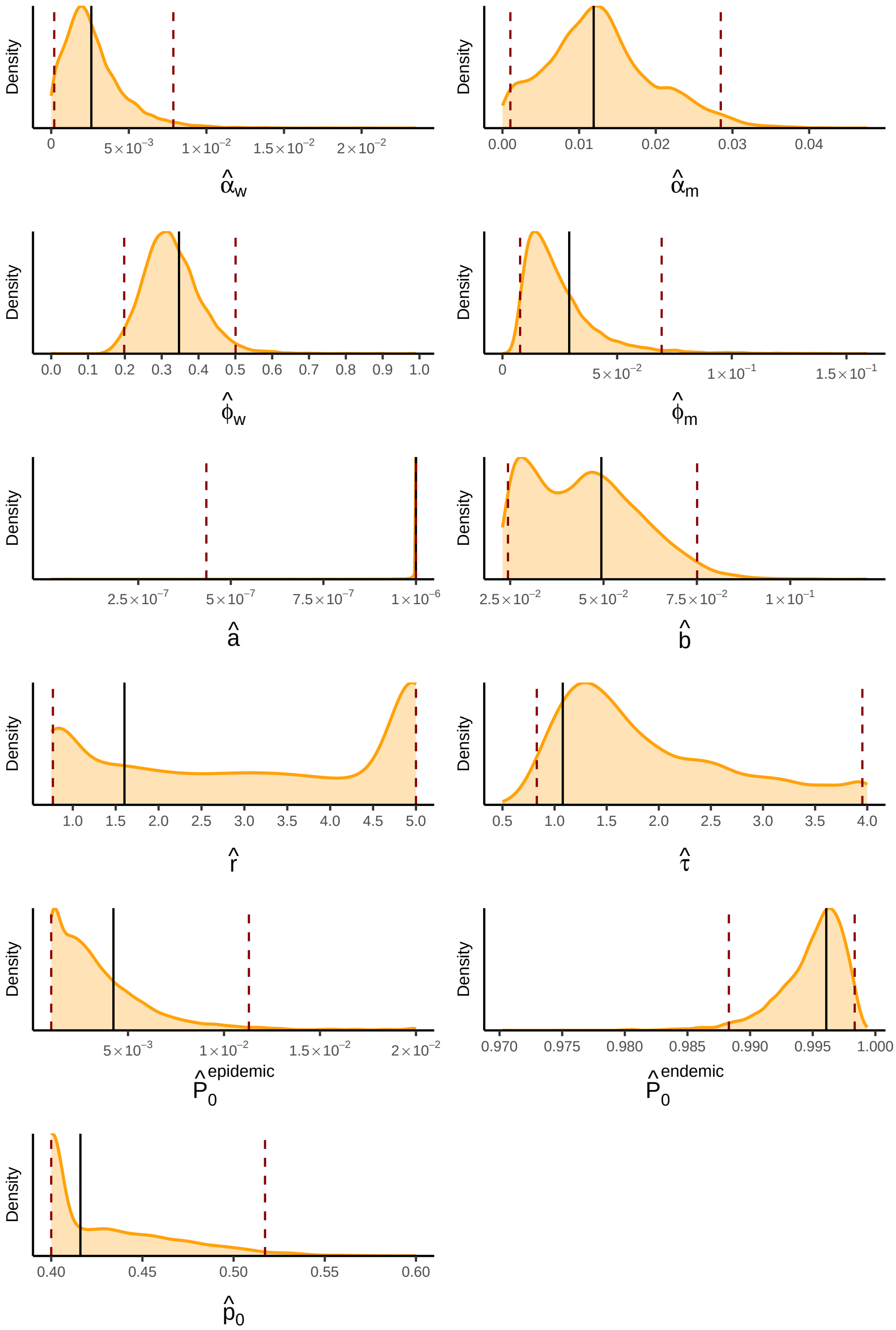


Figure S8: **Density distributions of parameter estimates.** Using sieve bootstrap on the residuals between experimental data and the best fit of our model (black vertical solid lines), we generate 1,000 new datasets on which we reiterate the estimation procedure to compute the joint distributions of estimated parameters; we only keep results with successful convergence ($n = 6,686$). Dashed vertical lines indicates 2.5% and 97.5% quantiles.

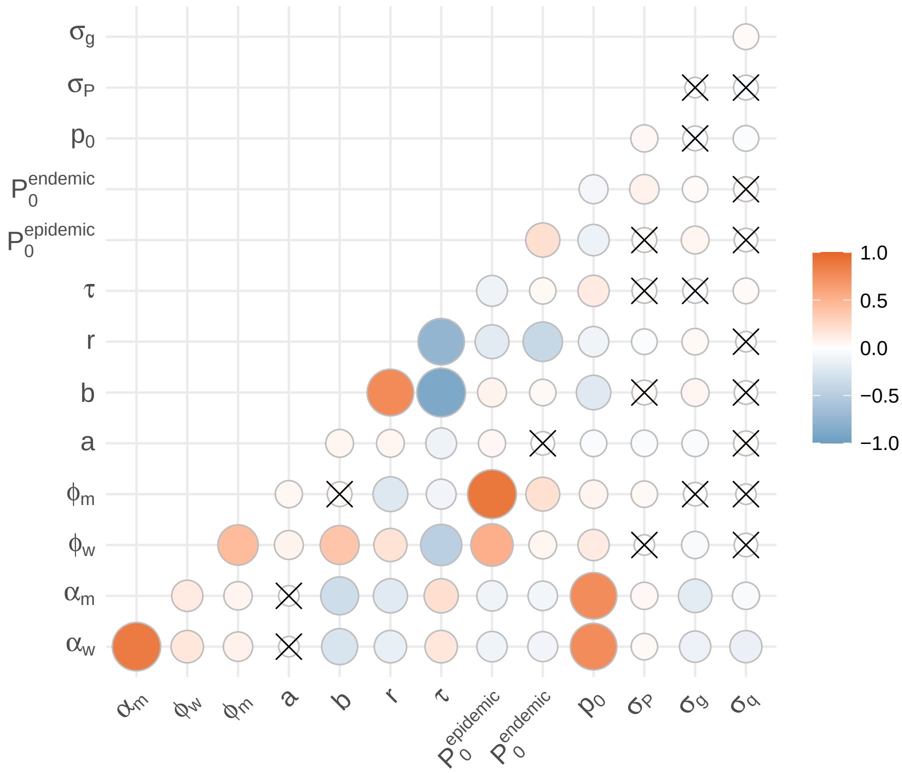


Figure S9: **Pairwise correlations.** We compute Pearson correlation coefficient of bootstrap-based distributions of parameter estimations (cf. **Fig. S8**, $n = 6,686$). Crosses indicate non-significant coefficients (p-value > 0.05).

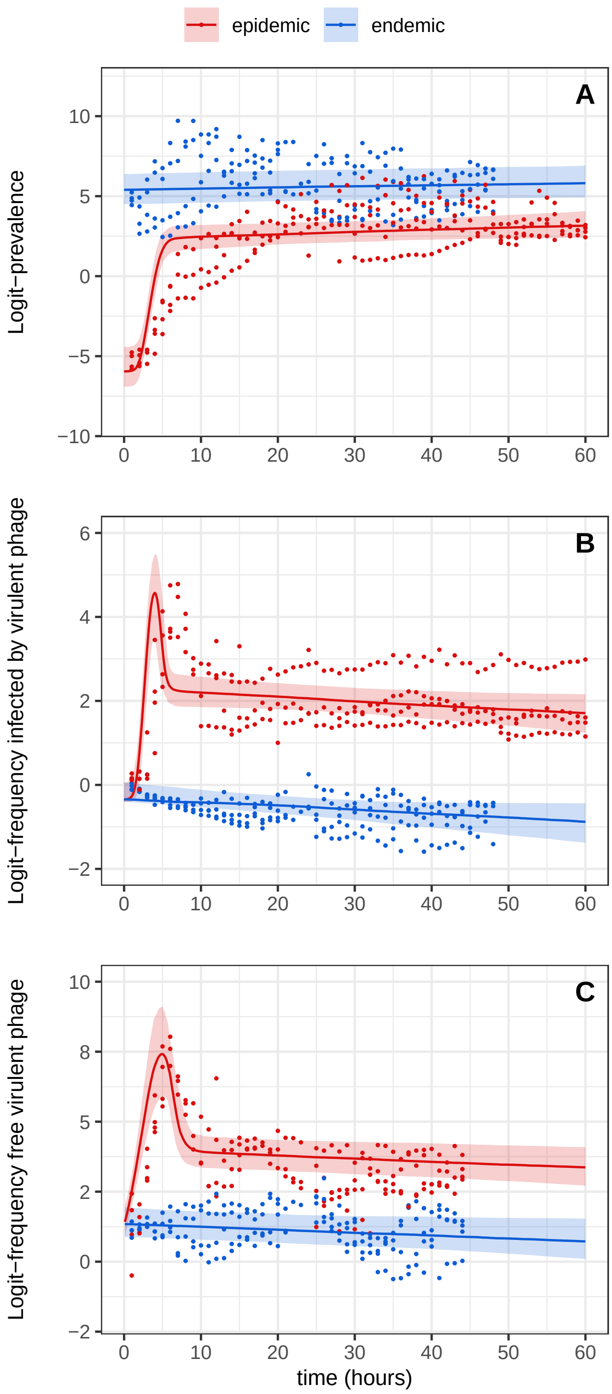


Figure S10: **Fitted values from sieve bootstrap on experimental data.** Distributions of fitted values – median (line) and 95% interval (shaded envelope) – were obtained using sieve bootstrap on the residuals between experimental data (points) and the best fit of our model. Initial prevalence is either low (endemic condition, in blue) or high (epidemic condition, in red). (A) Logit-prevalence of infection; (B) logit-frequency of cells infected (either lysogenic or lytic) by the the mutant (virulent) phage; (C) logit-frequency of the mutant (virulent) in the culture medium (free virus stage).

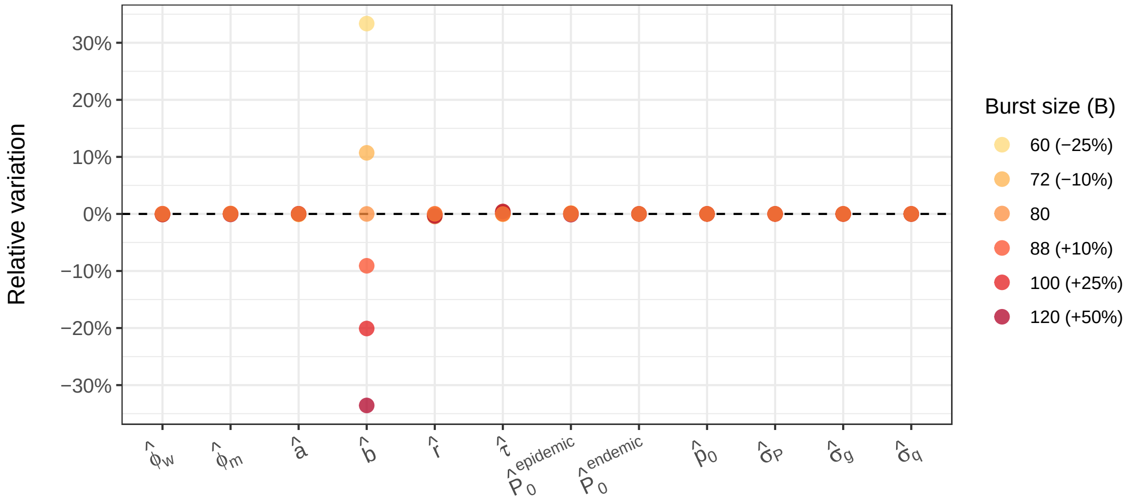


Figure S11: **Sensitivity of the inference of estimated parameters to the fixed burst size.** We apply ±10%, ±25% and +50% perturbations on the fixed burst size B (original fixed value: $B = 80\ virus\,.\,cell^{-1}$) and we reiterate non-linear optimizations as before to compute new point estimates. Relative variations (y-axis) refer to the percentage of variation of these new estimates compared to the original best MLE estimates (without perturbation). All parameters show high robustness to these perturbations, except for parameter b as expected from **Fig. S6** (but the product $b \times B$ is always around 3.95).

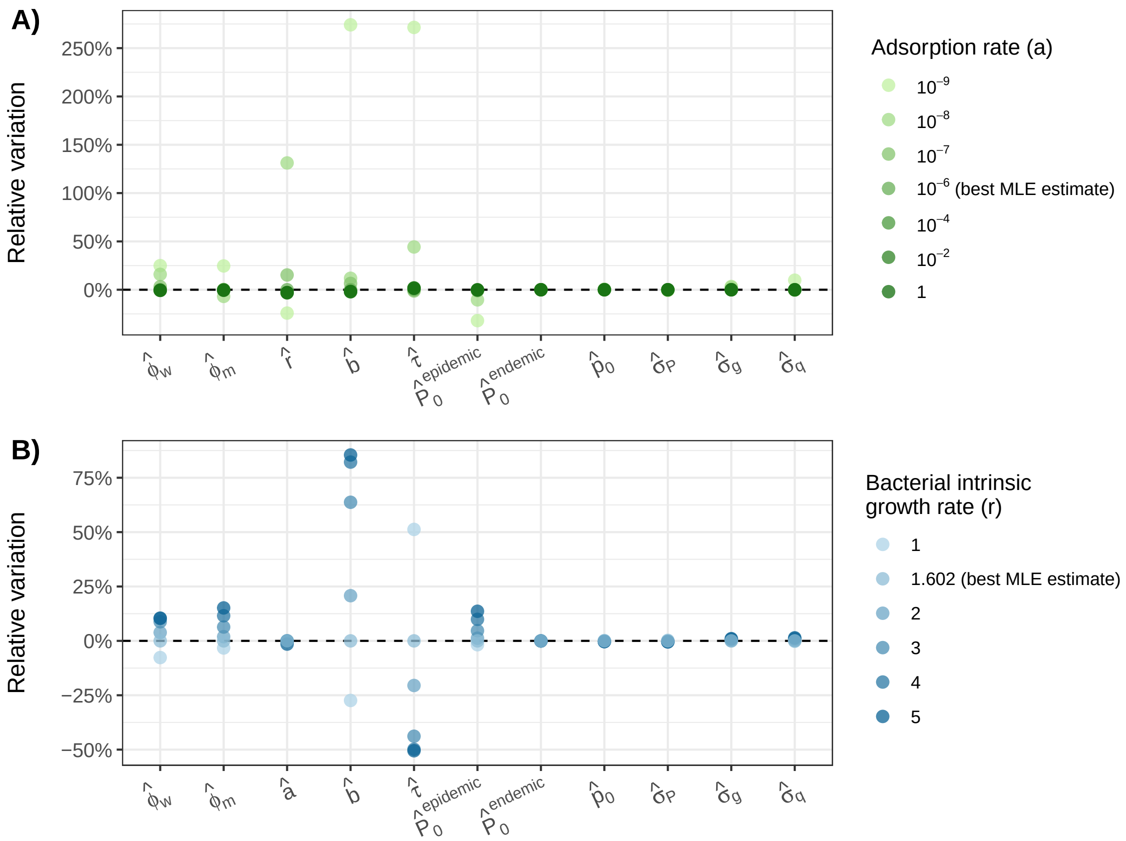


Figure S12: **Sensitivity of the estimated parameters to perturbations in the adsorption rate or in the bacterial intrinsic growth rate.** Both these parameters – a and r, respectively – were poorly estimated with experimental data. We thus investigate *a posteriori* how the other estimated parameters are impacted when a and r deviate from their best MLE estimates (both expressed in h^{-1}). Iteratively fixing a (A) or r (B) to some values, we then reiterate non-linear optimizations as before to compute new point estimates. Relative variations (y-axis) refer to the percentage of variation of these new estimates compared to best MLE estimates.

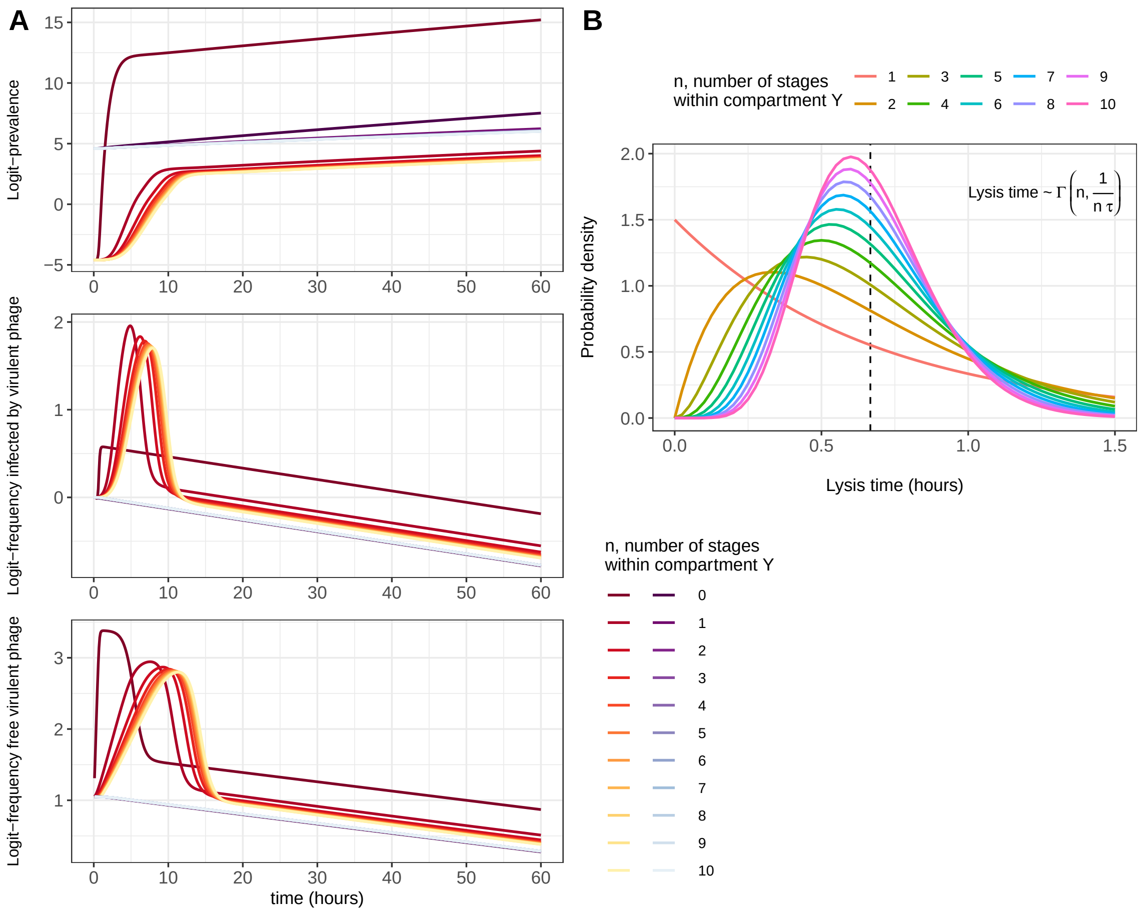


Figure S13: **Comparisons between different lysis time distributions.** Using linear/Gamma chain trick, compartments Y (both Y_w and Y_m) are stratified into n successive stages. As ODEs implicitly assume exponentially distributed sojourn time, the lysis time is thus the sum of n i.i.d. exponential distributions, that is a gamma distribution with shape parameter n and scale parameter $1/(n\tau)$. The case $n = 1$ corresponds to the exponential distribution – which is the one we use in the main text – and we also include the case $n = 0$ (no compartment Y), as in Berngruber et al., 2013. See **Table S1** for parameter values. The mean lysis time is given by $1/\tau$ (vertical dashed line in B). At $t = 0$, bacteria are at carrying capacity K with initial prevalence 1% (epidemic treatment) or 99% (endemic treatment). The initial prophage ratio for the two strains is 1:1.

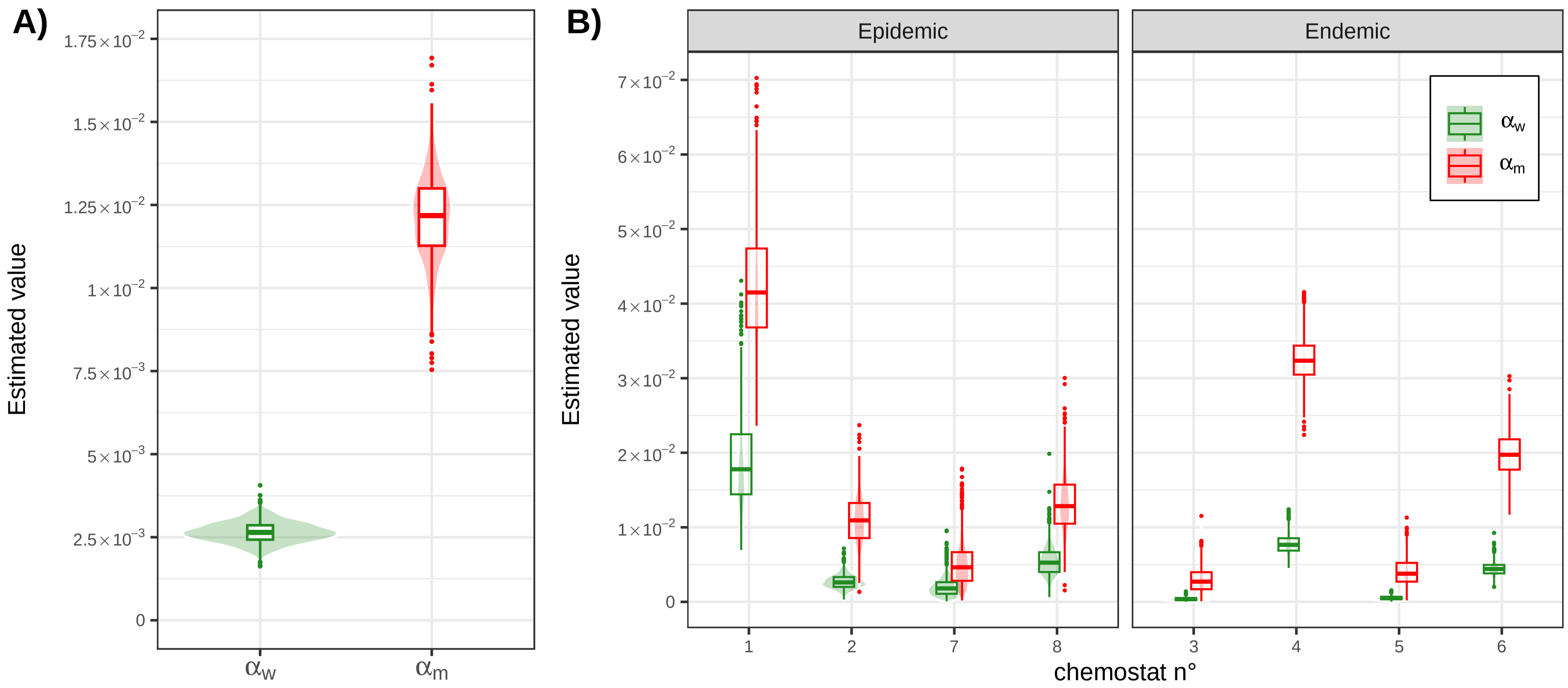


Figure S14: **Bayesian inference of the rates of prophage reactivations.** We estimate α_w and α_m from experimental data using a Bayesian approach (see details in **SI Appendix §3.2**). (A) Parameters α_w and α_m are assumed to be the same for all chemostats – $\hat{\alpha}_w = 2.65 \times 10^{-3}$ (mean, 95% credible interval $[2.03 \times 10^{-3},\ 3.32 \times 10^{-3}]$) and $\hat{\alpha}_m = 1.21 \times 10^{-2}$ (mean, 95% credible interval $[9.48 \times 10^{-3},\ 1.47 \times 10^{-2}]$) –; (B) parameters α_w and α_m are allowed to vary across chemostats.

Table S1: **Parameter values used in the simulations.** The subscripts w and m refer to the wildtype strain and the mutant (or virulent) strain λcI857 of phage λ, respectively. $P_0^{epidemic}$ and $P_0^{endemic}$ correspond to the initial conditions (at $t = 0$) of the prevalence for the endemic and epidemic treatment, respectively; p_0 corresponds to the initial condition of the frequency of lysogenic hosts (L) infected by the virulent phage (infected cells are all lysogenic at $t = 0$). See **Table 1** for notations.

Parameter	**Value**	**Unit**
$P_0^{epidemic}$	1%	–
$P_0^{endemic}$	99%	–
p_0	0.5	–
α_w	7×10^{-3}	h^{-1}
α_m	2×10^{-2}	h^{-1}
ϕ_w	0.2	–
ϕ_m	2×10^{-2}	–
a	3×10^{-9}	$h^{-1} \,.\, cell^{-1}$
b	0.1	–
τ	1.5	h^{-1}
B	80	$virus \,.\, cell^{-1}$
r	1.4	h^{-1}
K	10^9	$cell$
δ	0.8	h^{-1}
$\sigma_P,\ \sigma_g,\ \sigma_q$	0.01/0.5	–

Table S2: **Bounds for non-linear optimizations.** Bounds are placed on estimated parameters to constrain optimizations to relevant ranges of values. Besides, starting values are uniformly drawn between these bounds.

Parameter	**Lower bound**	**Upper bound**	**Unit**
$P_0^{epidemic}$	0.95	1	–
$P_0^{endemic}$	10^{-3}	2×10^{-2}	–
p_0	0.4	0.6	–
$\phi_w,\ \phi_m$	0	1	–
a	10^{-9}	10^{-6}	$h^{-1} \,.\, cell^{-1}$
b	0	0.2	–
r	0	5	h^{-1}
τ	0.5	4	h^{-1}
$\sigma_P,\ \sigma_g,\ \sigma_q$	10^{-3}	2	–

Table S3: **Examples of values for phage parameters from previous studies.** We do not include the probability of fusion b, as it is not separately identifiable from the burst size B in our model and as it is not really considered in most studies. We also do not include the probability of lysogenization of the virulent phage ϕ_m as, to our knowledge, it has not been estimated elsewhere.

Parameter	**Value**	**Unit**	**Reference**
α_w	3.4×10^{-4}	h^{-1}	(De Paepe et al., 2016)
	$6.3 \times 10^{-3} - 3.6 \times 10^{-2}$		(De Paepe et al., 2016)
	2.4×10^{-5}		(Little et al., 1999)
	$\sim 10^{-7} - 10^{-6}$		(Zong et al., 2010)
α_m	1.3×10^{-2} (i)	h^{-1}	(De Paepe et al., 2016)
	$\sim 10^{-3}$		(Zong et al., 2010)
ϕ_w	0.19	–	(De Paepe et al., 2016)
	0.63 (ii)		(Little et al., 1999)
	$0.2 - 0.4$ (iii)		(Zeng et al., 2010)
a	2.7×10^{-8}	$h^{-1} \,.\, cell^{-1}$	(De Paepe & Taddei, 2006)
	$5 \times 10^{-8} - 3 \times 10^{-7}$		(De Paepe et al., 2016)
	$\sim 10^{-9} - 10^{-8}$ (iv)		(Lindberg et al., 2014)
	$\sim 10^{-8} - 10^{-7}$		(Shao & Wang, 2008)
B	115	$virus \,.\, cell^{-1}$	(De Paepe & Taddei, 2006)
	12.1		(De Paepe et al., 2016)
	$37 - 590$ (iv)		(Lindberg et al., 2014)
	56		(Little et al., 1999)
	$9.7 - 255$		(Wang, 2006)
$1/\tau$	0.7	h	(De Paepe & Taddei, 2006)
	> 0.67		(De Paepe et al., 2016)
	$0.78 - 1.67$ (iv)		(Lindberg et al., 2014)
	0.49-1.13		(Shao & Wang, 2008)
	0.47-1.05		(Wang, 2006)

(i) Virulent mutant λcI*; (ii) MOI=6-8; (iii) MOI=1; (iv) Phages of *P. aeruginosa* from environmental water source.

Supplementary references

Berngruber, T. W., Froissart, R., Choisy, M., & Gandon, S. (2013). Evolution of virulence in emerging epidemics. *PLoS pathogens*, *9*(3), e1003209. https://doi.org/10.1371/journal.ppat.1003209

De Paepe, M., & Taddei, F. (2006). Viruses' life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS biology*, *4*(7), e193. https://doi.org/10.1371/journal.pbio.0040193

De Paepe, M., Tournier, L., Moncaut, E., Son, O., Langella, P., & Petit, M.-A. (2016). Carriage of λ latent virus is costly for its bacterial host due to frequent reactivation in monoxenic mouse intestine. *PLoS genetics*, *12*(2), e1005861. https://doi.org/10.1371/journal.pgen.1005861

Lindberg, H. M., McKean, K. A., & Wang, I.-N. (2014). Phage fitness may help predict phage therapy efficacy. *Bacteriophage*, *4*(4), e964081. https://doi.org/10.4161/21597073.2014.964081

Little, J. W., Shepley, D. P., & Wert, D. W. (1999). Robustness of a gene regulatory circuit. *The EMBO journal*, *18*(15), 4299–4307. https://doi.org/10.1093/emboj/18.15.4299

Shao, Y., & Wang, I.-N. (2008). Bacteriophage adsorption rate and optimal lysis time. *Genetics*, *180*(1), 471–482. https://doi.org/10.1534/genetics.108.090100

Wang, I.-N. (2006). Lysis timing and bacteriophage fitness. *Genetics*, *172*(1), 17–26. https://doi.org/10.1534/genetics.105.045922

Zeng, L., Skinner, S. O., Zong, C., Sippy, J., Feiss, M., & Golding, I. (2010). Decision making at a subcellular level determines the outcome of bacteriophage infection. *Cell*, *141*(4), 682–691. https://doi.org/10.1016/j.cell.2010.03.034

Zong, C., So, L.-h., Sepúlveda, L. A., Skinner, S. O., & Golding, I. (2010). Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. *Molecular systems biology*, *6*(1), 440. https://doi.org/10.1038/msb.2010.96

— Supplementary Information (SI Appendix) —

Evolution of virulence in emerging epidemics: from theory to experimental evolution and back

Wakinyan Benhamou, François Blanquart, Marc Choisy, Thomas W. Berngruber, Rémi Choquet and Sylvain Gandon

May 23, 2024

S1 A model coupling epidemiology and evolution

In the main text, we model in continuous cultures of *E. coli* the competition between two strains of phage λ: the wildtype strain – hereafter denoted w – vs. the mutant strain (or variant) λcI857 – hereafter denoted m. The variant m is known to be more virulent than the wildtype strain w due to a point mutation in the transcriptional repressor protein cI (St-Pierre & Endy, 2008; Sussman & Jacob, 1962).

S1.1 Epidemiology

We recall here the epidemiological model (system of ODEs) we use in the main text (see **Table 1** for notations):

$$\begin{cases} \dot{S}(t) &= rS(t)\left(1-\dfrac{N(t)}{K}\right) - \Big(abV(t)+\delta\Big)S(t) \\ \dot{L}(t) &= rL(t)\left(1-\dfrac{N(t)}{K}\right) + \overline{\phi}(t)abV(t)S(t) - \Big(\overline{\alpha}(t)+\delta\Big)L(t) \\ \dot{Y}(t) &= \big(1-\overline{\phi}(t)\big)abV(t)S(t) + \overline{\alpha}(t)L(t) - \Big(\tau+\delta\Big)Y(t) \\ \dot{V}(t) &= \tau Y(t)B - \Big(aN(t)+\delta\Big)V(t) \end{cases} \tag{S1}$$

where the overlines refer to mean values of the life-history traits after averaging over the distribution of strain frequencies:

$$\begin{cases} \overline{\alpha}(t) &= p(t)\alpha_m + \big(1-p(t)\big)\alpha_w \\ \overline{\phi}(t) &= q(t)\phi_m + \big(1-q(t)\big)\phi_w \end{cases} \tag{S2}$$

Furthermore, at each time point t, the prevalence is given by:

$$P(t) \;=\; \frac{Y(t)+L(t)}{N(t)} \;=\; 1-\frac{S(t)}{N(t)}, \tag{S3}$$

and its differentiation with respect to time yields:

$$\dot{P}(t) = \big(1-P(t)\big)\left[abV(t) - \left(r\left(1-\frac{N(t)}{K}\right)+\tau\right)\frac{Y(t)}{N(t)}\right]. \tag{S4}$$

S1.2 Evolution

We refer in the main text to the different frequencies associated with the mutant strain as follows:

- $p(t) = L_m(t)/L(t)$, the frequency of L cells infected by the mutant strain;
- $q(t) = V_m(t)/V(t)$, the frequency of the mutant strain in the free virus stage (V);
- $f(t) = Y_m(t)/Y(t)$, the frequency of Y cells infected by the mutant strain;
- $g(t) = \dfrac{Y_m(t) + L_m(t)}{Y(t) + L(t)}$, the frequency of cells infected (either Y or L) by the mutant strain.

Note that we also have:

$$g(t) = f(t)\left(\frac{Y(t)}{Y(t) + L(t)}\right) + p(t)\left(\frac{L(t)}{Y(t) + L(t)}\right). \tag{S5}$$

Using our model (S1), we can then easily calculate the temporal dynamics of each of these frequencies, which yields:

$$\begin{cases} \dot{p}(t) = & \Big(\big(q(t) - p(t)\big)\phi_w + q(t)\big(1 - p(t)\big)\Delta\phi\Big)\dfrac{abV(t)S(t)}{L(t)} - p(t)\big(1 - p(t)\big)\Delta\alpha \\ \dot{q}(t) = & \Big(f(t) - q(t)\Big)\dfrac{Y(t)}{V(t)}\tau B \\ \dot{f}(t) = & \Big(\big(q(t) - f(t)\big)\big(1 - \phi_w\big) - q(t)\big(1 - f(t)\big)\Delta\phi\Big)\dfrac{abV(t)S(t)}{Y(t)} + \\ & \Big(\big(p(t) - f(t)\big)\alpha_w + p(t)\big(1 - f(t)\big)\Delta\alpha\Big)\dfrac{L(t)}{Y(t)} \\ \dot{g}(t) = & \Big(p(t) - g(t)\Big)\left(r\left(1 - \dfrac{N(t)}{K}\right) - \delta\right)\dfrac{L(t)}{Y(t) + L(t)} + \Big(q(t) - g(t)\Big)\dfrac{abV(t)S(t)}{Y(t) + L(t)} - \\ & \Big(f(t) - g(t)\Big)\big(\delta + \tau\big)\dfrac{Y(t)}{Y(t) + L(t)} \end{cases} \tag{S6}$$

with $\Delta\alpha = \alpha_m - \alpha_w$ and $\Delta\phi = \phi_m - \phi_w$.

Taken together, equations (S1)-(S6) yield the coupled evolutionary-epidemiological dynamics of this phage-bacteria system. Focusing instead on logit-frequencies, that is the log odds ln(frequency of the variant / frequency of the wildtype):

$$
\begin{cases}
\dfrac{\mathrm{d}\,\mathrm{logit}(p(t))}{\mathrm{d}t} & = & \left(\dfrac{q(t)-p(t)}{p(t)\big(1-p(t)\big)}\phi_w + \dfrac{q(t)}{p(t)}\Delta\phi\right)\dfrac{abV(t)S(t)}{L(t)} - \Delta\alpha \\
\dfrac{\mathrm{d}\,\mathrm{logit}(q(t))}{\mathrm{d}t} & = & \dfrac{f(t)-q(t)}{q(t)\big(1-q(t)\big)}\dfrac{Y(t)}{V(t)}\tau B \\
\dfrac{\mathrm{d}\,\mathrm{logit}(f(t))}{\mathrm{d}t} & = & \left(\dfrac{q(t)-f(t)}{f(t)\big(1-f(t)\big)}(1-\phi_w) - \dfrac{q(t)}{f(t)}\Delta\phi\right)\dfrac{abV(t)S(t)}{Y(t)} + \\
& & \left(\dfrac{p(t)-f(t)}{f(t)\big(1-f(t)\big)}\alpha_w + \dfrac{p(t)}{f(t)}\Delta\alpha\right)\dfrac{L(t)}{Y(t)} \\
\dfrac{\mathrm{d}\,\mathrm{logit}\,(g(t))}{\mathrm{d}t} & = & \dfrac{p(t)-g(t)}{g(t)\big(1-g(t)\big)}\left(r\left(1-\dfrac{N(t)}{K}\right)-\delta\right)\dfrac{L(t)}{Y(t)+L(t)} + \dfrac{q(t)-g(t)}{g(t)\big(1-g(t)\big)}\dfrac{abV(t)S(t)}{Y(t)+L(t)} - \\
& & \dfrac{f(t)-g(t)}{g(t)\big(1-g(t)\big)}\Big(\delta+\tau\Big)\dfrac{Y(t)}{Y(t)+L(t)}
\end{cases}
\tag{S7}
$$

S2 Theoretical analyses

S2.1 Equilibrium states & basic reproduction number

In the absence of the virus, (S1) converges trivially to the following equilibrium state: $\big(S(\infty),\ L(\infty),\ Y(\infty),\ V(\infty)\big) = \big(K(1-\delta/r),\ 0,\ 0,\ 0\big)$ if $r > \delta$, $\big(0,\ 0,\ 0,\ 0\big)$ otherwise. When a single strain of the virus (with phenotypes α and ϕ) is introduced in the bacterial population (fully susceptible, with density S_0), the fate of this phage-bacteria system depends on the basic reproduction number $\mathscr{R}_0$ of the pathogen – i.e., the expected number of secondary infections caused by one primary infected bacteria in an otherwise fully susceptible population. Using the next-generation-matrix method (Diekmann et al., 2010), we decompose the life cycle of phage λ into transmission (matrix $\mathbf{T}$) and transition (matrix $\mathbf{\Sigma}$) components:

$$
\mathbf{T} = \begin{pmatrix} r\left(1-\frac{S_0}{K}\right) & 0 & \phi abS_0 \\ 0 & 0 & (1-\phi)abS_0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{pmatrix} -(\alpha+\delta) & 0 & 0 \\ \alpha & -(\tau+\delta) & 0 \\ 0 & \tau B & -(aS_0+\delta) \end{pmatrix},
$$

such that :

$$
\Big(\dot{L}(t) \quad \dot{Y}(t) \quad \dot{V}(t)\Big)^\top = (\mathbf{T}+\mathbf{\Sigma})\Big(L(t) \quad Y(t) \quad V(t)\Big)^\top
$$

The basic reproduction number $\mathscr{R}_0$ is given by the spectral radius of the matrix $-\mathbf{T}\mathbf{\Sigma}^{-1}$, that is:

$$\mathscr{R}_0 = \frac{A + \sqrt{A^2 - 4r\left(1 - \frac{S_0}{K}\right)(1-\phi)abS_0\tau B(aS_0+\delta)(\alpha+\delta)(\tau+\delta)}}{2(aS_0+\delta)(\alpha+\delta)(\tau+\delta)}, \tag{S8}$$

with $A = r\left(1 - \frac{S_0}{K}\right)(aS_0+\delta)(\tau+\delta) + abS_0\tau B(\alpha + (1-\phi)\delta)$. Note that, if $S_0 = K$, equation (S8) reduces to:

$$\mathscr{R}_0 = \frac{abK\tau B(\alpha + (1-\phi)\delta)}{(aK+\delta)(\alpha+\delta)(\tau+\delta)},$$

When $\mathscr{R}_0 < 1$, phages goes extinct and the bacterial population converges to the virus-free equilibrium above. Alternatively, when $\mathscr{R}_0 > 1$, an epidemic breaks out and eventually stabilises to the following endemic equilibrium:

$$\begin{cases} S(\infty) &= 0 \\ L(\infty) &= \dfrac{K\Big(r - (\delta+\alpha)\Big)}{\delta+\alpha+\tau}\ \dfrac{\delta+\tau}{r} \\ Y(\infty) &= \dfrac{K\Big(r - (\delta+\alpha)\Big)}{\delta+\alpha+\tau}\ \dfrac{\alpha}{r} \\ V(\infty) &= \dfrac{K\Big(r - (\delta+\alpha)\Big)}{\delta+\alpha+\tau}\ \dfrac{B\ \alpha\ \tau}{aK\Big(r-(\delta+\alpha)\Big) + \delta\ r} \end{cases} \tag{S9}$$

S2.2 Viral dynamics in an emerging epidemic

In the epidemic case, susceptible hosts are initially highly abundant. To simplify, we investigate the case where the density of susceptible hosts remains constant over time, i.e., $\forall t,\ S(t) = S_0$. Starting with a bacterial population at carrying capacity K, we also assume that the population size remains constant over time, i.e., $N(t) = K$, leading to the following simplified epidemiological system:

$$\begin{cases} \dot{S}(t) &= 0 \\ \dot{L}(t) &= \overline{\phi}(t)abV(t)S_0 - \Big(\overline{\alpha}(t) + \delta\Big)L(t) \\ \dot{Y}(t) &= \big(1 - \overline{\phi}(t)\big)abV(t)S_0 + \overline{\alpha}(t)L(t) - \Big(\tau + \delta\Big)Y(t) \\ \dot{V}(t) &= \tau Y(t)B - \Big(aK + \delta\Big)V(t) \end{cases} \tag{S10}$$

To study the selection gradient $\mathscr{S}$ of the virulent phage – i.e., the rate at which it grows or declines in frequency on the logit scale –, we focus on compartment Y. As shown in (S7), the temporal dynamics of $\text{logit}\big(f(t)\big)$ depends on the ratio $V(t)/Y(t)$, whose differentiation with respect to time is given by the quadratic polynomial:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{V(t)}{Y(t)}\right) = -\big(1-\overline{\phi}(t)\big)abS_0\left(\frac{V(t)}{Y(t)}\right)^2 - \left(\overline{\alpha}(t)\frac{L(t)}{Y(t)} + aK - \tau\right)\frac{V(t)}{Y(t)} + \tau B. \tag{S11}$$

We now use an argument of separation of time scale (Rinaldi & Scheffer, 2000; Verhulst, 2007) under the assumption of weak selection: epidemiological dynamics, such as $L(t)$, $Y(t)$ and $V(t)$, are treated as fast variables while evolutionary dynamics – i.e., strain frequencies – are treated as slow variables because phenotypic differences between the wildtype and the virulent strain are assumed to be small. Setting the right-hand side of (S11) to 0, we obtain the positive solution:

$$\frac{V(t)}{Y(t)} \approx \frac{\tau - \overline{\alpha}(t)\frac{L(t)}{Y(t)} - aK + \sqrt{\left(\tau - \overline{\alpha}(t)\frac{L(t)}{Y(t)} - aK\right)^2 + 4\big(1-\overline{\phi}(t)\big)abS_0\tau B}}{2\big(1-\overline{\phi}(t)\big)abS_0}.$$

In the early state of the epidemic, the system is mainly governed by the lytic pathway, so we assume that $L(t)/Y(t) \approx 0$ (**Fig. S3-B**). Furthermore, under the assumption of weak selection, we substitute $\overline{\phi}(t)$ by ϕ_w. We denote Z this approximation of the ratio $V(t)/Y(t)$:

$$Z = \frac{\tau - aK + \sqrt{(\tau - aK)^2 + 4(1-\phi_w)abS_0\tau B}}{2(1-\phi_w)abS_0}. \tag{S12}$$

The temporal dynamics of $\mathrm{logit}\big(f(t)\big)$ becomes:

$$\frac{\mathrm{d}\,\mathrm{logit}\big(f(t)\big)}{\mathrm{d}t} \approx \underbrace{\frac{\tau - aK + \sqrt{(\tau - aK)^2 + 4(1-\phi_w)abS_0\tau B}}{2(1-\phi_w)}}_{abS_0 Z}\left(\frac{q(t)-f(t)}{f(t)\big(1-f(t)\big)}(1-\phi_w) - \frac{q(t)}{f(t)}\Delta\phi\right). \tag{S13}$$

We now look at the term in brackets. Posing $D(t) = \frac{q(t)-f(t)}{f(t)\left(1-f(t)\right)}$, we obtain $\frac{q(t)}{f(t)} = 1 + D(t)\big(1-f(t)\big)$ and:

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = \frac{\mathrm{d}\,\mathrm{logit}\big(q(t)\big)}{\mathrm{d}t}\frac{q(t)\big(1-q(t)\big)}{f(t)\big(1-f(t)\big)} - \frac{\mathrm{d}\,\mathrm{logit}\big(f(t)\big)}{\mathrm{d}t}\Big(1 + D(t)\big(1-2f(t)\big)\Big) \tag{S14}$$

We set the right-hand side of (S14) to 0 (quasi-equilibrium) and only keep the solution for which the term $\mathcal{O}(0)$ is equal to 0. Again, under the assumption of weak selection, phenotypic differences are small and $\mathcal{O}(\varepsilon)$. A Taylor expansion for the selected solution about the neutral case - i.e., when both strains share the same phenotype – to first order then yields:

$$D(t) \approx \frac{abS_0\left(\frac{V(t)}{Y(t)}\right)^2 \Delta\phi}{\tau B + abS_0\left(\frac{V(t)}{Y(t)}\right)^2 \big(1-\phi_w\big)} + \mathcal{O}(\varepsilon^2). \tag{S15}$$

Plugging approximations (S12) and (S15) into (S13), we obtain after some rearrangements:

$$\frac{\mathrm{d}\,\mathrm{logit}(f(t))}{\mathrm{d}t} \approx -\Delta\phi\ abS_0 Z\left(\frac{\tau B + abS_0 Z^2\big(1-f(t)\big)\Delta\phi}{\tau B + abS_0 Z^2\big(1-\phi_w\big)}\right). \tag{S16}$$

We use (S16) as an approximation of the selection gradient of the virulent phage $\mathscr{S}$. As $f(t)$ is here always between 0.5 and 1 at the beginning of the epidemic:

$$-\Delta\phi\ abS_0 Z\left(\frac{\tau B + abS_0 Z^2\Delta\phi/2}{\tau B + abS_0 Z^2\big(1-\phi_w\big)}\right) \leq \underbrace{\frac{\mathrm{d}\,\mathrm{logit}(f(t))}{\mathrm{d}t}}_{\mathscr{S}} \leq -\Delta\phi\ \frac{abS_0 Z\tau B}{\tau B + abS_0 Z^2\big(1-\phi_w\big)}, \tag{S17}$$

which may provide good approximations to predict the trajectory of the logit-frequency of the virulent phage in both compartment Y and V, while $S(t)$ does not vary over time (**Fig. S3-C**).

Or, using once again an assumption of weak selection, a Taylor expansion of (S16) to first order in $\Delta\phi$ leads to:

$$\underbrace{\frac{\mathrm{d}\,\mathrm{logit}(f(t))}{\mathrm{d}t}}_{\mathscr{S}} \approx -\Delta\phi\ \frac{abS_0 Z\tau B}{\tau B + abS_0 Z^2\big(1-\phi_w\big)} + \mathcal{O}(\varepsilon^2), \tag{S18}$$

which is the approximation we use in the main text when we propose that, at the beginning of the epidemic:

$$\mathscr{S} \propto -\Delta\phi. \tag{S19}$$

Therefore, the virulent phage ($\Delta\phi < 0$) is selected for at the beginning of the epidemic.

S2.3 Viral dynamics at the end of the epidemic

At the end of the epidemic, the prevalence is high and the pool of susceptible host is depleted, so that $P(t) \approx 1$ and $S(t) \approx 0$; the system (S1) reduces then to:

$$
\begin{cases}
\dot{S}(t) & = 0 \\
\dot{L}(t) & = rL(t)\left(1 - \dfrac{N(t)}{K}\right) - \Big(\overline{\alpha}(t) + \delta\Big)L(t) \\
\dot{Y}(t) & = \overline{\alpha}(t)L(t) - \Big(\tau + \delta\Big)Y(t) \\
\dot{V}(t) & = \tau Y(t)B - \Big(aN(t) + \delta\Big)V(t)
\end{cases}
\tag{S20}
$$

This phage-bacteria system is now driven by the lysogenic pathway so that the density of Y cells becomes negligible compared to the density of L cells. Indeed, using (S9):

$$
\lim_{S(t)\,\to\,0} \frac{Y(t)}{L(t)} = \frac{\alpha_w}{\delta + \tau} \approx 0,
$$

as $\alpha_w \ll \tau$ (time elapsed between phage integration and reactivation being much longer than lysis). According to (S5), this also means that the frequency of the virulent strain among infected hosts is now almost entirely driven by its frequency in lysogenic cells solely, that is $g(t) \approx p(t)$.

Rewriting the system (S20) in matrix form for the strain k ($k \in \{w,\ m\}$):

$$
\begin{pmatrix} \dot{L}_k(t) \\ \dot{Y}_k(t) \\ \dot{V}_k(t) \end{pmatrix} = \underbrace{\begin{pmatrix} r\left(1 - \frac{N(t)}{K}\right) - (\delta + \alpha_k) & 0 & 0 \\ \alpha_k & -(\tau + \delta) & 0 \\ 0 & \tau B & -(aN(t) + \delta) \end{pmatrix}}_{\mathbf{R}_k} \begin{pmatrix} L_k(t) \\ Y_k(t) \\ V_k(t) \end{pmatrix},
\tag{S21}
$$

the dominant eigenvalue of the Jacobian $\mathbf{R}_k$ is equals to $r\left(1 - \frac{N(t)}{K}\right) - (\delta + \alpha_k)$. In these conditions, one would expect the selection gradient $\mathscr{S}$ of the mutant strain to be given in each compartment by the difference in eigenvalues which yields:

$$
\begin{aligned}
\mathscr{S} &= \left(r\left(1 - \frac{N(t)}{K}\right) - (\delta + \alpha_m)\right) - \left(r\left(1 - \frac{N(t)}{K}\right) - (\delta + \alpha_w)\right) \\
&= -\alpha_m + \alpha_w \;=\; -\Delta\alpha.
\end{aligned}
\tag{S22}
$$

As a result, the virulent phage is counter-selected in the long-term ($\Delta\alpha > 0 \;\Leftrightarrow\; \mathscr{S} < 0$) and, in each compartment, decreases in frequency at a rate of $\mathscr{S}$ on the logit scale (**Fig. S1-B**). We can once again predict the future trajectory of the logit-frequency of the virulent that linearly declines with negative slope $\mathscr{S} = -\Delta\alpha$. Note that recovering this result for compartment L is straightforward as one just needs to set $S(t)$ to 0 in $\mathrm{d}\,\mathrm{logit}(p(t))/\mathrm{d}t$ in (S7).

S2.4 Differentiation across compartments

We define the differentiation of the virulent strain between free phages (V) and prophages (L) as:

$$\mathcal{Q}^{VL}(t) = \frac{q(t)}{1-q(t)}\frac{1-p(t)}{p(t)}, \tag{S23}$$

such that: $\ln\big(\mathcal{Q}^{VL}(t)\big) = \mathrm{logit}(q(t)) - \mathrm{logit}(p(t))$. Note that we also have $\mathcal{Q}^{VL}(t) = \frac{V_m(t)}{V_w(t)}\frac{L_w(t)}{L_m(t)}$ and that 1 corresponds to no differentiation. Using the eigenvectors associated with the dominant eigenvalues of the Jacobian matrices $\mathbf{R}_w$ and $\mathbf{R}_m$ (cf. equation (S21)), the differentiation $\mathcal{Q}^{YL}(t)$ converges towards:

$$\mathcal{Q}^{VL} = \frac{\alpha_m\big(r(1-N(t)/K) - \alpha_w + \tau\big)}{\alpha_w\big(r(1-N(t)/K) - \alpha_m + \tau\big)} \;\approx\; \frac{\alpha_m}{\alpha_w} \;=\; 1 + \frac{\Delta\alpha}{\alpha_w}, \tag{S24}$$

as α_m and α_w are very small values.

Likewise, the differentiation between Y and L cells:

$$\mathcal{Q}^{YL}(t) = \frac{f(t)}{1-f(t)}\frac{1-p(t)}{p(t)}, \tag{S25}$$

converges towards:

$$\mathcal{Q}^{YL} = \frac{\alpha_m\big(r(1-N(t)/K) - \alpha_w + aN(t)\big)\big(r(1-N(t)/K) - \alpha_w + \tau\big)}{\alpha_w\big(r(1-N(t)/K) - \alpha_m + aN(t)\big)\big(r(1-N(t)/K) - \alpha_m + \tau\big)} \;\approx\; \frac{\alpha_m}{\alpha_w} \;=\; 1 + \frac{\Delta\alpha}{\alpha_w} \tag{S26}$$

and the differentiation between free phages (V) and Y cells:

$$\mathcal{Q}^{VY}(t) = \frac{q(t)}{1-q(t)}\frac{1-f(t)}{f(t)} \tag{S27}$$

converges towards:

$$\mathcal{Q}^{VY} = \frac{r(1-N(t)/K) + aN(t) - \alpha_w}{r(1-N(t)/K) + aN(t) - \alpha_m} \;\approx\; 1. \tag{S28}$$

When the system stabilizes (endemic or late stage of the epidemic case), the virulent strain is therefore more frequent among free viruses and Y cells than among L cells (prophages) but we also expect almost no differentiation between free viruses and Y cells (**Fig. S4**).

S3 Statistical inference of the rates of prophage reactivation

From the previous analysis of the model, we show that when the system reaches high prevalence the selection gradient $\mathscr{S}$ of the variant is simply given by: $\mathscr{S} = \alpha_w - \alpha_m = -\Delta\alpha$ (S22); and the differentiation $\mathcal{Q}^{VL}$ of the variant between free phages and prophages by: $\mathcal{Q}^{VL} = \alpha_m/\alpha_w = 1 + \Delta\alpha/\alpha_w$ (S24).

Three quantities are tracked over time in the experiment: (i) the prevalence $P(t)$, (ii) the frequency of hosts infected by the virulent phage $g(t)$ and (iii) the frequency of the virulent phage in the free virus stage $q(t)$. We only keep data from the time point the system has reached high prevalence ($\geq 95\%$). We recall that, in these conditions, $g(t)$ is almost entirely driven by the frequency of L cells infected by the virulent phage, that is $g(t) \approx p(t)$ (cf. **§S2.3**).

S3.1 Frequentist approach

We fit a linear mixed-effects model on the logit-frequency infected by the virulent phage $\text{logit}\,(g(t))$ to estimate the slope $\mathscr{S}$. For treatment i (epidemic vs. endemic), in chemostat j and at time point t, we have (t, i and j are now noted as indexes for clarity):

$$\underbrace{\text{logit}\,(g)_{i,j,t}}_{\text{Response variable}} = \textit{intercept} + \mathscr{S} \times t + \beta_i + \chi_j + \varepsilon_{i,j,t},$$

with:

- $\textit{intercept}$, the common fixed effect (reference);
- β_i, the fixed effect of treatment i on the intercept of the model;
- $\chi_j \overset{i.i.d}{\sim} \mathcal{N}\left(0,\ \gamma^2\right)$, the random effect (with variance γ^2) of the j^{th} chemostat on the intercept of the model;
- $\varepsilon_{i,j,t} \overset{i.i.d}{\sim} \mathcal{N}\left(0,\ \sigma_g^2\right)$, the residual error (with variance σ_g^2).

We fit this mixed-effects model using the function '*lmer*' from the R package '*lme4*'. Alongside, we also propose a version where the rates of prophage reactivation α_w and α_m (and therefore the slope $\mathscr{S}$) are allowed to vary across chemostats; this slightly changes the previous linear model to:

$$\text{logit}\,(g)_{i,j,t} = \textit{intercept} + (\mathscr{S} + \kappa_j) \times t + \beta_i + \chi_j + \varepsilon_{i,j,t},$$

with κ_j the fixed effect of the j^{th} chemostat on the slope of the model.

In parallel, we estimate $\mathcal{Q}^{VL}$, starting from the computation of $\ln\left(\mathcal{Q}^{VL}(t)\right) = \text{logit}\,(q(t)) - \text{logit}\,(p(t))$, in which we substitute $\text{logit}\,(p(t))$ by $\text{logit}\,(g(t))$. Using the same notations for the indexes we have:

$$\underbrace{\text{logit}\,(q)_{i,j,t} - \text{logit}\,(g)_{i,j,t}}_{\ln(\mathcal{Q}^{VL})_{i,j,t},\ \text{response variable}} = \ln\left(\mathcal{Q}^{VL}\right) + \varepsilon_{i,j,t},$$

where $\varepsilon \overset{i.i.d}{\sim} \mathcal{N}\left(0,\ \sigma_Q^2\right)$ is the residual error (with variance σ_Q^2). We thus estimate $\mathcal{Q}^{VL}$ by computing, all chemostats combined, the exponential of the arithmetic mean of $\ln\left(\mathcal{Q}^{VL}(t)\right)$ – this is completely equivalent to computing the geometric mean of $\mathcal{Q}^{VL}(t)$.

Alternatively, if we want to allow α_w and α_m (and therefore $\mathcal{Q}^{VL}$) to vary across chemostats:

$$\text{logit}\,(q)_{i,j,t} - \text{logit}\,(g)_{i,j,t} = \ln\left(\mathcal{Q}^{VL}\right)_j + \varepsilon_{i,j,t},$$

and we just need to compute the mean for each chemostat.

Finally, we obtain point estimates of both rates of prophage reactivation α_w and α_m combining equations (S22) and (S24):

$$\begin{cases} \alpha_w &= \dfrac{\mathscr{S}}{1 - \mathcal{Q}^{VL}} \\ \alpha_m &= \dfrac{\mathscr{S} \times \mathcal{Q}^{VL}}{1 - \mathcal{Q}^{VL}} \end{cases} \tag{S29}$$

in which we use the estimated values of $\mathscr{S}$ and $\mathcal{Q}^{VL}$. From our experimental data, all chemostats combined, we get (expressed in h^{-1}): $\hat{\alpha}_w = 2.58 \times 10^{-3}$ and $\hat{\alpha}_m = 1.19 \times 10^{-2}$.

S3.2 Bayesian approach

Besides, we also use a Bayesian approach to estimate parameters α_w and α_m along with their 95% credible intervals. For treatment i (epidemic vs. endemic), in chemostat j and at time point t, we consider the following likelihoods for the response variables (t, i and j are now noted as indexes for clarity):

$$\begin{cases} \text{logit}\left(g\right)_{i,j,t} \sim \mathcal{N}\Big(\beta_i + \chi_j + \overbrace{(\alpha_w - \alpha_m)}^{\mathscr{S}} \times t,\ \sigma_g^2\Big) \\ \Big(\text{logit}\,(q)_{i,j,t} - \text{logit}\,(g)_{i,j,t}\Big) \sim \mathcal{N}\Big(\underbrace{\ln(\alpha_m) - \ln(\alpha_w)}_{\ln(\mathcal{Q}^{VL})},\ \sigma_Q^2\Big) \end{cases}$$

with, as before:

- β_i, the fixed effect of treatment i on the intercept of the model;

- $\chi_j \sim \mathcal{N}\left(0,\ \gamma^2\right)$, the random effect (with variance γ^2) of the j^{th} chemostat on the intercept of the model;

- σ_g^2 and σ_Q^2, the variance parameters of the response variables.

We choose the following prior distributions:

$$\begin{aligned}
\alpha_w &\sim \mathcal{U}\left([0,\ 0.1]\right) \\
\alpha_m &\sim \mathcal{U}\left([0,\ 0.1]\right) \\
\beta_{epidemic} &\sim \mathcal{U}\left([0.25,\ 3]\right) \\
\beta_{endemic} &\sim \mathcal{U}\left([-0.25,\ 0.25]\right) \\
\gamma &\sim \mathcal{U}\left([0,\ 3]\right) \\
\sigma_Q &\sim \mathcal{U}\left([0,\ 5]\right) \\
\sigma_g &\sim \mathcal{U}\left([0,\ 1]\right)
\end{aligned}$$

Again, if we allow parameters α_w and α_m to vary across chemostats:

$$\begin{cases} \text{logit}(g)_{i,j,t} \sim \mathcal{N}\Big(\beta_i + \chi_j + (\alpha_{w,j} - \alpha_{m,j}) \times t,\ \sigma_g^2\Big) \\ \Big(\text{logit}\,(q)_{i,j,t} - \text{logit}\,(g)_{i,j,t}\Big) \sim \mathcal{N}\Big(\ln(\alpha_{m,j}) - \ln(\alpha_{w,j}),\ \sigma_Q^2\Big) \end{cases}$$

where the prior of each $\alpha_{w,j}$ and $\alpha_{m,j}$ is the same as for α_w and α_m above.

We independently run 4 Monte-Carlo Markov chains using JAGS (Plummer et al., 2003) version 4.3.0 and function '*jags*' from the R package '*R2jags*'. Each chain is 20,000 iterations long (length of the burn-in period 8,000) with thinning rate 50. As initial conditions: α_w and α_m are drawn uniformly between 0 and 3×10^{-2} such that $\alpha_w < \alpha_m$; $\beta_{epidemic}$ is drawn uniformly between 0.25 and 3, and $\beta_{endemic}$ between -0.25 and 0.25; χ_j is drawn from a standard Normal distribution; and variance parameters σ_Q, σ_g and γ are drawn uniformly between 0 and 1. In the end, we assess convergence of posterior distributions, especially we check that Gelman-Rubin statistics are below 1.1 and that effective sample sizes are above 100.

From our experimental data, all chemostats combined, we get (expressed in h^{-1}): $\alpha_w = 2.65 \times 10^{-3}$ (mean, 95% credible interval $[2.03 \times 10^{-3},\ 3.32 \times 10^{-3}]$) and $\alpha_m = 1.21 \times 10^{-2}$ (mean, 95% credible interval $[9.48 \times 10^{-3},\ 1.47 \times 10^{-2}]$) (**Fig. S14-A**); see **Fig. S14-B** for posterior distributions by chemostat.

Supplementary references

Diekmann, O., Heesterbeek, J., & Roberts, M. G. (2010). The construction of next-generation matrices for compartmental epidemic models. *Journal of the royal society interface*, *7*(47), 873–885. https://doi.org/10.1098/rsif.2009.0386

Plummer, M., et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*, *124*(125.10), 1–10.

Rinaldi, S., & Scheffer, M. (2000). Geometric analysis of ecological models with slow and fast processes. *Ecosystems*, *3*, 507–521. https://doi.org/10.1007/s100210000045

St-Pierre, F., & Endy, D. (2008). Determination of cell fate selection during phage lambda infection. *Proceedings of the National Academy of Sciences*, *105*(52), 20705–20710. https://doi.org/10.1073/pnas.0808831105

Sussman, R., & Jacob, F. (1962). [On a thermosensitive repression system in the escherichia coli lambda bacteriophage]. *Comptes rendus hébdomadaires des séances de l'Académie des sciences*, *254*, 1517–1519.

Verhulst, F. (2007). Singular perturbation methods for slow–fast dynamics. *Nonlinear Dynamics*, *50*, 747–753. https://doi.org/10.1007/s11071-007-9236-z

Chapter four

Host movements and pathogen evolution

Wakinyan Benhamou iD, Rémi Choquet iD and Sylvain Gandon iD

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

September 6, 2024

Abstract

Pathogen adaptation erodes our ability to mitigate epidemics and represents a major threat for public health. Upon the acquisition of beneficial mutations, novel variants emerge and sometimes invade the population. The strength of selection of an emerging variant is classically quantified using its frequency changes. Such approach is particularly well suited assuming an homogeneous-mixing population. Yet, spatially separated host populations are most often interconnected through movements ("migration") of susceptible and infected individuals. Migration is a key force that may lead to new pathogen introductions, global pathogen persistence, or affect the spread of epidemics. In addition to its effects on the epidemiology, migration may also interfere with the evolution of a polymorphic pathogen population. Yet, little is known about how migration shapes the evolutionary dynamics of the pathogen. In this study, we track the transient dynamics of the frequency and differentiation of a variant in competition with the wildtype strain in a two-patch metapopulation in which hosts commute. We consider a scenario with homogeneous selection in both populations, and a scenario with heterogeneous selection. Overall, we emphasize the usefulness of mechanistic dynamical models to disentangle the effect of migration and selection; migration can blur the quantification of the strength of selection and lead to erroneous estimations of the selective advantages of variants.

1 Introduction

Spatial heterogeneity can play a significant role in epidemiological dynamics. In particular, host movements between different areas may allow infectious diseases to persist globally by countering local extinctions (Lloyd & May, 1996; Post et al., 1983). On the other hand, outbreaks are expected to be less explosive within fragmented populations because pathogens have access to fewer susceptible hosts than within spatially uniform populations (Post et al., 1983). Spatial transmission of diseases spread by direct contagion is bound to host movements. In open environments, the force of infection is driven by local transmissions – as within closed environments – but also by transmissions between individuals from distinct areas. In public health, human mobility can be a particularly important feature for the spread of infectious diseases and a considerable number of epidemiological studies has thus often taken such spatial interactions into account. In the context of a highly interconnected and interdependent world (Hufnagel et al., 2004), human infectious diseases spread faster and at larger scales, as exemplified by the recent COVID-19 (Coronavirus Disease 2019) pandemic. Nowadays, human traveling behavior can be accurately quantified, for instance using travel history data (Butera et al., 2021; Lemey et al., 2020), air-traffic data (Brockmann & Helbing, 2013; Hufnagel et al., 2004) or cell-phone mobility data (Kraemer et al., 2021; Le Treut et al., 2022).

From a modelling perspective, epidemic compartment models popularized by (Kermack & McKendrick, 1927) such as the classical SIR model typically assume an homogeneous-mixing population; each individual has the same probability to meet any other individual, so that there is no spatial effect on the spread of the disease (Lipshtat et al., 2021). Using a system of partial differential equations, classical compartmental models can be extended by adding a spatial diffusion term to model the dynamics of the continuous spatial distributions of hosts (Murray, 2003; Postnikov & Sokolov, 2007), with solutions in form of traveling waves. Diffusion models typically hold for spatially limited dispersal, that is with rather short distances compared to geographical distances (in contrast with air transportation for instance) (Hufnagel et al., 2004; Le Treut et al., 2022). A very simple way to take into account the arrival of new infected hosts would be to add an immigration term of infected hosts in the analysis of the dynamics of a focal population (e.g., (Engbert & Drepper, 1994)). A more advanced approach would be to consider a metapopulation – i.e., a network of spatially separated patches interconnected through migration flows –; here, "patch" is a generic name for a group of hosts which, depending on its definition, can refer to households, cities, regions, countries, etc... (Grenfell & Harwood, 1997). Spatial heterogeneity is considered at the scale of the metapopulation, while each component patch is still homogeneous. Such models has been used widely in the context of infectious diseases (e.g., (Brockmann & Helbing, 2013; Grenfell & Harwood, 1997; Hethcote, 1978; Lajmanovich & Yorke, 1976; Post et al., 1983; Yuksel et al., 2021)), and very recently for the COVID-19 pandemic (e.g., (Le Treut et al., 2022; Roques et al., 2020)). In many cases, increasing distances between patches are assumed to lead to decreasing contact probabilities – e.g., due to travel time and cost –, so that transmission is often weighted by some functions of the distances (e.g., power law decay or exponential distance weighting (Roques et al., 2020; Xia et al., 2004)). Besides, host movements between large populations are typically much more frequent than between small populations. Host movements depend therefore on both distances between populations and population sizes (Erlander & Stewart, 1990; Xia et al., 2004). Using gravity models – from transportation theory, and inspired from Newton's law of universal gravitation (Erlander & Stewart, 1990) –, transmissions between two populations are expected to be positively correlated with population sizes and negatively correlated with the distance separating the two populations (Xia et al., 2004).

The effects of dispersal/migration has also been extensively studied in evolutionary studies, and especially in population genetics, but not much in evolutionary epidemiology. The effects of spatial heterogeneity and migration has for instance been investigated for the evolution of host life-history traits (Débarre et al., 2012), the emergence of drug resistance (Débarre et al., 2007), the emergence of specialist or generalist strains (Débarre et al., 2013; Ronce & Kirkpatrick, 2001), the evolution of pathogen virulence (Berngruber et al., 2015; Boots & Sasaki, 1999) or the phenotypic evolution of pathogen with vaccination coverage (Walter et al., 2024; Zurita-Gutiérrez & Lion, 2015). However, in most cases, the interest lies in the long-term outcome of the competition between different pathogenic strains. Yet, migration can also interfere with the short-term rise of variants and affect evolutionary dynamics. For example, the differential growth of the Delta variant of SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) in India and in different regions of England (Volz, 2023) may challenge the hypothesis that the dynamics of the variant frequency is solely driven by differences in fitness, and such discrepancy might be explained by other processes such as migration. Classically, the strength of selection of emerging variants is quantified by estimating the slope of the changes of the variant frequency on the logit scale, that is the log odds ln(frequency of the variant/frequency of the wildtype) (Boyle et al., 2022; Otto et al., 2021; Volz, 2023). However, movements of susceptible and infected hosts can also affect the evolutionary dynamics of pathogens.

In this study, we model a polymorphic pathogen population – competition between the wildtype and a variant – in an Susceptible-Infected-Recovered-Susceptible (SIRS) model. We first recall some epidemiological and evolutionary dynamics within a closed host population. We then focus on a two-patch host metapopulation. Using an evolutionary epidemiology approach, we track the short (transient) and long-term outcome of the competition and, in particular, the dynamics of the variant frequencies and of the variant differentiation between the two host populations.

2 Models

2.1 An SIRS model within a closed population

We begin with a simple compartmental SIRS model in a closed population of constant total density N. We model the dynamics of a directly and horizontally transmitted infectious disease. Hosts are either susceptible (S), infected and infectious (I) or recovered and immune (R). We denote β, the *per capita* transmission rate, which captures both the host contact rate and the probability of transmission per contact with an infected host; transmission is assumed to be direct, horizontal and frequency-dependant – i.e., the force of infection is given by $\beta I(t)/N$, where $I(t)$ is the density of infected hosts at time t. Then, infected hosts recover with a *per capita* recovery rate (or infection clearance) γ and acquire a full immunity that wane at a *per capita* rate ζ. We consider that two strains of a pathogen are co-circulating in the host population: the wildtype (w) and the mutant strain (m), or variant. Therefore, an individual may either be infected by the wildtype (I_w) or by the variant (I_m) (no coinfections). The variant may differ phenotypically from the wildtype in terms of transmission rate ($\beta_m = \beta_w + \Delta\beta$) and/or of recovery rate ($\gamma_m = \gamma_w + \Delta\gamma$). We describe these epidemiological dynamics with the following system of ordinary differential equations (ODEs), where the dots refer to differentiation with respect to time:

$$\begin{cases} \dot{S}(t) &= -\overline{\beta}(t)\dfrac{I(t)}{N}S(t) + \zeta R(t) \\ \dot{I}(t) &= \overline{\beta}(t)\dfrac{I(t)}{N}S(t) - \overline{\gamma}(t)I(t) \\ \dot{R}(t) &= \overline{\gamma}(t)I(t) - \zeta R(t) \end{cases} \tag{1}$$

where the overlines refer to the mean values of phenotypic traits:

$$\begin{cases} \overline{\beta}(t) &= \big(1 - q(t)\big)\beta_w + q(t)\beta_m \\ \overline{\gamma}(t) &= \big(1 - q(t)\big)\gamma_w + q(t)\gamma_m \end{cases} \tag{2}$$

with $q(t) = I_m(t)/I(t)$, the frequency of the variant.

2.2 An SIRS model within a two-patch metapopulation

Building upon the previous SIRS model, we now assume that two spatially separated host populations A and B are interconnected by migration (two-patch metapopulation, see **Fig. 1**). Throughout, we add a superscript A or B to distinguish the two populations. The probability that a host from population A (resp. B) visit population B (resp. A) is denoted by ω^A (resp. ω^B). Visits are assumed to be only temporary and "migrants" to return shortly to their home population, as commuters (Lipshtat et al., 2021). Strictly speaking, we thus model the migration of the disease rather than the migration of hosts (Post et al., 1983). Due to differences in local conditions, we assume that phenotypic traits β and γ can vary between population A and B, so that the phenotype of the variant in population A (resp. B) is given by $\beta_m^A = \beta_w^A + \Delta\beta^A$ and $\gamma_m^A = \gamma_w^A + \Delta\gamma^A$ (resp. $\beta_m^B = \beta_w^B + \Delta\beta^B$ and $\gamma_m^B = \gamma_w^B + \Delta\gamma^B$). We then refer to these differences between populations as heterogeneous selection.

The force of infection – i.e., the *per capita* infection rate – becomes the sum of the contributions of four types of intra or inter-community interactions (Le Treut et al., 2022). The intra-community contributions correspond to the transmissions between a susceptible and an infected host that belongs to the same population, which can occur either (i) in the local population or (ii) in the visited population (visitor-to-visitor infection). On the contrary, the inter-community contributions correspond to the transmissions between a susceptible and an infected host that do not belong to the same population, which can occur (i) when a susceptible visitor gets infected by a native or (ii) when a susceptible native gets infected by a visitor. Epidemiological dynamics for population A are now given by the following system of ODEs:

$$
\begin{cases}
\dot{S}^A(t) &= -\Big[\left(1-\omega^A\right)\lambda^A(t)+\omega^A\lambda^B(t)\Big]S^A(t)+\zeta R^A(t)\\
\dot{I}^A(t) &= \Big[\left(1-\omega^A\right)\lambda^A(t)+\omega^A\lambda^B(t)\Big]S^A(t)-\displaystyle\sum_{k\in\{w,m\}}\gamma_k^A I_k^A(t)\\
\dot{R}^A(t) &= \displaystyle\sum_{k\in\{w,m\}}\gamma_k^A I_k^A(t)-\zeta R^A(t)\\
\\
\dot{S}^B(t) &= -\Big[\left(1-\omega^B\right)\lambda^B(t)+\omega^B\lambda^A(t)\Big]S^B(t)+\zeta R^B(t)\\
\dot{I}^B(t) &= \Big[\left(1-\omega^B\right)\lambda^B(t)+\omega^B\lambda^A(t)\Big]S^B(t)-\displaystyle\sum_{k\in\{w,m\}}\gamma_k^B I_k^B(t)\\
\dot{R}^B(t) &= \displaystyle\sum_{k\in\{w,m\}}\gamma_k^B I_k^B(t)-\zeta R^B(t)
\end{cases}
\tag{3}
$$

with $\lambda^A(t)$ and $\lambda^B(t)$, the forces of infection in population A and population B, respectively (**Fig. 1**). The force of infection in population A $\lambda^A(t)$ is given by:

$$
\lambda^A(t)=\sum_{k\in\{w,m\}}\beta_k^A\frac{\left(1-\omega^A\right)I_k^A(t)+\omega^B I_k^B(t)}{\left(1-\omega^A\right)N^A+\omega^B N^B}. \tag{4}
$$

Substituting A by B and vice-versa gives the expression for the force of infection in population B $\lambda^B(t)$. We now assume that migration probabilities ω^A and ω^B are small, of order $\varepsilon \ll 1$ ($\mathcal{O}(\varepsilon)$). Treating ω^A as $\varepsilon\omega^A$ and ω^B as $\varepsilon\omega^B$ to emphasize that migration probabilities are small, a Taylor expansion about $\varepsilon = 0$ yields:

$$
\lambda^A(t)=\sum_{k\in\{w,m\}}\beta_k^A\frac{\left(1-\omega^B\frac{N^B}{N^A}\right)I_k^A(t)+\omega^B I_k^B(t)}{N^A}+\mathcal{O}\left(\varepsilon^2\right). \tag{5}
$$

Note that the term $\mathcal{O}(\varepsilon)$ depends on ω^B but not on ω^A. Under the assumption that migration probabilities are small, a Taylor expansion of the ODE system (3) about $\varepsilon = 0$ yields the following epidemiological dynamics for infected hosts from population A:

$$
\begin{aligned}
\dot{I}^A(t) = &\Bigg[\overbrace{\left(1-\omega^A-\omega^B\frac{N^B}{N^A}\right)\sum_k\beta_k^A\frac{I_k^A(t)}{N^A}}^{\text{Local intra-community infections}}+\overbrace{\omega^B\sum_k\beta_k^A\frac{I_k^B(t)}{N^A}+\omega^A\sum_k\beta_k^B\frac{I_k^B(t)}{N^B}}^{\text{Inter-community infections}}\Bigg]S^A(t)\\
&-\sum_k\gamma_k^A I_k^A(t)+\mathcal{O}\left(\varepsilon^2\right).
\end{aligned}
\tag{6}
$$

Note that the visitor-to-visitor contribution to the force of infection is included in the term $\mathcal{O}\left(\varepsilon^2\right)$.

In the following, $q(t) = I_m(t)/I(t)$ refers to the overall frequency of the variant (i.e., at the scale of the metapopulation) at the current time t, and $q^A(t) = I_m^A(t)/I^A(t)$ and $q^B(t) = I_m^B(t)/I^B(t)$, the frequency of the variant in population A and B, respectively. Besides, we define the differentiation of the variant $\mathcal{Q}$ between population A and B by:

$$\mathcal{Q}(t) = \frac{q^A(t)}{1 - q^A(t)} \frac{1 - q^B(t)}{q^B(t)}, \tag{7}$$

such that:

$$\ln\left(\mathcal{Q}(t)\right) = \text{logit}\left(q^A(t)\right) - \text{logit}\left(q^B(t)\right).$$

When there is no differentiation between the two populations, $\mathcal{Q}(t) = 1$ and $\ln\left(\mathcal{Q}(t)\right) = 0$.

3 Results

3.1 Within a closed host population

Using (1) and (2), we derive the temporal dynamics of the frequency of the variant $q(t)$ within a closed population:

$$\dot{q}(t) = \underbrace{q(t)\big(1 - q(t)\big)}_{\text{genetic variance}} \underbrace{\left(\Delta\beta\frac{S(t)}{N} - \Delta\gamma\right)}_{\mathcal{S}(t),\ \text{selection gradient}}. \tag{8}$$

More conveniently, we then focus on the logit-frequency of the variant – i.e., the log odds ln(frequency of the variant/frequency of the wildtype) –:

$$\frac{\mathrm{d}\,\text{logit}\left(q(t)\right)}{\mathrm{d}t} = \mathcal{S}(t). \tag{9}$$

The direction and speed of selection is governed by the sign and magnitude of the selection gradient $\mathcal{S}(t)$, respectively, which depends on the phenotypic differences $\Delta\beta$ and $\Delta\gamma$ and, when $\Delta\beta \neq 0$, on the availability of susceptible hosts (Day & Gandon, 2006, 2007; Day et al., 2020). When one strain (let say with phenotypes β_k and γ_k, such that $\beta_k > \gamma_k$) outcompetes the other, the system converges in the long term towards the following epidemiological attractor (endemic equilibrium of strain k):

$$\begin{cases} S(\infty) &= \dfrac{\gamma_k}{\beta_k} N \\ I(\infty) &= \dfrac{\zeta\left(\beta_k - \gamma_k\right)}{\beta_k\left(\gamma_k + \zeta\right)} N \\ R(\infty) &= \dfrac{\gamma_k\left(\beta_k - \gamma_k\right)}{\beta_k\left(\gamma_k + \zeta\right)} N \end{cases} \tag{10}$$

3.2 Within a two-patch host metapopulation

3.2.1 Homogeneous selection

We first consider the scenario where the phenotypic traits of each strain are identical in both populations, which yields $\Delta\beta^A = \Delta\beta^B = \Delta\beta$ and $\Delta\gamma^A = \Delta\gamma^B = \Delta\gamma$ (we relax this assumption later to account for heterogeneous selection). Using equation (6) (two-patch metapopulation) with homogeneous selection, the temporal dynamics of the logit-frequency of the variant in population A is given by:

$$\frac{\mathrm{d}\,\mathrm{logit}\left(q^A(t)\right)}{\mathrm{d}t} = \Delta(t) + \Omega(t) + \mathcal{O}\left(\varepsilon^2\right), \tag{11}$$

with:

$$\begin{cases} \Delta(t) & = \Delta\beta\left[\dfrac{1-\omega^A - \frac{N^B}{N^A}\omega^B}{N^A} + \left(\dfrac{\omega^B}{N^A} + \dfrac{\omega^A}{N^B}\right)\dfrac{q^B(t)}{q^A(t)}\dfrac{I^B(t)}{I^A(t)}\right] S^A(t) - \Delta\gamma \\ \Omega(t) & = -\dfrac{q^A(t)-q^B(t)}{q^A(t)\left(1-q^A(t)\right)}\beta_w\left(\dfrac{\omega^B}{N^A} + \dfrac{\omega^A}{N^B}\right)\dfrac{I^B(t)}{I^A(t)} S^A(t) \end{cases}$$

$\Delta(t)$ represents the effect of selection and $\Omega(t)$, the effect of pure migration – i.e., independently of any phenotypic difference. The slope of the variant logit-frequency depends on the phenotypic differences between the two strains and the availability of susceptible hosts – as within closed populations –, but also on the migration probabilities and on demographic ratios between the two populations. The direction of $\Omega(t)$ is opposite to the sign of $q^A(t) - q^B(t)$, that is migration tends to homogenize the frequency of the variant between the two populations. When there is no migration ($\omega^A = \omega^B = 0$), equation (11) reduces to equation (9).

To simplify further, let us assume that only hosts from population A may visit population B ($\omega^A = \omega$ and $\omega^B = 0$) and the variant has reached fixation in population B ($q^B = 1$), we obtain:

$$\frac{\mathrm{d}\,\mathrm{logit}\left(q^A(t)\right)}{\mathrm{d}t} = \underbrace{\Delta\beta\left[\frac{1-\omega}{N^A} + \frac{\omega}{N^B}\frac{I^B(t)}{q^A(t)I^A(t)}\right]S^A(t) - \Delta\gamma^A}_{\Delta(t)} \underbrace{+\frac{\omega}{q^A(t)}\frac{\beta_w}{N^B}\frac{I^B(t)}{I^A(t)}S^A(t)}_{\Omega(t)} + \mathcal{O}\left(\varepsilon^2\right).$$

As we assume that migration is low ($\omega \ll 1$), one might thus expect that $\Omega(t)$ would be small or negligible compared to $\Delta(t)$. However, when the local frequency of the variant is even smaller ($q^A(t) \ll \omega$), $\Omega(t)$ can be very large. The effect of migration can thus amplify significantly the effect of selection at the beginning of the sweep of the variant; migration effects then diminishes as the variant increases in frequency (**Fig. 2**).

Instead of focusing on the variant frequency in one focal population, one can also focus on the whole metapopulation. The temporal dynamics of the overall logit-frequency of the variant (i.e., at the scale of the metapopulation) is given by:

$$\begin{aligned}\frac{\mathrm{d}\,\mathrm{logit}\,(q(t))}{\mathrm{d}t} =&\Delta\beta\left[\left[\frac{1}{N^A}\left(\left(1-\omega^A-\omega^B\frac{N}{N^A}\right)\frac{q^A(t)}{q(t)}\frac{I^A(t)}{I(t)}+\omega^B\right)+\frac{\omega^A}{N^B}\frac{q^B(t)}{q(t)}\frac{I^B(t)}{I(t)}\right]S^A(t)+\right.\\ &\left.\left[\frac{1}{N^B}\left(\left(1-\omega^B-\omega^A\frac{N}{N^B}\right)\frac{q^B(t)}{q(t)}\frac{I^B(t)}{I(t)}+\omega^A\right)+\frac{\omega^B}{N^A}\frac{q^A(t)}{q(t)}\frac{I^A(t)}{I(t)}\right]S^B(t)\right]-\Delta\gamma\\ &-\frac{q^B(t)-q^A(t)}{q(t)\left(1-q(t)\right)}\beta_w\left[\left(\left(1-\omega^B-\omega^A\frac{N}{N^B}\right)\frac{1}{N^B}-\omega^B\frac{1}{N^A}\right)S^B(t)-\right.\\ &\left.\left(\left(1-\omega^A-\omega^B\frac{N}{N^A}\right)\frac{1}{N^A}-\omega^A\frac{1}{N^B}\right)S^A(t)\right]\frac{I^A(t)}{I(t)}\frac{I^B(t)}{I(t)}\;+\;\mathcal{O}\left(\varepsilon^2\right).\end{aligned} \tag{12}$$

We then focus on the dynamics of the differentiation $\mathcal{Q}$ of the variant between the two populations. Using equation (11) and its counterpart for population B, we obtain the following expression for the temporal dynamics of $\ln\left(\mathcal{Q}(t)\right)$:

$$\begin{aligned}\frac{\mathrm{d}\ln\left(\mathcal{Q}(t)\right)}{\mathrm{d}t} =&\Delta\beta\left[\left(\frac{1}{N^A}\left(1-\omega^A-\omega^B\frac{N^B}{N^A}\right)+\left(\frac{\omega^B}{N^A}+\frac{\omega^A}{N^B}\right)\frac{q^B(t)}{q^A(t)}\frac{I^B(t)}{I^A(t)}\right)S^A(t)-\right.\\ &\left.\left(\frac{1}{N^B}\left(1-\omega^B-\omega^A\frac{N^A}{N^B}\right)+\left(\frac{\omega^A}{N^B}+\frac{\omega^B}{N^A}\right)\frac{q^A(t)}{q^B(t)}\frac{I^A(t)}{I^B(t)}\right)S^B(t)\right]\\ &-\left(\mathcal{Q}(t)-1\right)\beta_w\left(\frac{\omega^B}{N^A}+\frac{\omega^A}{N^B}\right)\left(\frac{q^B(t)}{q^A(t)}\frac{I^B(t)}{I^A(t)}S^A(t)+\frac{1-q^A(t)}{1-q^B(t)}\frac{I^A(t)}{I^B(t)}S^B(t)\right)+\mathcal{O}\left(\varepsilon^2\right).\end{aligned} \tag{13}$$

The first two lines of (13) represents the effect of selection driven by the phenotypic differences between the two strains; the third line represents the effect of pure migration, independently of the phenotypic differences, whose direction depends on the sign of $\mathcal{Q}(t)-1$ which tends to homogenize the two populations. In the long term, $\mathcal{Q}(t)$ is thus expected to converge towards 1, that is $q^A(t)=q^B(t)$.

We run some simulations of the SIRS model (3) with homogeneous selection. Starting with population A at the endemic equilibrium of the wildtype and population B at the endemic equilibrium of a variant that is selected for, migration accelerates the rise of the variant in population A as well as the homogenization between the two populations, especially at the beginning of the sweep (**Fig. 3-A**) – the slope of the logit-frequency then converges towards that without migration. In contrast with closed populations, the variant can still increase in frequency in population A in the neutral case ($\Delta(t)=0$), driven solely by migration ($\Omega(t)>0$) because the variant is more frequent in population B ($q^A(t)<q^B(t)$) (**Fig. 3-B.1**). This scenario emphasizes that the slope of a variant logit-frequency can be completely misleading about the potential adaptive advantage of the variant in real-life situations. The rise of the variant is also accelerated when the total density of the population in which the variant is more frequent is larger ($N^B>N^A$, **Fig. 3-C.1**); on the other hand, the dynamics of the differentiation of the variant is always faster when an asymmetry between the total population densities exists (**Fig. 3-C.2**).

We also consider a scenario where the pathogen is introduced in equal density and frequency in both populations, so that initially there is no differentiation between population A and B ($\mathcal{Q}(0) = 1$). Such scenario could for example occur during the introduction of a small quantity of the pathogen. Interestingly, an asymmetry between the total population densities ($N^A \neq N^B$) can result in a transient differentiation between population A and B (**Fig. 4**), even though selection is homogeneous. As the pathogen is introduced with identical density in both populations, susceptible hosts do not represent the same fraction of the population if total population densities differ. This differential availability of susceptible hosts leads to differential selection that transiently disrupts the differentiation of the variant between the two populations. In the long term, however, there is no longer any differentiation.

3.2.2 Heterogeneous selection

We recall that the temporal dynamics of the logit-frequency of the variant in population A is given by:

$$\frac{\mathrm{d}\,\mathrm{logit}\left(q^A(t)\right)}{\mathrm{d}t} = \Delta(t) + \Omega(t) + \mathcal{O}\left(\varepsilon^2\right).$$

with $\Delta(t)$, the effect of selection, and $\Omega(t)$, the effect of pure migration. When the phenotypic differences between the variant and the wildtype differ in the two populations, we obtain:

$$\begin{cases} \Delta(t) &= \left[\dfrac{\Delta\beta^A}{N^A}\left(1 - \omega^A - \dfrac{N^B}{N^A}\omega^B\right) + \left(\Delta\beta^A \dfrac{\omega^B}{N^A} + \Delta\beta^B \dfrac{\omega^A}{N^B}\right) \dfrac{q^B(t)}{q^A(t)} \dfrac{I^B(t)}{I^A(t)}\right] S^A(t) - \Delta\gamma^A \\ \Omega(t) &= -\dfrac{q^A(t) - q^B(t)}{q^A(t)\left(1 - q^A(t)\right)} \left(\beta_w^A \dfrac{\omega^B}{N^A} + \beta_w^B \dfrac{\omega^A}{N^B}\right) \dfrac{I^B(t)}{I^A(t)} S^A(t) \end{cases} \tag{14}$$

Likewise, the temporal dynamics of the overall logit-frequency of the variant is given by:

$$\begin{aligned} \frac{\mathrm{d}\,\mathrm{logit}\left(q(t)\right)}{\mathrm{d}t} = &\left[\frac{\Delta\beta^A}{N^A}\left(\left(1 - \omega^A - \omega^B \frac{N}{N^A}\right) \frac{q^A(t)}{q(t)} \frac{I^A(t)}{I(t)} + \omega^B\right) + \Delta\beta^B \frac{\omega^A}{N^B} \frac{q^B(t)}{q(t)} \frac{I^B(t)}{I(t)}\right] S^A(t) + \\ &\left[\frac{\Delta\beta^B}{N^B}\left(\left(1 - \omega^B - \omega^A \frac{N}{N^B}\right) \frac{q^B(t)}{q(t)} \frac{I^B(t)}{I(t)} + \omega^A\right) + \Delta\beta^A \frac{\omega^B}{N^A} \frac{q^A(t)}{q(t)} \frac{I^A(t)}{I(t)}\right] S^B(t) - \\ &\left(\frac{q^A(t)}{q(t)} \frac{I^A(t)}{I(t)} \Delta\gamma^A + \frac{q^B(t)}{q(t)} \frac{I^B(t)}{I(t)} \Delta\gamma^B\right) \\ &- \frac{q^A(t) - q^B(t)}{q(t)\left(1 - q(t)\right)} \left[\left(\left(1 - \omega^B - \omega^A \frac{N}{N^B}\right) \frac{\beta_w^B}{N^B} - \omega^B \frac{\beta_w^A}{N^A}\right) S^B(t) - \right. \\ &\qquad \left. \left(\left(1 - \omega^A - \omega^B \frac{N}{N^A}\right) \frac{\beta_w^A}{N^A} - \omega^A \frac{\beta_w^B}{N^B}\right) S^A(t) - \left(\gamma_w^B - \gamma_w^A\right)\right] \frac{I^A(t)}{I(t)} \frac{I^B(t)}{I(t)} \\ &+ \mathcal{O}\left(\varepsilon^2\right) \end{aligned} \tag{15}$$

The temporal dynamics of $\ln\left(\mathcal{Q}(t)\right)$ is given by:

$$
\begin{aligned}
\frac{\mathrm{d}\ln\left(\mathcal{Q}(t)\right)}{\mathrm{d}t} = &\left(\frac{\Delta\beta^A}{N^A}\left(1-\omega^A-\omega^B\frac{N^B}{N^A}\right)+\left(\Delta\beta^A\frac{\omega^B}{N^A}+\Delta\beta^B\frac{\omega^A}{N^B}\right)\frac{q^B(t)}{q^A(t)}\frac{I^B(t)}{I^A(t)}\right)S^A(t)- \\
&\left(\frac{\Delta\beta^B}{N^B}\left(1-\omega^B-\omega^A\frac{N^A}{N^B}\right)+\left(\Delta\beta^B\frac{\omega^A}{N^B}+\Delta\beta^A\frac{\omega^B}{N^A}\right)\frac{q^A(t)}{q^B(t)}\frac{I^A(t)}{I^B(t)}\right)S^B(t) \\
&-\left(\Delta\gamma^A-\Delta\gamma^B\right) \\
&-\left(\mathcal{Q}(t)-1\right)\left(\beta_w^A\frac{\omega^B}{N^A}+\beta_w^B\frac{\omega^A}{N^B}\right)\left(\frac{q^B(t)}{q^A(t)}\frac{I^B(t)}{I^A(t)}S^A(t)+\frac{1-q^A(t)}{1-q^B(t)}\frac{I^A(t)}{I^B(t)}S^B(t)\right) \\
&+\mathcal{O}\left(\varepsilon^2\right). \qquad (16)
\end{aligned}
$$

For the sake of simplicity, let us consider the scenario where the variant and the wildtype only differ in terms of recovery rates, that is $\Delta\beta^A = \Delta\beta^B = 0$. Under the assumption that phenotypic differences are even smaller than migration probabilities – i.e., migration is faster than selection –, we expect the dynamics of the differentiation $\mathcal{Q}$ to rapidly reach a quasi-equilibrium value. Setting the right-hand side of (16), we obtain:

$$
\mathcal{Q}(t) \approx 1 - \frac{\Delta\gamma^A - \Delta\gamma^B}{\left(\beta_w^A\frac{\omega^B}{N^A}+\beta_w^B\frac{\omega^A}{N^B}\right)\left(\frac{q^B(t)}{q^A(t)}\frac{I^B(t)}{I^A(t)}S^A(t)+\frac{1-q^A(t)}{1-q^B(t)}\frac{I^A(t)}{I^B(t)}S^B(t)\right)}
$$

Migration spatially homogenizes the frequencies of the variant ($\mathcal{Q}(t) = 1$), but these quantities are disrupted by heterogeneous selection (**Fig. 5**). When differentiation reaches a quasi-equilibrium and remains constant, the variant logit-frequency changes with the same slope in each population. This is surprising since selection is heterogeneous, but is due to migration that rapidly homogenizes the variant frequency between the two populations.

Let us now consider the case where the variant is adapted in population A but maladapted in population B. We use parameter values: $\beta_w^A = \beta_m^A = \beta_w^B = \beta_m^B = 0.25$, $\gamma_w^A = \gamma_w^B = 0.1$; $\gamma_m^A = 0.05$, $\gamma_m^B = 0.15$ and $\zeta = 0.01$. Therefore, the phenotypic differences are $\Delta\gamma^A = -0.05$ and $\Delta\gamma^B = +0.05$ along with no transmission difference ($\Delta\beta^A = \Delta\beta^B = 0$). First, migration can reverse the direction of selection, that is the variant can increase in frequency in the population where it is maladapted just because of host movements between differentiated populations (**Fig. 6**). Second, in the long term, local polymorphism maintenance becomes possible when hosts commute between the two population; polymorphism is sometimes however not maintained when migration is stronger and the total population densities very asymmetric (**Fig. 7**). Interestingly, the differentiation is little affected by population size asymmetry.

4 Discussion

Estimating the selective advantage of emerging variants is essential for evaluating epidemic risk and optimizing control strategies. In closed populations, the selective advantage of pathogens depends on the phenotypic differences between competing strains and is also shaped by the environment – in particular, the availability of susceptible hosts and the control measures implemented to mitigate the spread of the epidemic – (Benhamou et al., 2023; Day et al., 2020; Otto et al., 2021). In metapopulations, populations are interconnected through movements of hosts ("migration"), which can affect the phenotypic evolution of pathogens. In this study, we investigate the interplay between such spatial heterogeneity and the evolutionary dynamics across time (variant logit-frequency) and space (differentiation between two populations). For this purpose, we consider a two-patch host metapopulation with commuting of susceptible and infected hosts. Migration is a force that drives the spatial homogenization of the variant. The direction and the magnitude of migration effects depends notably on the differentiation of the variant between the two populations. How can such differentiation be induced? When selection is homogeneous – i.e., the phenotype of each strain does not depend on the population in which the strain is found – differentiation is only transient but can arise due to differences in initial conditions (i.e., the variant is introduced at different time in each populations) or in the environment (i.e., differential availability of susceptible host). Heterogeneous selection, on the other hand, can disrupts the spatial homogenization of the variant, maintains in the long term the differentiation between the two populations, and lead to polymorphism maintenance. These effects depend on the amount of migration, with increased migration eroding local adaptation (Blanquart et al., 2013).

The selective advantage of a novel variant is typically estimated during a sweep by fitting a linear regression on times series of the variant logit-frequency (Boyle et al., 2022; Otto et al., 2021; Volz, 2023). In simple cases, the value of the selective advantage is given by the slope of the linear regression. The estimation of different slopes in different locations can reflect real differences in terms of selection – e.g., unequal availability of susceptible hosts, stringency of non-pharmaceutical interventions, vaccination coverage – but can also be shaped by migration processes. It is therefore challenging to estimate the selective advantage of emerging variants when hosts move between different areas. Here, we emphasize that such migration can blur the strength of selection and can be misleading about the potential real advantage of novel variants, in particular when populations are highly differentiated. Therefore, estimation could be biased when migration is neglected, at the scale of a one population or at the scale of the metapopulation.

The evolutionary dynamics (frequency and differentiation) are conditional on the environment and it is therefore essential to understand the demography to quantify the evolution, especially for transient dynamics. Dynamical mechanistic models are particularly useful to investigate different scenarios in order to understand the interplay between demographic and evolutionary processes. To deepen how neglecting migration can bias the estimation of selective advantages, we can conduct a simulation study and analyze the discrepancy between estimates obtained with a model that neglects migration and a model that takes it into account. This study is a work in progress. We also should further investigate how theoretical approximations, using for instance a separation of timescale between migration and evolutionary processes, can provide useful approximations and help the estimation process with real times series data.

Table 1: **Notations.** The subscript $k \in \{w,\ m\}$ refers to the strain of the pathogen: wildtype strain w or the mutant strain m (or variant). The superscripts i and j ($(i,\ j) \in \{A,\ B\}^2$) refer to populations A or B.

Term	**Definition**
N^i	Population i (constant population size)
S^i	Susceptible host in population i
I^i	Infected/infectious host in population i
q^i	Frequency of the variant m among I^i hosts
β_k^i	*Per capita* transmission rate
γ_k^i	*Per capita* infection clearance/recovery rate
ζ	Rate of immunity waning
ω^i	Probability for hosts from population i to be transiently in population j (commuting)
$\Delta\beta^i,\ \Delta\gamma^i$	Phenotypic differences between the variant and the wildtype; $\Delta\beta^i = \beta_m^i - \beta_w^i$ and $\Delta\gamma^i = \gamma_m^i - \gamma_w^i$
Δ	Effect of the phenotypic differences on the dynamics of the variant logit-frequency
Ω	Effect of migration independently of any phenotypic difference on the dynamics of the variant logit-frequency

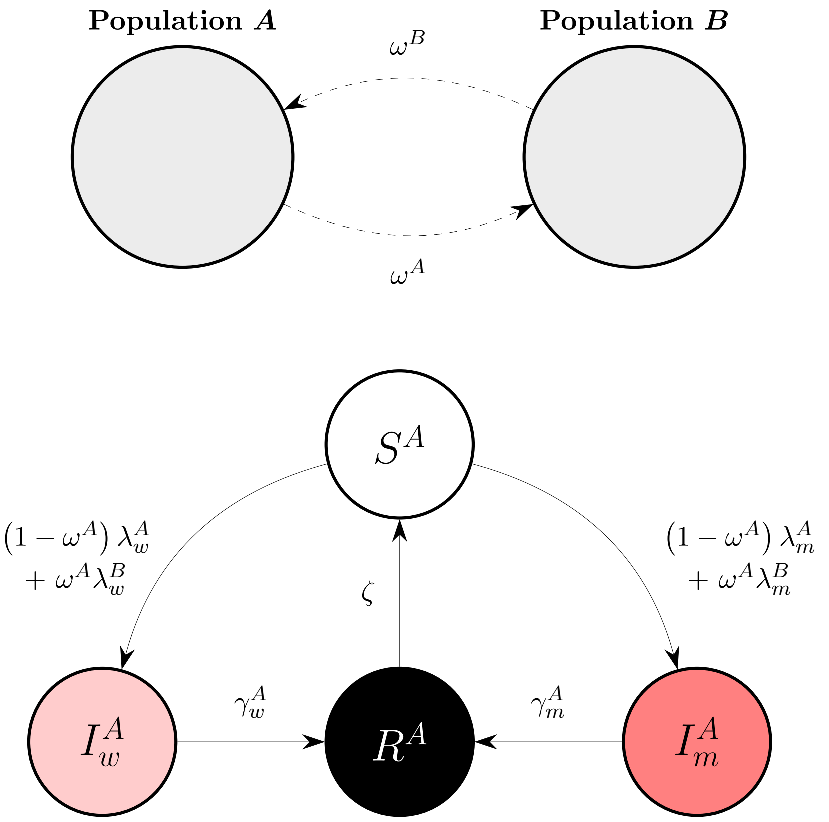


Figure 1: **Flow chart of the epidemic compartment model**. Top panel: we model a two-patch host metapopulation; commuters from population A (resp. B) may visit population B (resp. A) with probability ω^A (resp. ω^B) or remain in the home population with complementary probabilities. Bottom panel: local SIRS model in population A. The subscript w denotes the wildtype strain while the subscript m denotes the mutant strain (or variant); λ_w^A and λ_m^A (resp. λ_w^B and λ_m^B) refer to the forces of infection by the wildtype and the mutant strain, respectively, experienced by susceptible hosts when they are in population A (resp. B).

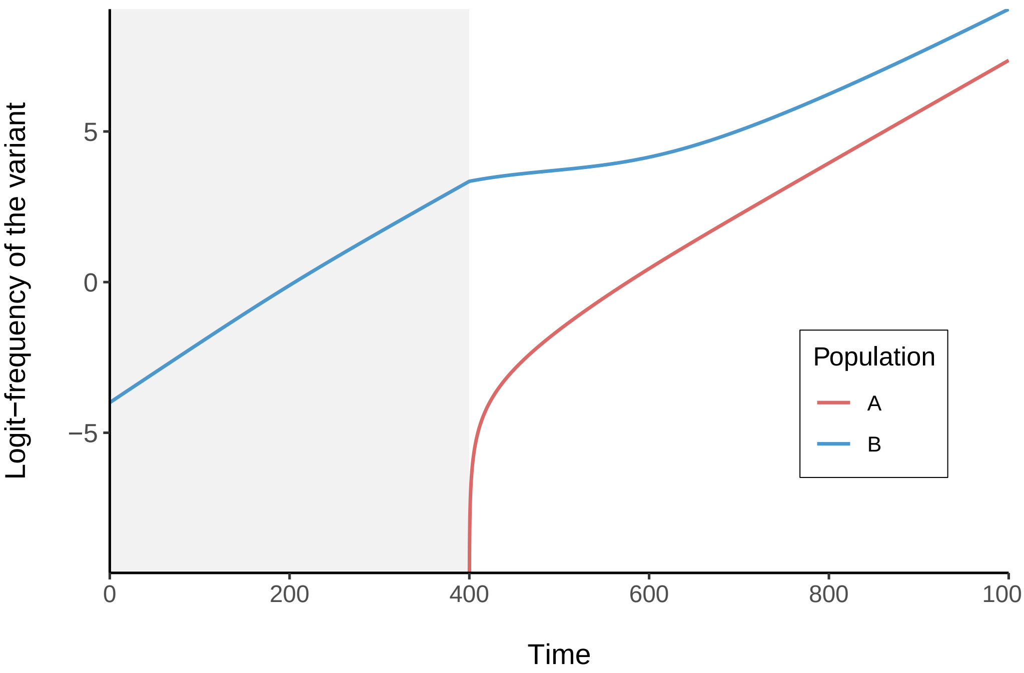


Figure 2: **Dynamics of the logit-frequency of a novel variant in a two-patch metapopulation**. The host metapopulation is divided in two populations, A and B. We use model (3) with parameter values: $\beta_w^A = \beta_w^B = 0.25$, $\beta_m^A = \beta_m^B = 0.3$, $\gamma_w^A = \gamma_w^B = \gamma_m^A = \gamma_m^B = 0.1$ and $\zeta = 0.01$. Note that there is no differential selection between the two populations and that the variant only differs from the wildtype in terms of transmission. The total densities of populations A and B are the same ($N^A = N^B = 1000$). At $t = 0$, both populations are at the endemic equilibrium of the wildtype strain and the variant is introduced at very low density ($I_m^B = 1$) in population B. From $t = 0$ to $t = 400$ (grey background), there is no migration between the two populations ($\omega_A = \omega_B = 0$), then, at $t = 400$, host from population A can visit hosts from population B ($\omega_A = 5 \times 10^{-3}$ and $\omega_B = 0$) where the variant is near fixation.

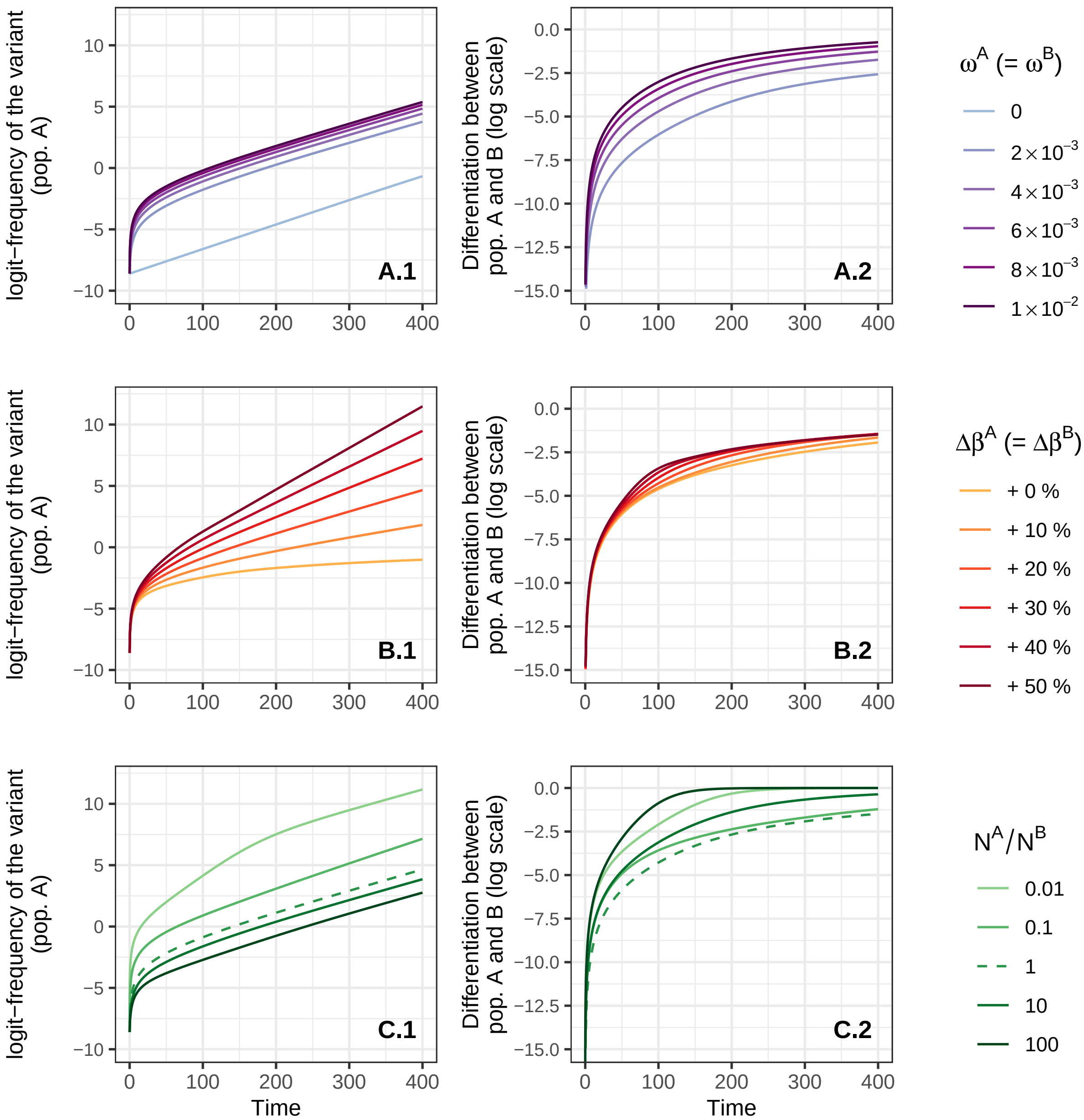


Figure 3: **Evolutionary dynamics over time and space (homogeneous selection)**. The host metapopulation is divided in two populations, A and B. We simulate model (3) in the case of homogeneous selection with default parameter values: $\beta_w = 0.25$, $\beta_m = 0.3$, $\gamma_w = \gamma_m = 0.1$, $\zeta = 0.01$ and $\omega^A = \omega^B = 5 \times 10^{-3}$. Default total densities of population A and B are $N^A = N^B = 1000$. We start from an endemic setting: at $t = 0$ population A is at the epidemiological attractor of the wildtype without migration and population B is at the epidemiological attractor of the variant without migration; we still introduced at $t = 0$ the variant at very low density in population A ($I_m^A(0) = 10^{-2}$). We track the logit-frequency of the variant in population A q^A (1, left column) and the differentiation $\mathcal{Q}$ between population A and B (2, right column) after varying (A) the probabilities of migration ω^A and ω^B simultaneously, (B) the phenotypic differences in transmission $\Delta\beta$ simultaneously (in all cases $\beta_w^A = \beta_w^B = 0.25$) and (C) the ratio of total population densities N^A/N^B (in all cases $N^A = 1000$).

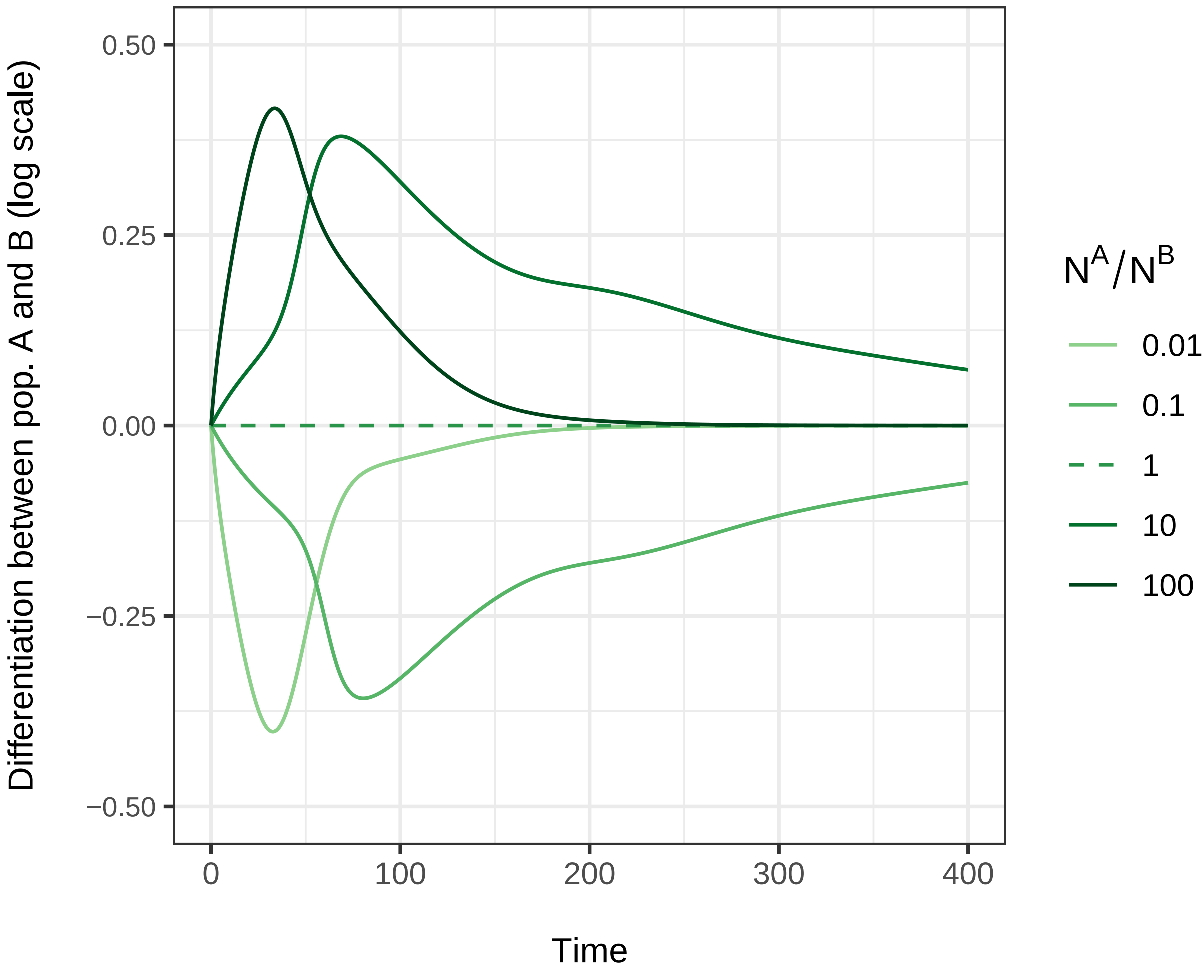


Figure 4: **Transient differentiation of the variant between two interconnected populations**. The host metapopulation is divided in two populations, A and B. We simulate model (3) (see **Fig. 3** for parameter values). The pathogen is introduced at very low densities and equal initial frequency in both populations ($I_w^A(0) = I_m^A(0) = I_w^B(0) = I_m^B(0) = 10^{-3}$), so that, initially, there is no differentiation between population A and B. We look at the effect of the asymmetry of the total population densities on the transient differentiation of the variant between population A and B (in all cases $N^A = 1000$).

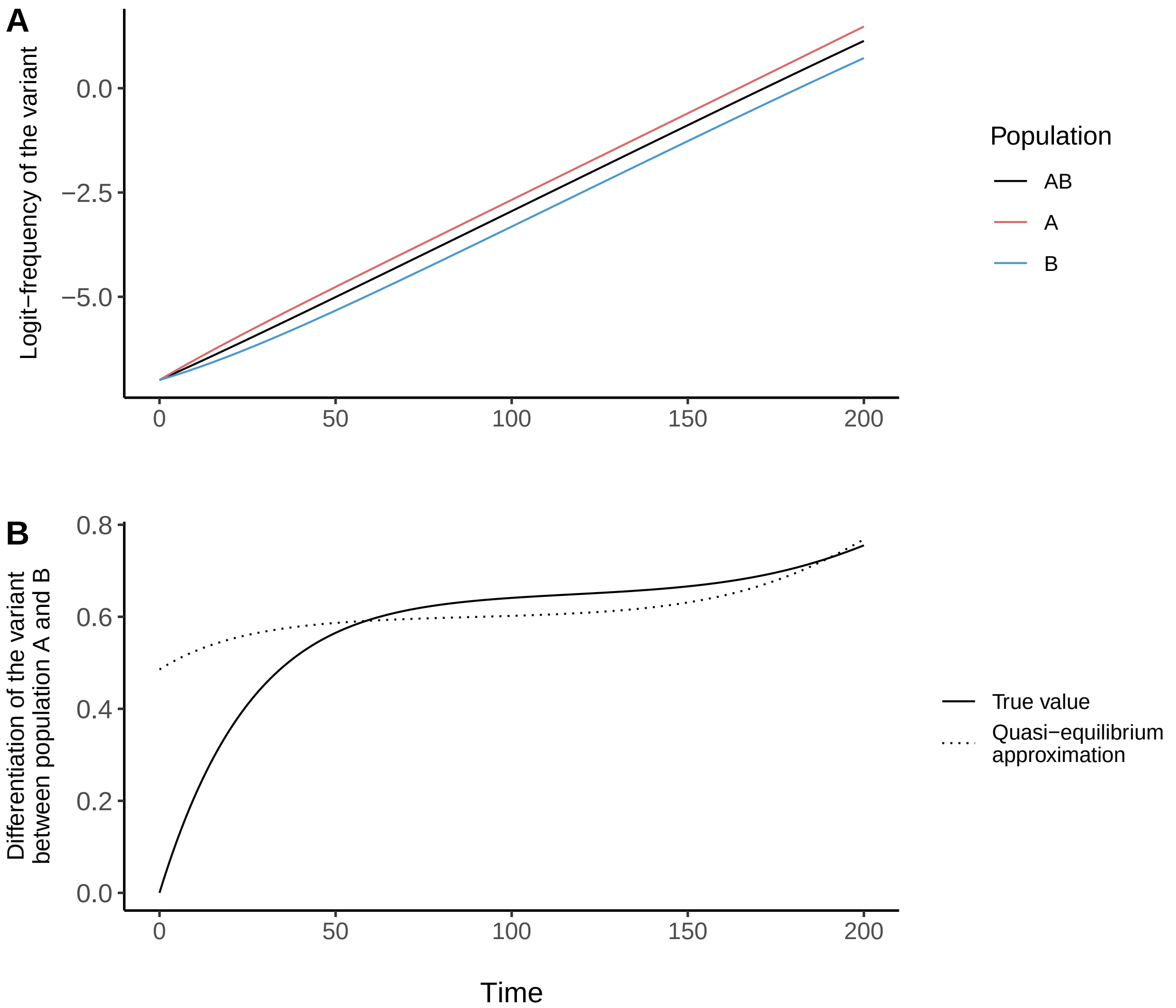


Figure 5: **Separation of timescale between migration and selection**. Temporal dynamics of (A) the logit-frequency of the variant and (B) the differentiation of the variant between population A and B. We use the SIRS model (3) with parameter values: $\beta_w^A = \beta_m^A = \beta_w^B = \beta_m^B = 0.25$, $\gamma_w^A = \gamma_w^B = 0.1$, $\gamma_m^A = 0.05$, $\gamma_m^B = 0.075$, $\zeta = 0.01$, $\omega^A = \omega^B = 0.1$. Therefore, the variant only differs from the wildtype in terms of recovery ($\Delta\beta^A = \Delta\beta^B = 0$) and selection is heterogeneous between the two populations ($\Delta\gamma^A = -0.05$ and $\Delta\gamma^B = -0.025$). Populations A and B have identical total density ($N^A = N^B = 10000$). At $t = 0$, both population are at the epidemiological attractor of the wildtype and we introduce a small density of the pathogen in both population with equal frequency between the wildtype and the variant ($I_w^A(t) = I_m^A(t) = I_w^B(t) = I_m^B(t) = 0.5$), so that initially there is no differentiation between the two populations. Under the assumption that migration is faster than selection, we expect the differentiation of the variant to rapidly reaches a quasi-equilibrium value.

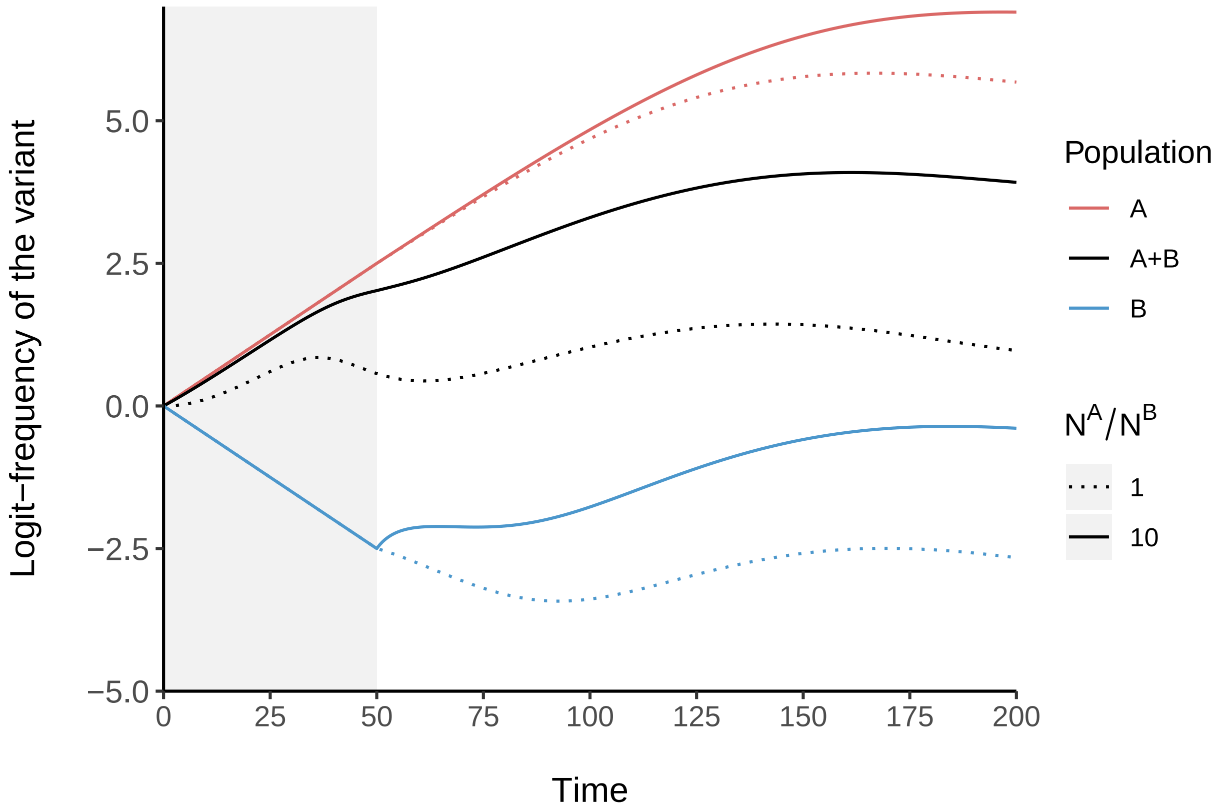


Figure 6: **Dynamics of the logit-frequency of a novel variant in a two-patch metapopulation with heterogeneous selection**. We use model (3) with parameter values: $\beta_w^A = \beta_m^A = \beta_w^B = \beta_m^B = 0.25$, $\gamma_w^A = \gamma_w^B = 0.1$, $\gamma_m^A = 0.05$, $\gamma_m^B = 0.15$ and $\zeta = 0.01$. Therefore, the variant only differs from the wildtype in terms of recovery ($\Delta\beta^A = \Delta\beta^B = 0$) and is adapted in population A ($\Delta\gamma^A = -0.05$) but maladapted in population B ($\Delta\gamma^B = +0.05$). At $t = 0$, we introduce the wildtype and the variant in both population with equal frequency (for each population, infected hosts initially represent 0.1% of the total population density). From $t = 0$ to $t = 50$ (grey background), there is no migration between population A and B ($\omega^A = \omega^B = 0$). From $t = 50$ (white background), the two populations are interconnected through host commuting ($\omega^A = \omega^B = 5 \times 10^{-3}$). Total population densities are either equal (dotted lines, $N^A = N^B = 10000$) or asymmetric (solid lines, $N^A = 10000$ and $N^B = 1000$).

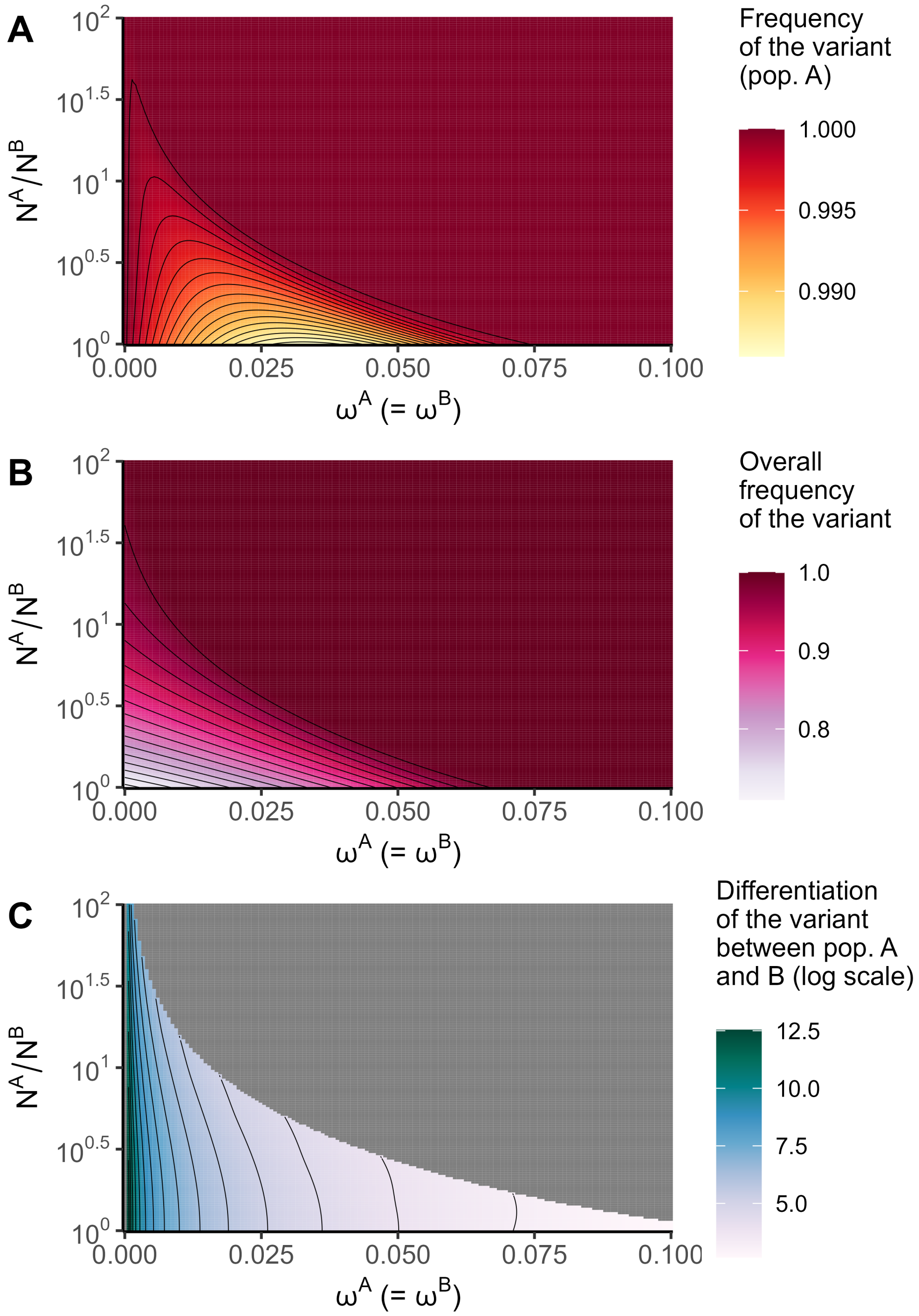


Figure 7: **Polymorphism maintenance for pathogen in a two-patch host metapopulation.** We use model (3) with parameter values: $\beta_w^A = \beta_m^A = \beta_w^B = \beta_m^B = 0.25$, $\gamma_w^A = \gamma_w^B = 0.1$, $\gamma_m^A = 0.05$, $\gamma_m^B = 0.15$ and $\zeta = 0.01$. Therefore, the variant only differs from the wildtype in terms of recovery ($\Delta\beta^A = \Delta\beta^B = 0$) and is adapted in population A ($\Delta\gamma^A = -0.05$) but maladapted in population B ($\Delta\gamma^B = +0.05$). At $t = 0$, population A is at the epidemiological attractor of the variant and population B at the epidemiological attractor of the wildtype. We run simulations for different strengths of migration, varying ω^A and ω^B simultaneously, and for different asymmetries between the total population densities, varying N^A/N^B (in all cases $N^A = 1000$). At final time $t = 4000$, we compute (A) the final frequency q^A of the variant in population A, (B) the final overall frequency q of the variant (in the metapopulation) and (C) the final differentiation $\mathcal{Q}$ of the variant between population A and B. The grey color in C reflects infinite log-differentiation.

References

Benhamou, W., Lion, S., Choquet, R., & Gandon, S. (2023). Phenotypic evolution of SARS-CoV-2: A statistical inference approach. *Evolution*, *77*(10), 2213–2223. https://doi.org/10.1093/evolut/qpad133

Berngruber, T. W., Lion, S., & Gandon, S. (2015). Spatial structure, transmission modes and the evolution of viral exploitation strategies. *PLoS pathogens*, *11*(4), e1004810. https://doi.org/10.1371/journal.ppat.1004810

Blanquart, F., Kaltz, O., Nuismer, S. L., & Gandon, S. (2013). A practical guide to measuring local adaptation. *Ecology letters*, *16*(9), 1195–1205. https://doi.org/10.1111/ele.12150

Boots, M., & Sasaki, A. (1999). 'small worlds' and the evolution of virulence: Infection occurs locally and at a distance. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*(1432), 1933–1938. https://doi.org/10.1098/rspb.1999.0869

Boyle, L., Hletko, S., Huang, J., Lee, J., Pallod, G., Tung, H.-R., & Durrett, R. (2022). Selective sweeps in SARS-CoV-2 variant competition. *Proceedings of the National Academy of Sciences*, *119*(47), e2213879119. https://doi.org/10.1073/pnas.2213879119

Brockmann, D., & Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *science*, *342*(6164), 1337–1342. https://doi.org/10.1126/science.1245200

Butera, Y., Mukantwari, E., Artesi, M., Umuringa, J. d., O'Toole, Á. N., Hill, V., Rooke, S., Hong, S. L., Dellicour, S., Majyambere, O., et al. (2021). Genomic sequencing of SARS-CoV-2 in Rwanda reveals the importance of incoming travelers on lineage diversity. *Nature communications*, *12*(1), 5705. https://doi.org/10.1038/s41467-021-25985-7

Day, T., & Gandon, S. (2006). Insights from price's equation into evolutionary epidemiology. *Disease evolution: models, concepts, and data analyses*, *71*, 23–44. https://doi.org/10.1090/dimacs/071/02

Day, T., & Gandon, S. (2007). Applying population-genetic models in theoretical evolutionary epidemiology. *Ecology Letters*, *10*(10), 876–888. https://doi.org/10.1111/j.1461-0248.2007.01091.x

Day, T., Gandon, S., Lion, S., & Otto, S. P. (2020). On the evolutionary epidemiology of SARS-CoV-2. *Current Biology*, *30*(15), R849–R857. https://doi.org/10.5683/SP2/VKH3LE

Débarre, F., Ronce, O., Gandon, & S. (2013). Quantifying the effects of migration and mutation on adaptation and demography in spatially heterogeneous environments. *Journal of Evolutionary Biology*, *26*(6), 1185–1202. https://doi.org/10.1111/jeb.12132

Débarre, F., Bonhoeffer, S., & Regoes, R. R. (2007). The effect of population structure on the emergence of drug resistance during influenza pandemics. *Journal of the Royal Society Interface*, *4*(16), 893–906. https://doi.org/10.1098/rsif.2007.1126

Débarre, F., Lion, S., Van Baalen, M., & Gandon, S. (2012). Evolution of host life-history traits in a spatially structured host-parasite system. *The American Naturalist*, *179*(1), 52–63. https://doi.org/10.1086/663199

Engbert, R., & Drepper, F. (1994). Chance and chaos in population biology—models of recurrent epidemics and food chain dynamics. *Chaos, Solitons & Fractals*, *4*(7), 1147–1169. https://doi.org/10.1016/0960-0779(94)90028-0

Erlander, S., & Stewart, N. F. (1990). *The gravity model in transportation analysis: Theory and extensions* (Vol. 3). Vsp.

Grenfell, B., & Harwood, J. (1997). (meta) population dynamics of infectious diseases. *Trends in ecology & evolution*, *12*(10), 395–399. https://doi.org/10.1016/S0169-5347(97)01174-9

Hethcote, H. W. (1978). An immunization model for a heterogeneous population. *Theoretical population biology*, *14*(3), 338–349. https://doi.org/10.1016/0040-5809(78)90011-4

Hufnagel, L., Brockmann, D., & Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *Proceedings of the national academy of sciences*, *101*(42), 15124–15129. https://doi.org/10.1073/pnas.0308344101

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, *115*(772), 700–721. https://doi.org/10.1098/rspa.1927.0118

Kraemer, M. U., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., McCrone, J. T., Baele, G., Parag, K. V., Battle, A. L., Gutierrez, B., et al. (2021). Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science*, *373*(6557), 889–895. https://doi.org/10.1126/science.abj0113

Lajmanovich, A., & Yorke, J. A. (1976). A deterministic model for gonorrhea in a nonhomogeneous population. *Mathematical Biosciences*, *28*(3-4), 221–236. https://doi.org/10.1016/0025-5564(76)90125-5

Le Treut, G., Huber, G., Kamb, M., Kawagoe, K., McGeever, A., Miller, J., Pnini, R., Veytsman, B., & Yllanes, D. (2022). A high-resolution flux-matrix model describes the spread of diseases in a spatial network and the effect of mitigation strategies. *Scientific Reports*, *12*(1), 15946. https://doi.org/10.1038/s41598-022-19931-w

Lemey, P., Hong, S. L., Hill, V., Baele, G., Poletto, C., Colizza, V., O'toole, Á., McCrone, J. T., Andersen, K. G., Worobey, M., et al. (2020). Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nature communications*, *11*(1), 5110. https://doi.org/10.1038/s41467-020-18877-9

Lipshtat, A., Alimi, R., & Ben-Horin, Y. (2021). Commuting in metapopulation epidemic modeling. *Scientific reports*, *11*(1), 15198. https://doi.org/10.1038/s41598-021-94672-w

Lloyd, A. L., & May, R. M. (1996). Spatial heterogeneity in epidemic models. *Journal of theoretical biology*, *179*(1), 1–11. https://doi.org/10.1006/jtbi.1996.0042

Murray, J. D. (2003). *Mathematical biology: II: Spatial models and biomedical applications* (Vol. 18). Springer.

Otto, S. P., Day, T., Arino, J., Colijn, C., Dushoff, J., Li, M., Mechai, S., Van Domselaar, G., Wu, J., Earn, D. J., et al. (2021). The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Current Biology*, *31*(14), R918–R929. https://doi.org/10.1016/j.cub.2021.06.049

Post, W., DeAngelis, D., & Travis, C. (1983). Endemic disease in environments with spatially heterogeneous host populations. *Mathematical Biosciences*, *63*(2), 289–302. https://doi.org/10.1016/0025-5564(82)90044-X

Postnikov, E. B., & Sokolov, I. M. (2007). Continuum description of a contact infection spread in a SIR model. *Mathematical biosciences*, *208*(1), 205–215. https://doi.org/10.1016/j.mbs.2006.10.004

Ronce, O., & Kirkpatrick, M. (2001). When sources become sinks: Migrational meltdown in heterogeneous habitats. *Evolution*, *55*(8), 1520–1531. https://doi.org/10.1111/j.0014-3820.2001.tb00672.x

Roques, L., Bonnefon, O., Baudrot, V., Soubeyrand, S., & Berestycki, H. (2020). A parsimonious approach for spatial transmission and heterogeneity in the COVID-19 propagation. *Royal Society Open Science*, *7*(12), 201382. https://doi.org/10.1098/rsos.201382

Volz, E. (2023). Fitness, growth and transmissibility of SARS-CoV-2 genetic variants. *Nature Reviews Genetics*, *24*(10), 724–734. https://doi.org/10.1038/s41576-023-00610-z

Walter, A., Gandon, S., & Lion, S. (2024). Effect of unequal vaccination coverage and migration on long-term pathogen evolution in a metapopulation. *Journal of Evolutionary Biology*, *37*(2), 189–200. https://doi.org/10.1093/jeb/voad016

Xia, Y., Bjørnstad, O. N., & Grenfell, B. T. (2004). Measles metapopulation dynamics: A gravity model for epidemiological coupling and dynamics. *The American Naturalist*, *164*(2), 267–281. https://doi.org/10.1086/422341

Yuksel, M. K., Remien, C. H., Karki, B., Bull, J. J., & Krone, S. M. (2021). Vector dynamics influence spatially imperfect genetic interventions against disease. *Evolution, Medicine, and Public Health*, *9*(1), 1–10. https://doi.org/10.1093/emph/eoaa035

Zurita-Gutiérrez, Y. H., & Lion, S. (2015). Spatial structure, host heterogeneity and parasite virulence: Implications for vaccine-driven evolution. *Ecology letters*, *18*(8), 779–789. https://doi.org/10.1111/ele.12455

Chapter five

General discussion | 5

Infectious diseases are major threats for public health. In modern and contemporary history, plague, flu, cholera, tuberculosis, malaria, HIV/AIDS*, EVD†, SARS or COVID-19 caused together at least tens of millions of deaths [100]. Epidemics in agriculture (crops or livestock) may also be devastating – e.g., the potato crop failures due to infections by late blight (oomycete *Phytophthora infestans*) caused the Great Famine in Ireland in the 1840s. Nowadays, the increase in land-use changes such as deforestation, as well as the consumption of animal products, exacerbates the risk of zoonosis emergence. In the second half of the 20th century, around two-thirds of the emergence of infectious diseases were caused by zoonotic spillovers, especially from wildlife reservoir [101], and the most recent pandemic to date, COVID-19, is another example of emergent zoonosis. Growing connectivity owing to globalization increases the potential for epidemics to spread faster and on larger scales [102], in a context of global health inequalities (e.g., healthcare, vaccination).

[100]: Sampath et al. (2021), 'Pandemics throughout the history'

[101]: Jones et al. (2008), 'Global trends in emerging infectious diseases'

[102]: Hufnagel et al. (2004), 'Forecast and control of epidemics in a globalized world'

Pathogens evolve, sometimes rapidly (high mutations rates, short generation times). The acquisition of adaptive mutations can accelerate the spread of pathogens and erodes our ability to control and mitigate epidemics – e.g. non-pharmaceutical interventions (NPIs), vaccination. Until recently, only demographic data from traditional surveillance of infectious diseases were available and studies focused on epidemiological dynamics. Using genomic sequencing, the question of pathogen adaptation has been investigated previously, during the 2013–2016 EVD epidemic [103, 104], but it was the successive emergence and sweep of VOCs of SARS-CoV-2 during the COVID-19 pandemic that harshly demonstrated the importance of evolution in epidemiology. Nowadays, the advent of sequencing methods revolutionizes this field of research as they allow to collect genetic data and track the spatio-temporal distribution of different variants. Statistical approaches are common in mathematical epidemiology, but coupling evolution and epidemiology has often been limited to theoretical approaches. Yet, it is essential to characterize the selective advantage of emerging variants, which requires the development of novel methods for estimating the phenotypic traits of pathogens.

[103]: Diehl et al. (2016), 'Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic'
[104]: Urbanowicz et al. (2016), 'Human adaptation of Ebola virus during the West African outbreak'

5.1 Summary

This thesis stands at the interface between evolutionary epidemiology theory and statistical analyses of empirical and experimental data. I have a background in biology but I have always been interested in interdisciplinary approaches, especially between biology and mathematics. I am also interested in the analysis of real data and statistics. The present work reflects discussions and collaboration with my PhD supervisors (statistician and theoretical biologist) with complementary expertise. This

* Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome
† Ebola Virus Disease

enables me to adopt an integrative perspective that provides a comprehensive understanding of how the phenotypic evolution of pathogens is shaped by epidemiological feedback and how the combination of information between epidemiological and genetic data can allow us to estimate key phenotypic traits of pathogens. As with all interdisciplinary work, this is a difficult task. Datasets are typically incomplete because of hidden processes and missing data, increasing the complexity of evo-epidemiological models and making even more challenging the process of estimating the model parameters.

Throughout my PhD, my work relied on the analyses of deterministic models based on dynamical systems of ordinary differential equations. I carried out three projects that allowed me to develop new tools to exploit incomplete datasets and extract so far inaccessible information on the dynamics of a pathogen spreading and evolving in heterogeneous environments. These new tools deal with missing data and rely on the explicit incorporation of hidden processes.

In the first research project (Chapter two), I focused on the selective advantage of the Alpha variant of SARS-CoV-2 relative to the previous dominant lineage in England. The underlying phenotypic variation was quantified considering two life-history traits: (i) the transmission rate and (ii) the duration of infectiousness (inverse of the recovery rate). Theoretical models predict that control measures diminishing contact rates and transmission reduce the selective advantage of variants with higher transmission but has little or no effect on variants with longer infectious periods [53, 65]. Based on a deterministic SEIR model, I used the time-varying stringency of NPIs [99] during the sweep of the Alpha variant in England to disentangle and estimate both phenotypic differences. I developed a two-step approach. In the first step, before the emergence of the variant, I estimated how the intensity of NPIs impacted the spread of the virus. I used these estimates in the second step, after the emergence of the variant, where I exploited the slow-fast dynamics of eco-evolutionary processes to complete the inference of the phenotypic advantage of the Alpha variant. I showed that the Alpha variant was more likely to have a higher transmission rate rather than a longer duration of infectiousness.

[53]: Day et al. (2020), 'On the evolutionary epidemiology of SARS-CoV-2'
[65]: Otto et al. (2021), 'The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic'
[99]: Hale et al. (2021), 'A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)'

In the second research project (Chapter three), I studied the competition between two strains of the temperate phage λ throughout experimental epidemics in continuous cultures of *E. coli*. These two strains exhibit distinct life-history strategies: the wildtype is more latent and relies more on vertical transmission (lysogenic cycle), while the variant (λcI857) is more virulent and relies mostly on horizontal transmission (lytic cycle). This dichotomy was assumed to be governed by two life-history traits: (i) the probability of lysogenization (phage integration) and (ii) the rate of prophage reactivation. This work was the direct continuation of [68], which used experimental evolution based on the biology of phage λ to qualitatively validate predictions from evolutionary epidemiology theory about the dynamics of selection on virulence in a broader context. In particular, data tracking both the epidemiology (prevalence) and the evolution of the virus (frequency of the virulent phage among viral particles and among infected bacteria) confirmed the theoretical prediction that variants with higher transmissibility and virulence can be

[68]: Berngruber et al. (2013), 'Evolution of virulence in emerging epidemics'

selected for in emerging epidemics, but counter-selected as soon as the host population reaches high prevalence. I went beyond the qualitative match between theoretical predictions and experimental time series data and used the data to improve the model and estimate the phenotypic traits of both strains. I developed a new inference approach to estimate the viral phenotypes at different stages of the epidemic – including phenotypic traits particularly challenging to estimate otherwise. Based on the knowledge of the biology of the system, I modelled hidden processes such as lysis and lysogeny and fitted this new model to an incomplete dataset.

In the third and last research project (Chapter four), I examined the interplay between migration and pathogen evolution. During the sweep of an emerging variant, selection is quantified by the slope of the variant frequency on the logit scale [65–67]. I used this approach in my first research project for the Alpha variant of SARS-CoV-2, assuming that the nine regions of England were independent closed populations. However, host populations are typically spatially structured and interconnected through movements ("migration") of susceptible and infected individuals. How such spatial heterogeneity affects the evolutionary dynamics of the variant and biases the estimation of its selective advantage are non-trivial. As in the first project (Chapter two), I considered that the variant and the wildtype may differ phenotypically in terms of (i) transmission rate and/or (ii) recovery rate. I considered an SIRS model two-patch host metapopulation and investigated how the commuting of susceptible and infected hosts affect the dynamics of the frequency of the variant across space and time. I showed that migration can blur the effects of selection and lead to misinterpretations about the real selective advantage of variants.

[65]: Otto et al. (2021), 'The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic'
[66]: Boyle et al. (2022), 'Selective sweeps in SARS-CoV-2 variant competition'
[67]: Volz (2023), 'Fitness, growth and transmissibility of SARS-CoV-2 genetic variants'

My work is a complementary approach to other methods currently used. In particular, the rise of phylogenetic approaches has allowed the development of useful and powerful methods that also combine epidemiological and genetic data to analyse the spread and the evolution of pathogens. Phylodynamics (as coined by Bryan Grenfell in 2004 for pathogens [105]) uses data from genomic surveillance (sequencing of collected samples) to reconstruct molecular phylogenies – i.e., evolutionary trees based on the relatedness of sampled genetic sequences – and population dynamics models to provide key insight on the dynamics of pathogens (e.g., estimate population sizes, reproduction numbers, elucidate pathogen transmission chains, identify superspreading events) at different scales [106–109]. These analyses are often based on the assumption that genomic variation is neutral. Phylodynamics can be limited by sampling biases and the small genetic variation observed among the genomes. Besides, the underlying epidemiological models are often simple (e.g., little or no host structure).

[105]: Grenfell et al. (2004), 'Unifying the epidemiological and evolutionary dynamics of pathogens'

[106]: Volz et al. (2013), 'Viral phylodynamics'
[107]: Pybus et al. (2009), 'Evolutionary analysis of the dynamics of viral infectious disease'
[108]: Attwood et al. (2022), 'Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic'
[109]: Alizon (2022), 'Phylodynamique'

5.2 Host structure and differentiation

In the general introduction (Chapter one), the dynamics were simple because the pathogen habitat (i.e., the infected hosts I) was assumed to be unstructured. As soon as the pathogen can be found in different compartments, one potentially needs to account for heterogeneous environments.

In the first project, the infected compartment was divided between the exposed class E – which does not allow the pathogen to be transmitted – and the infectious class I. I was not able to use the differentiation of the variant between classes E and I because no data were available but I found a way to estimate the phenotypic differences using the overall frequency of the variant – i.e., in both E and I. To do so, I accounted for these two classes by weighting the variant frequencies by class frequencies but also by reproductive values – i.e., relative long-term contributions to the future of the population – [78, 110]. This method relies on a separation of timescales argument: under the assumption of weak selection, epidemiological dynamics are fast and rapidly reach quasi-equilibrium values that can be then used to simplify the analysis of the evolutionary dynamics. The validity of this approach can be challenged, as it is not always guaranteed that epidemiology is faster than evolution. In particular, phenotypic differences are assumed to be small, though those are the quantities we want to estimate.

[78]: Lion (2018), 'Class structure, demography, and selection: reproductive-value weighting in nonequilibrium, polymorphic populations'
[110]: Lion et al. (2022), 'Evolution of class-structured populations in periodic environments'

In the second project, the host pathogen habitat was structured between the free virus stage V (the culture medium) and two stages among infected bacteria (lysogens L and cells prior to lysis Y). I did not track the overall dynamics of the variant but I tracked its frequency in each compartment. Theoretical analyses were simplified by neglecting the contribution of the lysogenic pathway at the beginning of the epidemic and the contribution of the lytic pathway at the end of the epidemic. Transitions from compartment V to compartment L or Y (horizontal transmissions) were different between the two strains, for they have different probabilities of lysogenization. Transitions from compartment L to Y (prophage reactivation) were also different between the two strains, for they have different reactivation rates. These phenotypic differences induced a differentiation between the different compartments. In the endemic case, the differentiation was simple and can be used to estimate the rates of prophage reactivations. On the other hand, the analysis of the early stage of the epidemic was challenging because epidemiology and evolution occur on similar timescales, so that they were difficult to decouple.

Eventually, in the third project, hosts were structured spatially in two populations interconnected through migration. Movements of infected hosts were assumed to have identical probabilities, regardless of whether the hosts were infected by the wildtype or the variant. I studied the differentiation of the variant between the two populations in the presence of migration homogeneous or heterogeneous selection.

Modelling and analysing heterogeneous environments is a much more difficult task than dealing with homogeneous environment, but it is often essential to capture properly the dynamics of host-pathogen systems. Crucially, I showed that the differentiation of the variant between compartments yields useful insights on the evolutionary epidemiology of infectious diseases and can be key to estimate model parameters such as phenotypic traits. Some data, unfortunately, do not exist (e.g., stratified between exposed and infectious hosts) and some data could be available but are not, or are very difficult to obtain (e.g., stratified by vaccination status). Accounting for the pathogen structure among different compartments is however essential to understand and to predict the evolutionary epidemiology of infectious diseases. This is ultimately bound to the availability of data stratified by compartment, and efforts should be pursued in this direction.

5.3 All models are wrong

All models rely on numerous assumptions, approximations and simplifications (often arbitrary) which can always be pointed out and challenged [111]. This is particularly true for models tailored to real biological systems and fitted to experimental/empirical data.

[111]: Box (1976), 'Science and statistics'

Simpler models are easier to analyze mathematically and to interpret, while more sophisticated models always improve the goodness of fit compared to simpler, nested models. However, the data may not be sufficiently rich to precisely infer the extra parameters and would lead to identifiability issues. For instance, one of the initial goals of the second project (Chapter three) was to compare the likelihood of an alternative model that accounts for the phenotypic plasticity of the probability of lysogenization – see original work [112]. When I took over this project, I built several models that account for different phenotypic plasticities. However, experimental data did not provide sufficient information to explore this level of detail and I eventually decided not to consider phenotypic plasticity any further. More broadly, at a certain point, seeking a better fit is no longer worthwhile, but it is often delicate to know where to draw the line between model details and parsimony. Combining effectively theoretical and statistical approaches is an iterative process, back and forth between models and data, between goodness of fit and parsimony. There is no general solution to optimize this process and each project should be examined on a case-by-case basis. In the following, I would like to look back at the models I used and highlight some limits.

[112]: Blanquart et al. (2020), 'Evolution of virulence in emerging epidemics: inference from an evolution experiment'

In this thesis, I only relied on deterministic dynamical models based on systems of ODEs. Yet, random fluctuations and inherent uncertainties can play a critical role in the dynamics of real-life systems. For instance, the early stage of an epidemic or the early stage of the sweep of a variant is characterized by a low density of infected hosts. It can thus be important to account for process stochasticity and see how far results deviate from deterministic cases. Stochastic versions of ODE systems can be derived using Gillespie algorithm [113, 114] (or its τ-leap approximation [115]), or using stochastic differential equations (e.g., [116, 117]). As for deterministic dynamical models, stochastic models can be fitted to time series data. For example, the R software package `pomp` (partially observed Markov process) [118] provides powerful tools and algorithms (such as iterated filtering [119]) to fit stochastic dynamical models to time series data and estimate model parameters. Interestingly, such methods have already been used in the context of competing strains [120].

[113]: Gillespie (1976), 'A general method for numerically simulating the stochastic time evolution of coupled chemical reactions'
[114]: Gillespie (1977), 'Exact stochastic simulation of coupled chemical reactions'
[115]: Gillespie (2001), 'Approximate accelerated stochastic simulation of chemically reacting systems'
[116]: Parsons et al. (2018), 'Pathogen evolution in finite populations: slow and steady spreads the best'
[117]: Day et al. (2020), 'The Price equation and evolutionary epidemiology'
[118]: King et al. (2015), 'Statistical inference for partially observed Markov processes via the R package pomp'
[119]: Ionides et al. (2006), 'Inference for nonlinear dynamical systems'
[120]: Bretó et al. (2009), 'Time series analysis via mechanistic models'

So why stick to deterministic models? Stochastic models are more complex, both mathematically and computationally, and deterministic models can provide an accurate approximation of the dynamics of real-life systems, when population sizes are large and/or when environmental conditions are controlled experimentally – as in the second project (Chapter three). Because deterministic models poorly capture the early stage of the sweep of an emerging variant, I discarded in the first project (Chapter two) the data for which the frequency of the Alpha variant was lower than 10%; in the third project (Chapter four), I sometimes had to introduce the variant later in the simulation to mimic the effects of stochasticity on the timing of emergence.

An implicit assumption of simple compartmental models based on ODEs is that sojourn times are exponentially distributed [31, 32]. The exponential distribution is Markovian, or memoryless, so that all the individuals within a compartment have the same probability to leave this compartment, regardless of how long they have already been in that state. Yet, many biological processes depend on how much time has elapsed. In particular, the duration of an infection (sojourn time in compartment I) is shaped by biological factors such as the host immune response, which typically increases over time. More flexible distributions are classically used, such as the Gamma or the Weibull distributions – e.g., [121, 122] used the Euler-Lotka/renewal equation framework popularized by [30] with Gamma-distributed generation interval to investigate the dynamics of SARS-CoV-2 VOCs. Another approach is to rely on systems of partial differential equations (PDEs). PDE systems is a non-Markovian approach in continuous time that incorporates a felxible temporal ("memory") structure that explicitly tracks elapsed time [123]. Kermack and McKendrick used mostly systems of PDEs and presented the SIR model based on ODEs only as a special case with constant rates [9, 124, 125].

[31]: Forien et al. (2021), 'Estimating the state of the COVID-19 epidemic in France using a model with memory'
[32]: Sofonea et al. (2021), 'Memory is key in capturing COVID-19 epidemiological dynamics'

[121]: Blanquart et al. (2022), 'Selection for infectivity profiles in slow and fast epidemics, and the rise of SARS-CoV-2 variants'
[122]: Park et al. (2022), 'The importance of the generation interval in investigating dynamics and control of new SARS-CoV-2 variants'
[30]: Wallinga et al. (2007), 'How generation intervals shape the relationship between growth rates and reproductive numbers'
[123]: Reyné et al. (2022), 'Non-Markovian modelling highlights the importance of age structure on Covid-19 epidemiological dynamics'
[9]: Kermack et al. (1927), 'A contribution to the mathematical theory of epidemics'
[124]: Kermack et al. (1932), 'Contributions to the mathematical theory of epidemics. II.—The problem of endemicity'
[125]: Kermack et al. (1933), 'Contributions to the mathematical theory of epidemics. III.—Further studies of the problem of endemicity'

In the first and second project of my thesis, I decided to use an alternative approach, still based on ODE systems, that consists in stratifying a compartment into n successive stages, so that sojourn time becomes the sum of n independent exponential distributions, i.e., a hypoexponential distribution (generalized Erlang distribution) or a Gamma distribution if the n distributions are i.i.d. (linear/Gamma chain trick). SEIR models in the first project are thus a very simple representation of an age-structured model. In the second project, I used this approach to explore the impact of Gamma-distributed lysis times – for the sake of simplicity, however, I kept exponentially distributed lysis time in the main text. I found that this intermediate approach strikes a balance between simplicity and biological realism.

5.4 Perspectives

5.4.1 New variants and strain structure

Throughout my PhD, I focused on the competition between two strains of pathogens (the wildtype and a single mutant). These analyses rely on *a priori* strain identification and classification. But when is a strain first described? This is a difficult task, especially during the early stage of an emerging variant. Some studies seek to estimate the selective advantage of variants without the need for prior strain classification [126]. Tracking the circulation of multiple strains and the frequencies of alternative alleles at different loci requires to extend the models used in this thesis to a multistrain and multilocus framework [127, 128]. Recombination and epistasis in fitness among multiple loci is expected to alter the epidemiology and evolution of the pathogen population [129, 130].

[126]: Donker et al. (2024), 'Estimation of SARS-CoV-2 fitness gains from genomic surveillance data without prior lineage classification'

[127]: Gupta et al. (1996), 'The maintenance of strain structure in populations of recombining infectious agents'

[128]: Gog et al. (2002), 'Dynamics and selection of many-strain pathogens'

[129]: Day et al. (2012), 'The evolutionary epidemiology of multilocus drug resistance'

[130]: McLeod et al. (2022), 'Effects of epistasis and recombination between vaccine-escape and virulence alleles on the dynamics of pathogen adaptation'

5.4.2 Natural immunity and vaccination

In all the projects of this thesis, immunity was modelled as an all-or-nothing response. However, imperfect immunity, cross-immunity, immune waning or immune escape shape both the epidemiological and evolutionary dynamics of infectious diseases. Adaptive immunity can be acquired through prior infections or through vaccination.

Vaccination is a highly effective way of protecting hosts from infections and to limit the spread of epidemics. Vaccines can for instance reduce host susceptibility, within-host pathogen growth, transmission rate of infected hosts or pathogen virulence [131, 132]. However, pathogens can adapt to vaccination. Vaccines against influenza viruses must for instance be updated annually [133]. For COVID-19, the vaccination campaign started in December 2020 in England and France. During the sweep of the first Omicron variant of SARS-CoV-2 in England, the variant was more frequent among vaccinated hosts than among naive hosts (**Figure 5.1**), revealing a higher ability to infect vaccinated hosts [134]. These data describe a dual structure for the environment of the pathogen. First,

[131]: Gandon et al. (2001), 'Imperfect vaccines and the evolution of pathogen virulence'

[132]: Walter et al. (2021), 'Epidemiological and evolutionary consequences of periodicity in treatment coverage'

[133]: Hannoun (2013), 'The evolving history of influenza viruses and influenza vaccines'

[134]: Paton et al. (2022), 'The rapid replacement of the SARS-CoV-2 Delta variant by Omicron (B.1.1.529) in England'

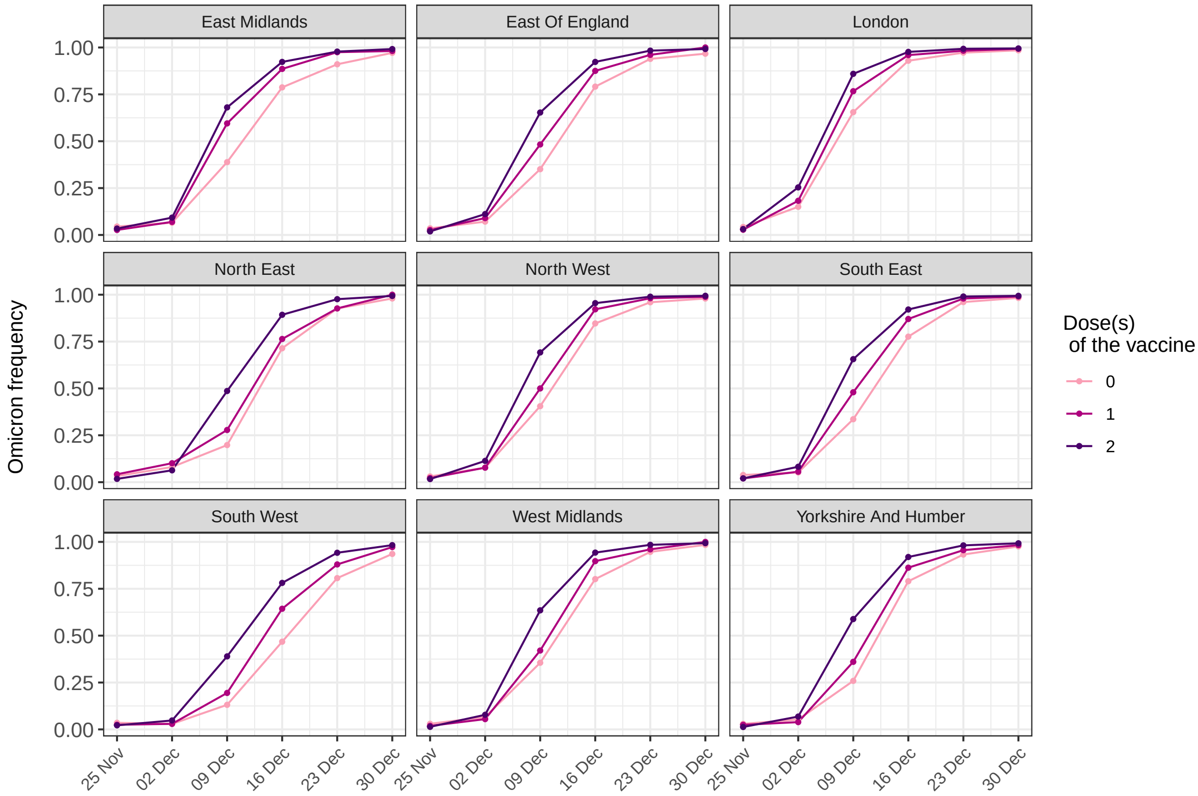


Figure 5.1: Dynamics of the frequencies of SARS-CoV-2 Omicron variant in England according to vaccination status. Regional, weekly data between November 25, 2021 and December 30, 2021 are from the UK Health Security Agency.

the population is stratified regionally (and regions are interconnected with host movements). Second, hosts are stratified between vaccinated and non-vaccinated individuals. Vaccination induces a heterogeneous selection between these two types of hosts.

One of the initial aims of the third project (Chapter four) was to include vaccination to investigate how migration and vaccination shapes the transient evolutionary dynamics of pathogens. Future extension of this work may thus include additional compartments to formally take vaccination into account.

5.4.3 Host coevolution

In this thesis, I focused on the phenotypic evolution of pathogens and I assumed that hosts do not evolve. The evolutionary potential of pathogens is often much higher than their host – metazoan hosts typically have much longer generation times than their pathogens. Yet, this is not always the case – e.g., many unicellular organisms such as bacteria – and the reciprocal selective pressures that hosts and pathogens exert on each other drive the dynamics of coevolution (red queen dynamics). For example, in [68], from which the data for the second project are taken, *E. coli* acquired resistance to phage λ later in the experiment. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) defense systems – which can be considered as an analogue of adaptive immunity in bacteria –, coevolve antagonistically with phage immune escape [135]. Adaptation is accelerated by the presence of plasmids and recombination events. More broadly, bacteria resistance evolution has dramatic impact on public health, and especially in the context of rising (multiple) antibiotic resistance. For this purpose, phage therapy uses lytic phages as a treatment coevolving with bacteria, offering an effective alternative to combat bacterial infections and emphasizing the urgent need to study host-pathogen coevolution.

[68]: Berngruber et al. (2013), 'Evolution of virulence in emerging epidemics'

[135]: Guillemet et al. (2022), 'Competition and coevolution drive the evolution and the diversification of CRISPR immunity'

Synthèse en français

Synthèse en français

Introduction

Fin 2019, moins de deux ans avant le début de cette thèse, le virus du SARS-CoV-2 (Syndrome Respiratoire Aigu Sévère – Coronavirus 2) a emergé dans le marché de Wuhan, dans la province de Hubei, en Chine [2]. La maladie respiratoire transmise par le virus (COVID-19, *Coronavirus disease 2019*) s'est rapidement propagée à l'échelle mondiale. En mars 2020, au moins 114 pays ont été touchés par la maladie qui a été classée comme pandémie par l'Organisation Mondiale de la Santé [3]. Comprendre et prédire la dynamique épidémiologique du COVID-19 a été un défi majeur de santé publique; les épidémiologistes devaient comprendre de toute urgence les dynamiques de sa transmission, évaluer les risques et les conséquences sanitaires potentielles de la pandémie, et concevoir des stratégies pour en freiner la propagation. Dans cet effort collectif, la modélisation mathématique a été un outil déterminant.

[2]: Lu et al. (2020), 'Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle'

[3]: WHO (2020), 'WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020'

Un modèle est une représentation simplifiée de la réalité : les processus sous-jacents sont approximés ou même négligés – pour ceux considérés comme les moins importants. Alors que les approches théoriques ont souvent été sous-estimées par rapport aux approches empiriques [4], les modèles mathématiques sont des outils particulièrement utiles et puissants qui permettent de formaliser, d'analyser, de comprendre et de fournir des prédictions qualitatives ou quantitatives sur les processus biologiques et de guider la prise de décision. On distingue classiquement les modèles pour comprendre et les modèles pour prédire. L'épidémiologie des maladies infectieuses est étroitement liée à la modélisation mathématique et ce depuis très longtemps. Dans la seconde moitié du XVIIIe siècle, Bernoulli a développé un modèle épidémiologique pour analyser les données de morbidité et de mortalité de la variole et a étudié le bénéfice de l'inoculation du pathogène [5]. À la fin du XIXe – début du XXe siècle, Ronald Ross (Prix Nobel 1902 de physiologie ou médecine) a proposé les premiers modèles mathématiques de transmission du paludisme; en particulier, il a démontré que le paludisme était transmis par les piqûres de moustiques anophèles infectés (maladie vectorielle) et a proposé des stratégies de contrôle de la maladie [6–8]. En 1927, Kermack et McKendrick ont publié un article qui a popularisé l'utilisation des modèles compartimentaux déterministes pour simuler les dynamiques épidémiologiques [9]. Ces modèles compartimentaux ont depuis été largement utilisés, notamment pour la pandémie de COVID-19. Les modèles mathématiques peuvent de plus être ajustés aux données (par exemple, le nombre de cas positifs ou de décès déclarés) pour estimer des paramètres clés d'une épidémie, comme par exemple le nombre moyen d'infections secondaires - estimé autour de 2,9 (intervalle de confiance à 95 % : 2,81 à 3,01) au début de l'épidémie de COVID-19 en France [10]. Les modèles mathématiques ont également été utilisés pour prévoir la dynamique future du COVID-19 dans différents scénarios – différentes stratégies de contrôle par exemple –, notamment pour anticiper la saturation des hôpitaux [11, 12].

[4]: Goldstein (2018), 'Are theoretical results 'Results'?'

[5]: Bernoulli (1760), 'An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it'

[6]: Ross (1899), 'Inaugural lecture on the possibility of extirpating malaria from certain localities by a new method'
[7]: Ross (1905), 'The logical basis of the sanitary policy of mosquito reduction'
[8]: Ross (1911), *The prevention of malaria*

[9]: Kermack et al. (1927), 'A contribution to the mathematical theory of epidemics'

[10]: Salje et al. (2020), 'Estimating the burden of SARS-CoV-2 in France'
[11]: Ferguson et al. (2020), 'Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand'
[12]: Paireau et al. (2022), 'An ensemble model based on early predictors to forecast COVID-19 health care demand in France'

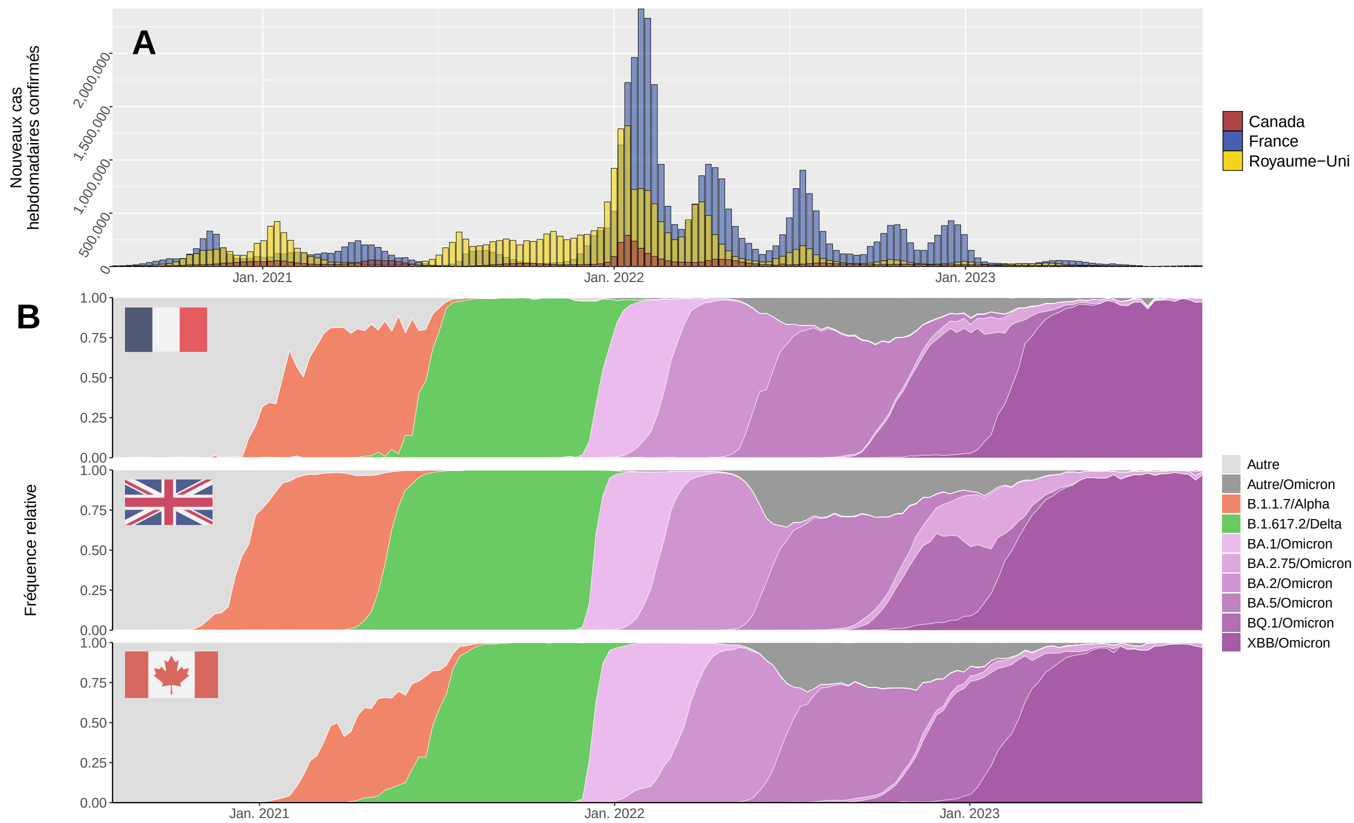


Figure 5.2: Épidémiologie et évolution du SARS-CoV-2 en France, au Royaume-Uni et au Canada. J'utilise des données publiques disponibles du 2020-08-01 au 2023-09-01. (A) Épidémiologie : nouveaux cas confirmés de COVID-19 par semaine (données de l'OMS, téléchargées depuis *Our World in Data*); (B) Évolution : fréquences relatives de plusieurs variants préoccupants (source des métadonnées : 4 020 732 séquences disponibles sur GISAID).

L'adaptation du SARS-CoV-2 n'a reçu que très peu d'attention au début de la pandémie [13, 14]. Les dynamiques évolutives sont généralement supposées se dérouler sur des échelles de temps beaucoup plus lentes que les dynamiques épidémiologiques. Toutefois, la mutation D614G (substitution de la protéine *Spike*, probablement associée à une transmission accrue) est apparue au début de la pandémie de COVID-19 (vers mai 2020) et est devenue la souche dominante [16, 17]. Les mutations identifiées étaient cependant d'abord considérées comme neutres ou faiblement délétères, l'augmentation en fréquence des mutations comme D614G pouvant potentiellement être attribuée à de la stochasticité démographique [18]. Plus tard en 2020, le variant Alpha du SARS-CoV-2 (lignée Pango B.1.1.7) a émergé en Angleterre [19, 20]. Sa propagation spectaculaire dans chaque pays où il a été introduit a brutalement mis l'évolution au devant de la scène. L'adaptation du virus est devenu un élément majeur de préoccupation dans la gestion de la pandémie de COVID-19, lançant des débats sur son évolution à court, moyen et long terme. Le variant Alpha a été classé comme variant préoccupant – c'est-à-dire possédant un avantage sélectif – et a été le premier d'une série de variants qui ont successivement émergé puis remplacé la lignée précédente – par exemple, Delta (lignée Pango B.1.617.2), ou Omicron (première lignée Pango B.1.1.529) - (**Figure 5.2** , à partir des données GISAID [21]). L'adaptation des pathogènes peut altérer notre capacité à contrôler les épidémies. La pandémie de COVID-19 illustre donc combien il est crucial de caractériser les phénotypes des pathogènes, de suivre leur

[13]: Grubaugh et al. (2020), 'We shouldn't worry when a virus mutates during disease outbreaks'
[14]: Rausch et al. (2020), 'Low genetic diversity may be an Achilles heel of SARS-CoV-2'
[16]: Plante et al. (2021), 'Spike mutation D614G alters SARS-CoV-2 fitness'
[17]: Volz et al. (2021), 'Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity'
[18]: Grubaugh et al. (2020), 'Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear'
[19]: Public Health England (2020), *Investigation of novel SARS-COV-2 variant 202012/01: technical briefing 5*
[20]: Volz et al. (2021), 'Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England'

[21]: Khare et al. (2021), 'GISAID's role in pandemic response'

évolution phénotypique et de comprendre les déterminants de cette évolution. Pourtant, alors que les approches statistiques sont très courantes en épidémiologie, coupler évolution et épidémiologie a le plus souvent été limité à des approches théoriques.

Dans cette thèse, je combine une approche théorique basée sur des modèles dynamiques mécanistiques et une approche statistique (**Figure 5.3**) pour estimer les paramètres de ces modèles – comme les traits phénotypiques des pathogènes – en santé publique et en microbiologie expérimentale. Données et observations peuvent être utilisées qualitativement pour adapter un modèle dynamique à la biologie d'un système hôte-pathogène en particulier. L'analyse théorique et les simulations numériques de ces modèles peuvent permettre de fournir des clés pour comprendre et prédire l'épidémiologie évolutive des maladies infectieuses. La confrontation entre données (démographiques et génétiques) et les résultats d'un modèle peut valider qualitativement des prédictions théoriques. De façon plus quantitative, les modèles statistiques construits à partir des modèles dynamiques peuvent être ajustés à des séries temporelles afin d'estimer les paramètres des modèles ou de comparer plusieurs modèles. Combiner efficacement les approches théoriques et statistiques est un processus itératif, un aller-retour entre modèles et données.

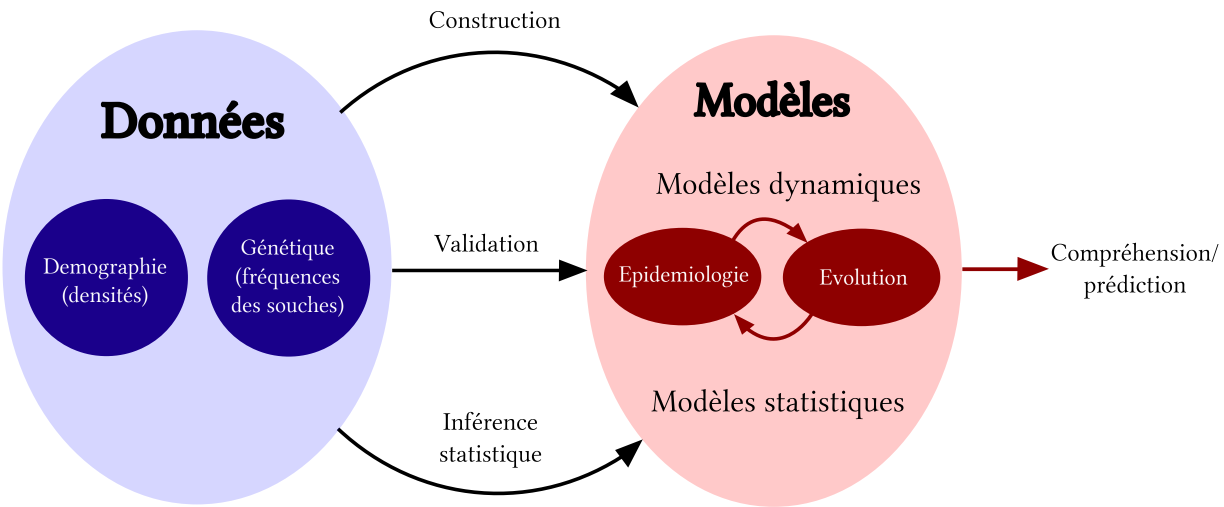


Figure 5.3: Relations schématisées entre modèles et données en épidémiologie évolutive. Les modèles mathématiques en épidémiologie évolutive sont des représentations simplifiées et formalisées du couplage entre processus épidémiologiques et évolutifs, permettant notamment de prendre en compte leurs rétroactions. Par des analyses théoriques et des simulations numériques, les modèles dynamiques/mécanistiques sont très utiles pour comprendre et prédire l'épidémiologie évolutive des maladies infectieuses. Ces prédictions théoriques peuvent être validées ou invalidées par des données empiriques ou expérimentales. Les données peuvent orienter les choix de modélisation pour un système hôte-pathogène en particulier. De plus, ajuster un modèle statistique permet d'estimer les valeurs de certains paramètres. Dans cette thèse, je combine à la fois des données démographiques (densités) et des données génétiques (fréquences des souches pathogènes).

Évolution phénotypique du SARS-CoV-2 : une approche par inférence statistique

Depuis son apparition fin 2019, le virus SARS-CoV-2 s'est propagé à l'échelle mondiale, provoquant la pandémie de COVID-19. À l'automne 2020, le variant Alpha a été détecté en Angleterre et s'est rapidement répandu, remplaçant la lignée précédente. Toutefois, les modifications phénotypiques sous-jacentes qui pourraient expliquer cet avantage sélectif ne sont pas très bien connues.

Dans cette étude, j'essaye de quantifier en quoi le variant Alpha différait de la lignée précédente pour deux traits phénotypiques : le taux de transmission et la durée de contagiosité. Dans cette optique, j'analyse les dynamiques épidémiologiques et évolutives conjointes en fonction du *Stringency Index* [99], un score qui mesure le degré de sévérité des mesures de contrôle mises en place pour freiner l'épidémie. En supposant que ces mesures réduisent les taux de contact et de transmission, j'ai développé une approche en deux étapes basée sur des modèles SEIR et l'analyse d'une combinaison d'informations épidémiologiques et évolutives. Dans une première étape, avant l'émergence du variant Alpha, je quantifie l'impact du *Stringency Index* sur la propagation du virus. Cette étape m'a permis d'inférer une fonction convexe qui permet de capturer l'effet du *Stringency Index* sur la réduction du nombre de contacts avec des hôtes sains. Dans une deuxième étape, après l'émergence du variant Alpha, j'analyse les changements de fréquence du variant Alpha. À partir d'un modèle SEIR, je propose une approximation du gradient de sélection reposant sur des hypothèses de sélection faible et de quasi-équilibre des variables épidémiologiques [63, 78, 110]. Je retrouve notamment un résultat classique des modèles SIR : l'intensité de la sélection pour des taux plus élevés de transmission dépend de la disponibilité des hôtes sains et des mesures de contrôles en place (par exemple, distanciation sociale, port du masque). En revanche, la sélection pour des durées de contagiosité plus longues est beaucoup moins sensible à ces mesures de contrôle. Grâce à mon estimation indépendante de l'efficacité des mesures de contrôle calculée lors de l'étape précédente, j'utilise mon approximation théorique du gradient de sélection pour estimer à la fois les différences de transmission et de guérison entre le variant Alpha et la lignée précédente. Je montre que l'avantage sélectif du variant Alpha résulte d'un taux de transmission accru (**Figure 5.4**). Une période de contagiosité plus longue semble beaucoup moins probable, bien que cette hypothèse ne puisse pas être complètement écartée. Ce résultat est en accord avec des études expérimentales qui ont confirmé l'avantage de transmission du variant Alpha [136].

[99]: Hale et al. (2021), 'A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)'

[63]: Gandon et al. (2022), 'Targeted vaccination and the speed of SARS-CoV-2 adaptation'
[78]: Lion (2018), 'Class structure, demography, and selection: reproductive-value weighting in nonequilibrium, polymorphic populations'
[110]: Lion et al. (2022), 'Evolution of class-structured populations in periodic environments'

[136]: Lai et al. (2023), 'Evolution of SARS-CoV-2 shedding in exhaled breath aerosols'

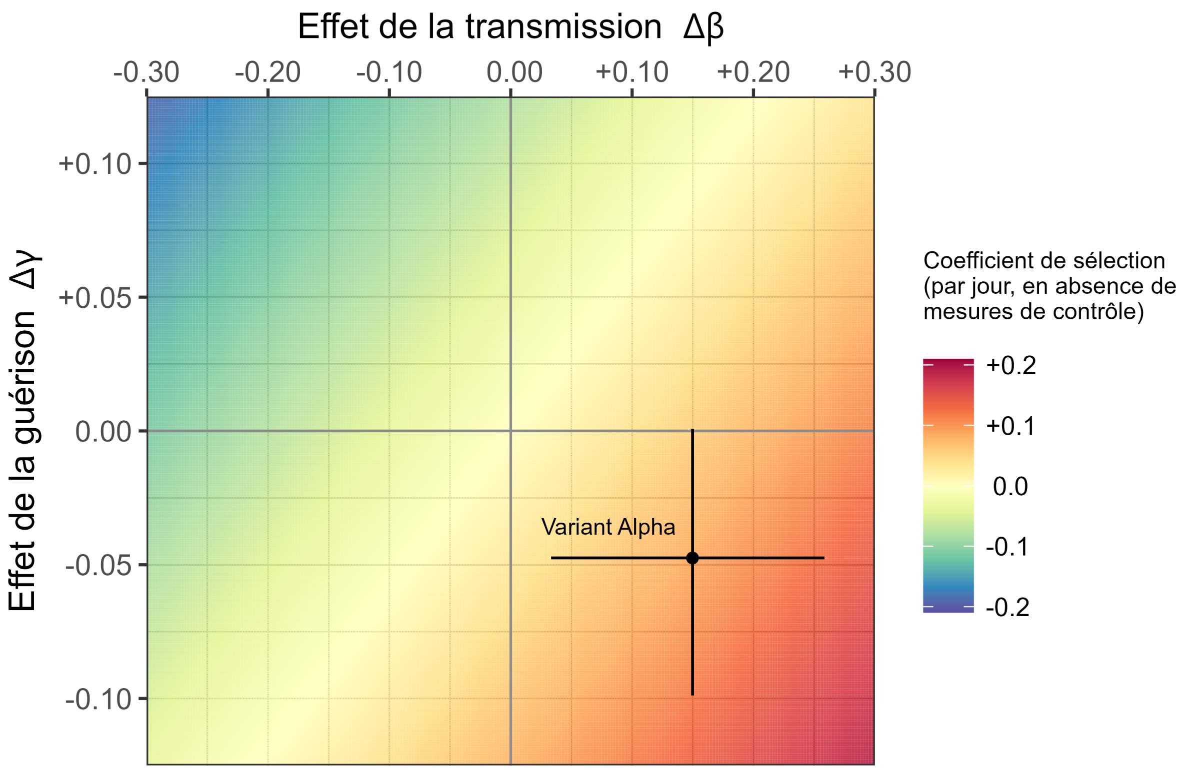


Figure 5.4: Profil phénotypique du variant Alpha (taux de transmission et de guérison) relativement à la souche résidente. Par définition, le phénotype de la souche résidente est situé à l'origine du graphique ($\Delta\beta = 0; \Delta\gamma = 0$). Les estimations obtenues à partir d'un modèle linéaire mixte (point noir, exprimé par jour) des différences phénotypiques en termes de transmission $\Delta\beta$ et de guérisons $\Delta\gamma$ ainsi que l'intervalle de confiance à 95% (IC95, croix noire) ont été obtenus à partir des meilleures estimations des paramètres du contrôle (phase 1). Nous avons estimé $\Delta\beta = 0.15$ (IC95 [0.033; 0.258]) et $\Delta\gamma = -0.047$ (IC95 [−0.099; +0.001]). Le fond coloré représente les valeurs du coefficient de sélection (en l'absence de mesures de contrôle) en fonction de $\Delta\beta$ et de $\Delta\gamma$; le coefficient de selection coefficient est ici autour de +0.11 par jour (ou +0.77 par semaine) pour le variant Alpha.

Évolution de la virulence dans les épidémies émergentes : aller-retour entre théorie et évolution expérimentale

La validation expérimentale de prédictions théoriques est une étape importante pour démontrer le pouvoir prédictif d'un modèle. Alors que les validations quantitatives sont courantes en épidémiologie des maladies infectieuses, la microbiologie expérimentale reste souvent limitée à l'évaluation d'une simple correspondance qualitative entre les prédictions d'un modèle et les résultats expérimentaux.

Dans cette étude, je développe une approche quantitative avec une population virale polymorphique. J'analyse les données d'évolution expérimentale d'une étude précédente sur l'évolution du bactériophage (ou phage) tempéré λ se propageant dans des cultures continues bactériennes d'*Escherichia coli* [68]. Ce travail expérimental a confirmé l'influence des dynamiques épidémiologiques sur l'évolution de la transmission et de la virulence du virus. Un variant ayant une plus grande propension à lyser les cellules bactériennes a été favorisé dans les épidémies émergentes (lorsque la densité de cellules sensibles était importante), mais contre-sélectionné lorsque la plupart des cellules étaient infectées. Bien que cette approche ait validé qualitativement une prédiction théorique

[68]: Berngruber et al. (2013), 'Evolution of virulence in emerging epidemics'

importante, aucune tentative n'a été faite pour ajuster le modèle aux données ni pour développer davantage le modèle afin d'améliorer la qualité de l'ajustement.

Je montre ici comment l'analyse théorique et l'ajustement de modèles aux données peuvent être utilisés pour estimer les paramètres clés du cycle de vie du phage λ et pour mieux comprendre l'épidémiologie évolutive du virus. Premièrement, j'ai constaté que le modèle original ne parvenait pas à capturer correctement les dynamiques évolutives transitoires dans le compartiment infecté. J'améliore ici l'ajustement avec les données expérimentales en distinguant deux types de cellules infectées : les cellules lysogéniques (L) et les cellules engagées dans la voie lytique (Y). L'ajout d'un compartiment supplémentaire permet de modéliser un pic transitoire dans la fréquence des hôtes infectés par le phage virulent lors du début de l'épidémie (**Figure 5.5-B**), cohérent avec les données expérimentales. Cet apport permet également de prendre en compte le temps de lyse (instantané dans le modèle original). Deuxièmement, je fais une analyse théorique de ce nouveau modèle pour aller au-delà de l'approche purement numérique utilisée dans [68]. Cela me permet d'obtenir des approximations analytiques sur la propagation du phage virulent au début et à la fin de l'épidémie. Troisièmement, je développe une approche d'inférence pour estimer les paramètres du modèle (**Figure 5.5**). J'implémente une procédure par maximum de vraissemblance en deux étapes : (i) j'estime grâce à un modèle linéaire les taux de réactivation des prophages des deux souches virales, (ii) je fixe ces valeurs des taux de réactivation puis j'effectue des optimisations non linéaires pour inférer les autres paramètres du modèle.

[68]: Berngruber et al. (2013), 'Evolution of virulence in emerging epidemics'

Cette approche d'inférence est très différente des études précédentes. Tout d'abord, j'analyse les dynamiques d'une population virale polymorphique. Deuxièmement, j'utilise trois types de données pour estimer les valeurs des paramètres du modèle : (i) des données épidémiologiques (la prévalence de l'infection), (ii) les changements de fréquence du variant et (iii) la différenciation du variant entre compartiments. Chaque type de données apporte des informations complémentaires et me permet d'estimer conjointement les traits d'histoire de vie des deux souches du phage λ. Cette nouvelle méthode est particulièrement bien adaptée pour estimer les taux de réactivation des prophages, pour lesquels seul le traitement endémique est nécessaire. L'estimation des probabilités de lysogénisation est cependant plus difficile car elle nécessite de prendre en compte les dynamiques épidémiologiques transitoires du traitement épidémique. La rapide rétroaction épidémiologique qui a lieu lors des épidémies émergentes – c'est-à-dire la baisse du nombre d'hôtes sains – rend nécessaire de considérer à la fois l'épidémiologie et l'évolution de l'infection.

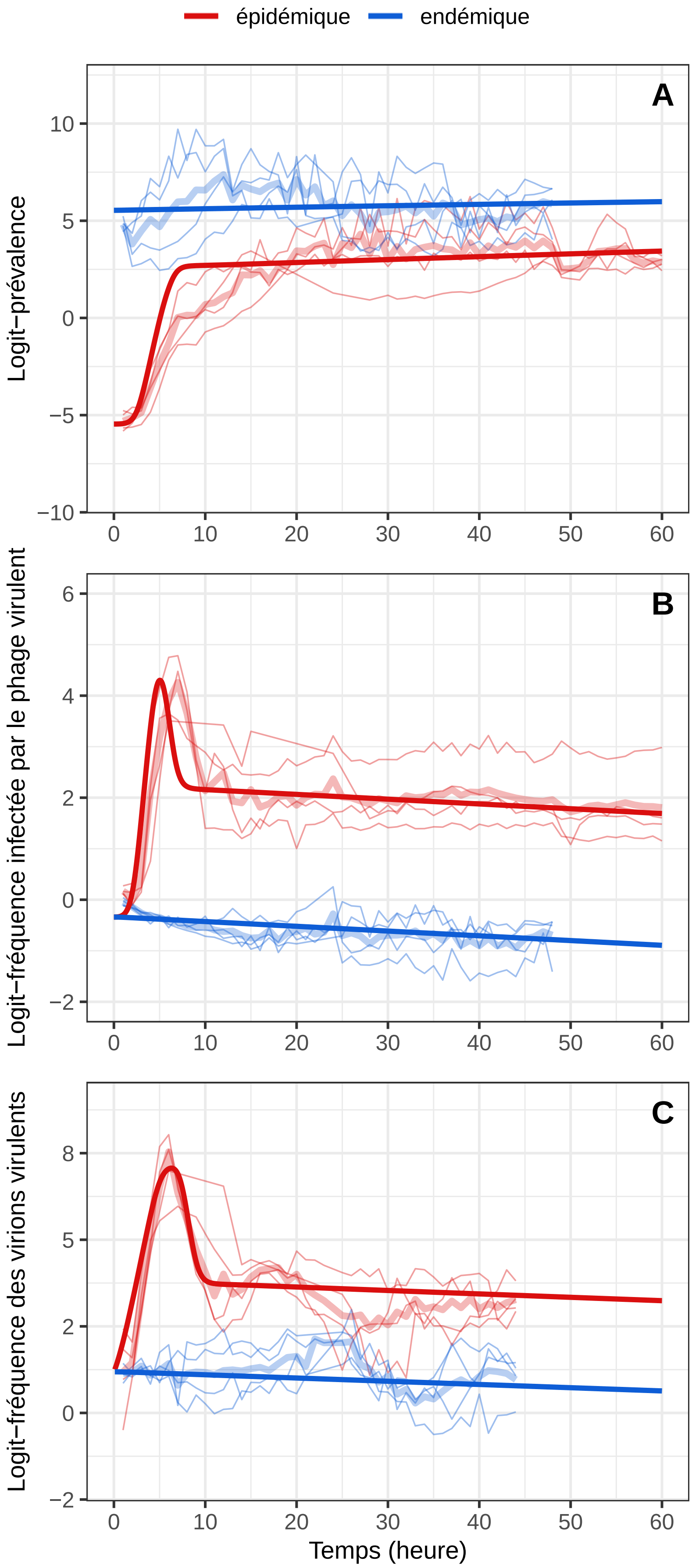


Figure 5.5: Modèle ajusté aux données expérimentales (projet Lambda). Les valeurs ajustées (courbes épaisses de nuance sombre) ont été simulées à partir des estimations obtenues par maximum de vraissemblance (données : courbes fines de nuance claire). La prévalence initiale est soit faible (autour de 1%, traitement épidémique, en rouge), soit haute (autour de 99%, traitement endémique, en bleu). Les courbes épaisses de nuance claire correspondent à la moyenne des valeurs sur l'échelle logit pour tous les réplicats par traitement. (A) Logit-prévalence de l'infection; (B) logit-fréquence des cellules infectées par le phage mutant (virulent); (C) logit-fréquence du phage mutant (virulent) dans le milieu de culture (stade virion ou virus libre).

Déplacements d'hôtes et évolution des pathogènes

L'adaptation des agents pathogènes affecte notre capacité à contrôler les épidémies et représente un enjeu majeur de santé publique. Lors de l'acquisition de mutations bénéfiques, de nouveaux variants apparaissent et remplaçent parfois les souches précédentes. La force de la sélection sur un variant émergent est classiquement quantifiée au cours de sa propagation à partir des séries temporelles de sa fréquence au sein des hôtes infectés. Cette approche marche très bien sous l'hypothèse d'une population homogène. Néanmoins, les populations d'hôtes sont généralement spatialement hétérogènes et interconnectés via des déplacements ("migration") d'hôtes d'une population à une autre. Cette migration peut conduire à l'introduction d'un pathogène dans de nouvelles populations ou à la persistence globale du pathogène ; la migration peut aussi affecter la vitesse et la manière dont se propager les épidémies. Outre ses effets sur l'épidémiologie, la migration peut également impacter l'évolution d'une population pathogène polymorphique. Cependant, peu de choses sont vraiment connus sur la manière dont les dynamiques évolutives des agents pathogènes sont façonnées par la migration des hôtes, en particulier à court-terme.

Dans cette dernière étude, j'étudie la dynamique transitoire de la fréquence et de la différenciation d'un variant en compétition avec la souche sauvage au sein d'une métapopulation. Dans cette métapopulation, deux populations sont interconnectées par des flux d'hôtes qui peuvent mutellement se rendre visite (**Figure 5.6**). Je considére ici un scénario avec une sélection homogène dans les deux populations, et un scénario avec une sélection hétérogène. Je met notamment l'accent sur l'utilité des modèles dynamiques mécanistiques pour séparer les effets de la migration et ceux de la sélection sur l'évolution des pathogènes. La migration peut notamment biaiser la force apparente de la sélection et ainsi conduire à des estimations erronées sur l'avantage sélectif réel des variants.

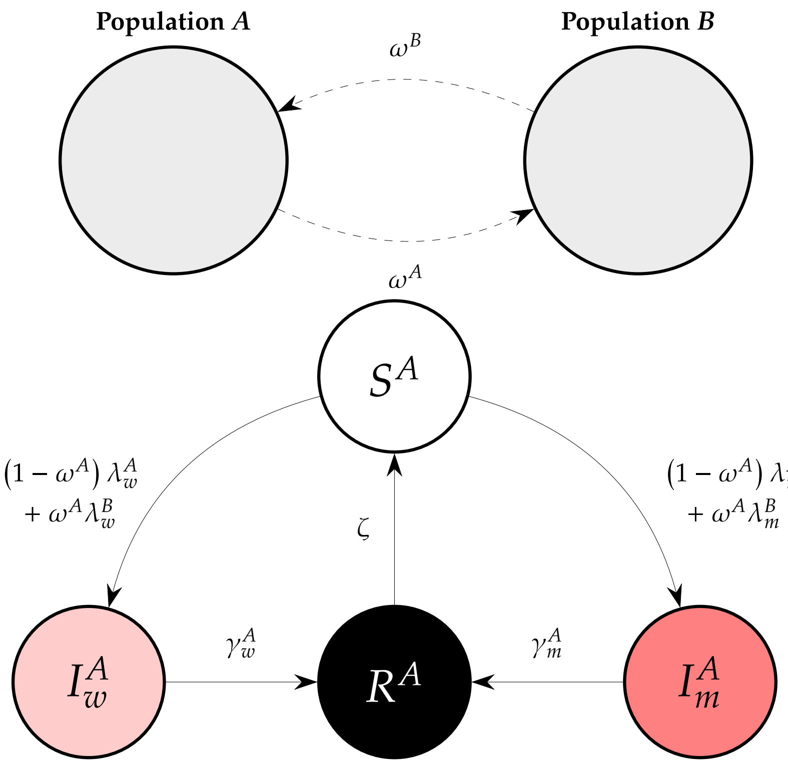


Figure 5.6: Diagramme du modèle SIRS dans une métapopulation. Panneau du haut : je modélise une métapopulation d'hôtes à deux populations; les hôtes de la population A (resp. B) peuvent transitoirement rendre visite à la population B (resp. A) avec une probabilité ω^A (resp. ω^B) ou rester dans la population locale avec des probabilités complémentaires. Panneau du bas : modèle SIRS pour la population locale A. L'indice w désigne la souche sauvage tandis que l'indice m désigne la souche mutante (ou variant); λ_w^A et λ_m^A (resp. λ_w^B et λ_m^B) représentent respectivement les forces d'infection par le sauvage et par le variant auxquelles sont soumis les hôtes sains de la population A (resp. B).

Conclusion

Cette thèse se situe à l'interface entre épidémiologie évolutive théorique et analyses statistiques à partir de données empiriques et expérimentales. J'ai une formation en biologie mais j'ai toujours été intéressé par les approches interdisciplinaires, en particulier entre la biologie et les mathématiques. Je m'intéresse également à l'analyse de données réelles et aux statistiques. Ce travail reflète les discussions et la collaboration avec mes directeurs de thèse (un statisticien et un biologiste théoricien) aux expertises complémentaires. Cela m'a permis d'adopter une perspective intégrative qui offre une compréhension fine de la manière dont l'évolution phénotypique des agents pathogènes est impactée par les rétroactions épidémiologiques et comment la combinaison d'informations entre données épidémiologiques et génétiques peut permettre d'estimer les traits phénotypiques des pathogènes. Comme pour tout travail interdisciplinaire, c'est une tâche difficile. Les jeux de données sont généralement incomplets en raison de processus cachés et de données manquantes, augmentant alors la complexité des modèles éco-épidémiologiques et rendant encore plus difficile le processus d'estimation des paramètres du modèle.

Tout au long de ma thèse, mon travail s'est porté sur des analyses de modèles déterministes reposant sur des systèmes dynamiques d'équations différentielles ordinaires. J'ai mené trois projets qui m'ont permis de développer de nouveaux outils pour exploiter davantage des jeux de données incomplets et extraire des informations jusqu'alors inaccessibles sur la dynamique de propagation et d'évolution d'agents pathogènes dans un environnement hétérogène. Ces nouveaux outils reposent sur l'incorporation explicite des processus cachés. Suivre l'évolution phénotypique des agents pathogènes est essentiel, notamment pour mieux comprendre et anticiper la dynamique de leur adaptation. L'inférence statistique construite à partir de modèles dynamiques mécanistiques y est alors très utile pour améliorer notre compréhension de la dynamique des maladies infectieuses. Dans cet effort, les allers-retours entre modèles et données stratifiées par compartiments sont des étapes importantes.

Ce travail de thèse pourrait être étendu avec une étude plus poussée de l'hétérogénéité induite par l'immunité de l'hôte – naturelle ou vaccinale –, la modélisation de la co-évolution des hôtes ou encore l'étude de la compétition entre de multiples souches.

Appendix

A

An introduction to evolutionary epidemiology theory

The following teaching material is the instructor version of a 3-hour practical course that was developed for the Winter school "Quantitative Viral Dynamics Across Scales" (March 2022, organizer: Joshua S. Weitz), held at the École Normale Supérieure (Ulm) in Paris. This work was carried out in collaboration with Martin Guillemet (former PhD student) and Sylvain Gandon.

An introduction to evolutionary epidemiology theory: Evolution of virulence and transmission

Wakinyan Benhamou **Martin Guillemet** **Sylvain Gandon**
Translation of R code into Python by **Adriana Lucia-Sanz***

Thursday, March 24th, 2022

*Georgia Institute of Technology, Atlanta, GA, USA

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 analytical approach with a numerical exploration of the models. The plan is to present/discuss briefly the analytical 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.

The course will consist of the following 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:

$$\dot{I}(t) = \Bigg(\beta S_0 - (\delta + \alpha + \gamma)\Bigg) I(t)$$

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) # for data manipulation
library(ggplot2) # for graphic visualizations
library(cowplot) # to arrange and show multiple subplots (function 'plot_grid')
library(deSolve) # to numerically integrate a system of ordinary differential equations
library(scales) # to manipulate the internal scaling infrastructure used by ggplot2
library(lattice) # to draw level plots
library(knitr) # to show nice tables (function 'kable')
```

```
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"] # Influx of new individuals
  delta <- parms["delta"] # Natural death rate
  beta <- parms["beta"] # Horizontal transmission rate
  alpha <- parms["alpha"] # Mortality rate induced by the infection (virulence)
  gamma <- parms["gamma"] # Recovery rate

  # 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
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

```
# 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

```
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
```

```
tail(simul_disease_free, n = 2) # 2 last rows of the table
```

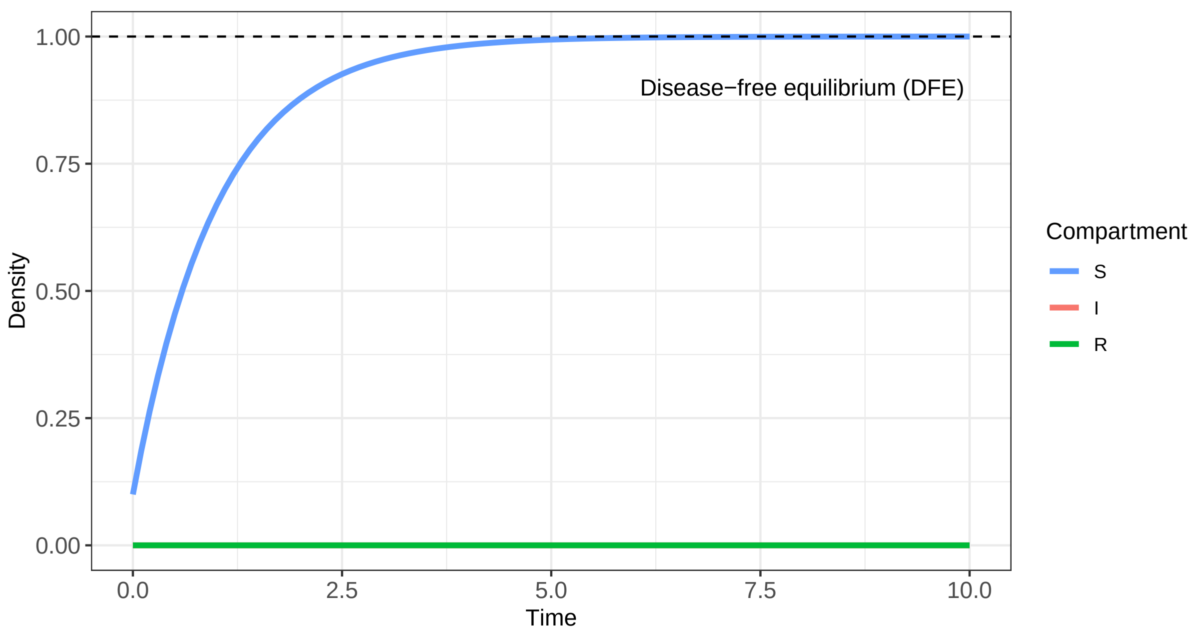
```
##        time         S I R
## [100,]  9.9 0.9999548 0 0
## [101,] 10.0 0.9999591 0 0
```

Formatting of simulated data & graphical visualization

```
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 = " ; ")))
```

```r
  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



Parameters: lambda = 1 ; delta = 1 ; beta = 5 ; alpha = 0.1 ; gamma = 0.1

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.

```r
# 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
```

```r
                   "I" = I_t0, # I(t0), low initial density of infected (I) individuals
                   "R" = 0)     # R(t0), no recovered (R) individuals
```

Numerical integration

```r
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
```

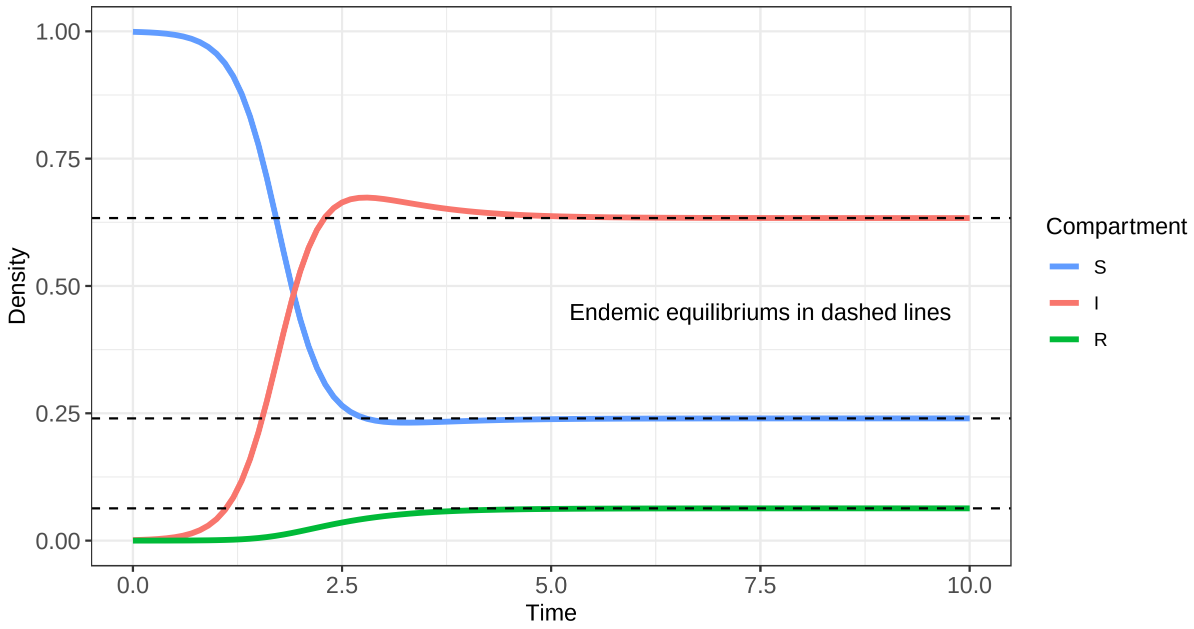
```r
tail(simul_disease, n = 2) # 2 last rows of the table
```

```
##        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

```r
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

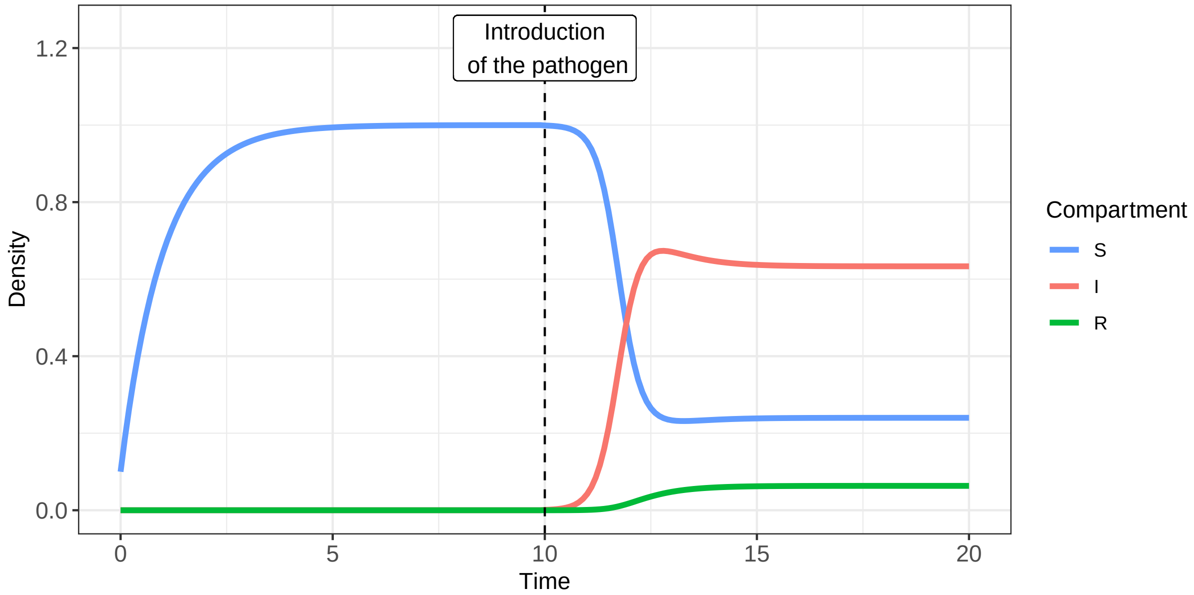
As expected from the analysis of the model, the density of infected hosts increases because $R_0 = 4.16 > 1$. After a transient phase, the dynamical variables $S(t)$, $I(t)$ and $R(t)$ converge toward the equilibrium values derived above (*i.e.*, S_e, I_e and R_e).

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



Parameters: lambda = 1 ; delta = 1 ; beta = 5 ; alpha = 0.1 ; gamma = 0.1

```
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
```

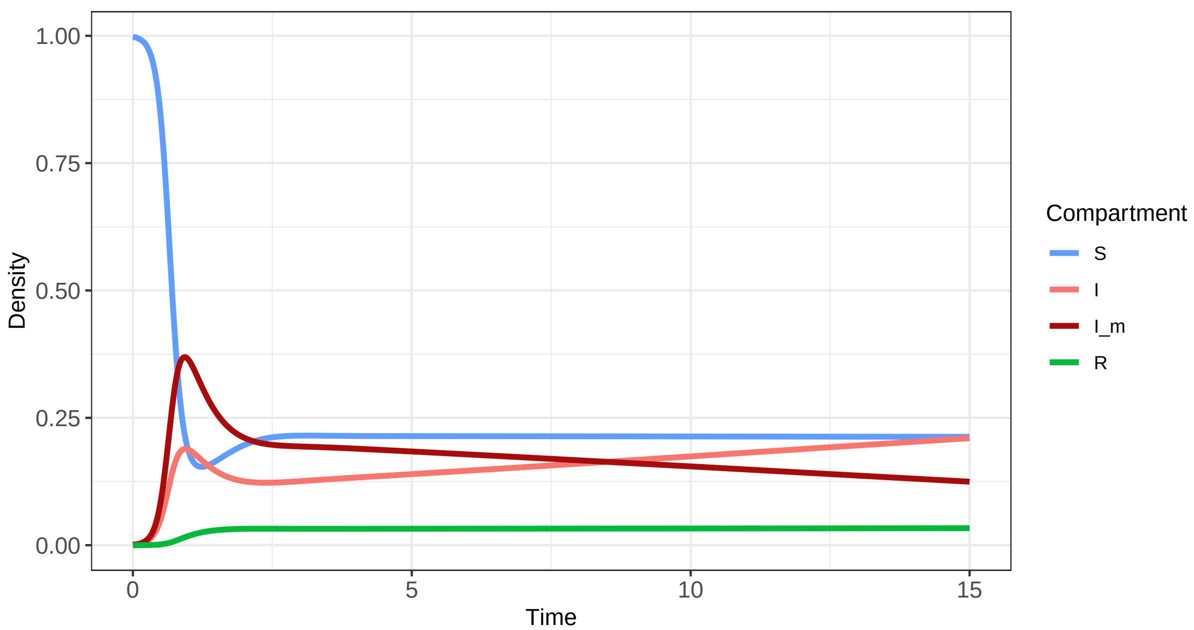
```r
tail(simul, n = 2) # 2 last rows of the table
```

```
##           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:

$$\begin{aligned} \dot{S}(t) &= \lambda - \overline{\beta}(t)I_T(t)S(t) - \delta S(t) \\ \dot{I}_T(t) &= \overline{\beta}(t)I_T(t)S(t) - (\delta + \overline{\alpha}(t) + \overline{\gamma}(t))I_T(t) \\ \dot{R}(t) &= \overline{\gamma}(t)I_T(t) - \delta R(t) \end{aligned} \tag{3a}$$

$$\text{where } I_T(t) = I(t) + I_m(t) \qquad \text{and} \qquad \begin{aligned} \overline{\beta}(t) &= (1 - p_m(t))\beta + p_m(t)\beta_m \\ \overline{\alpha}(t) &= (1 - p_m(t))\alpha + p_m(t)\alpha_m \\ \overline{\gamma}(t) &= (1 - p_m(t))\gamma + p_m(t)\gamma_m \end{aligned} \qquad \text{with } p_m(t) = \frac{I_m(t)}{I_T(t)}$$

$$\dot{p}_m(t) = \underbrace{p_m(t)(1 - p_m(t))}_{\text{genetic variance}} \underbrace{(r_m(t) - r(t))}_{\text{selection coefficient}} \tag{3b}$$

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

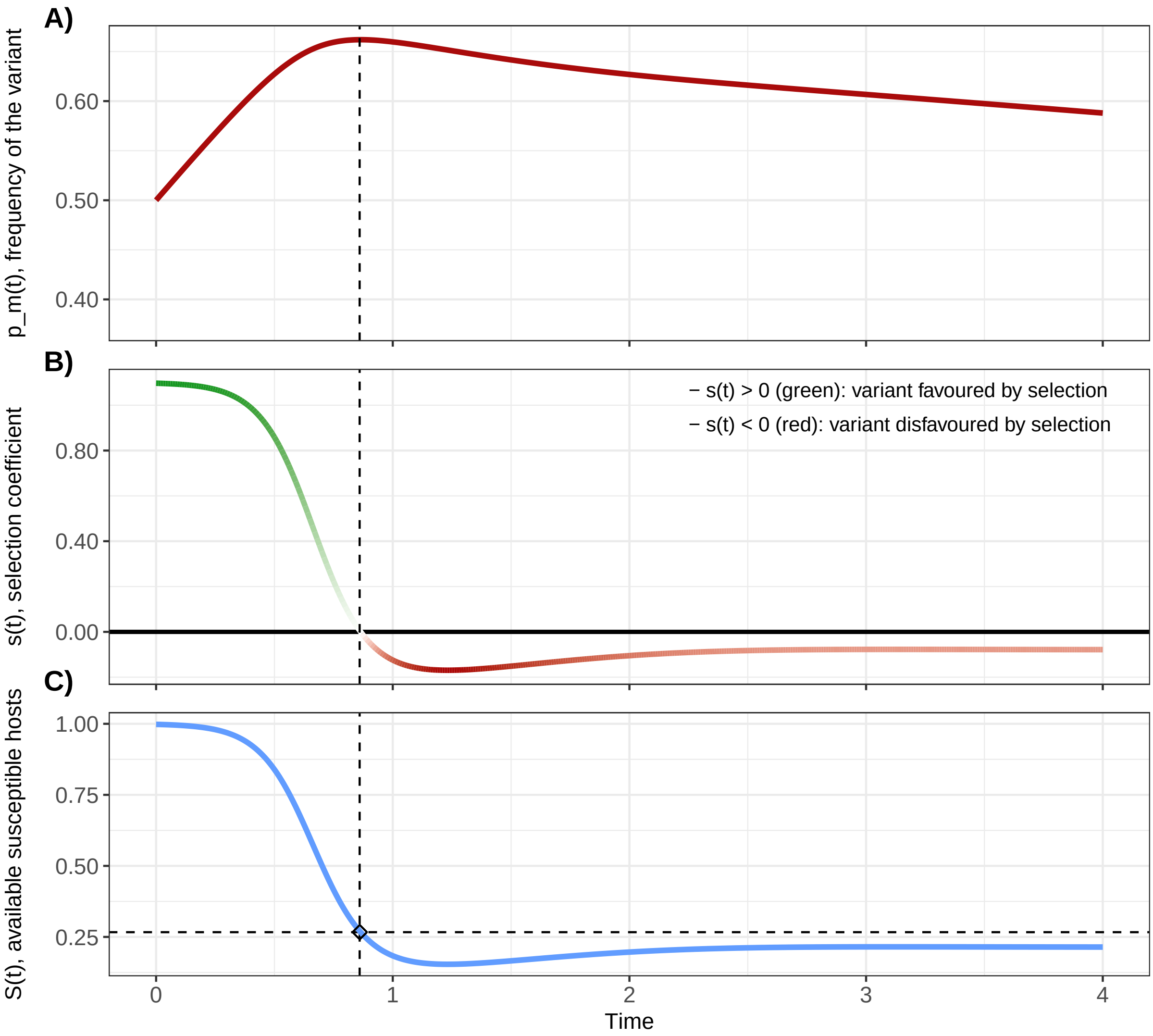
Note that the first term acts on the production of new infections (*i.e.* birth rate of the infection) while the last two points act on the duration of infection (*i.e.* lower death rate of the infection).

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.

```
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)) +
```

```r
  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).

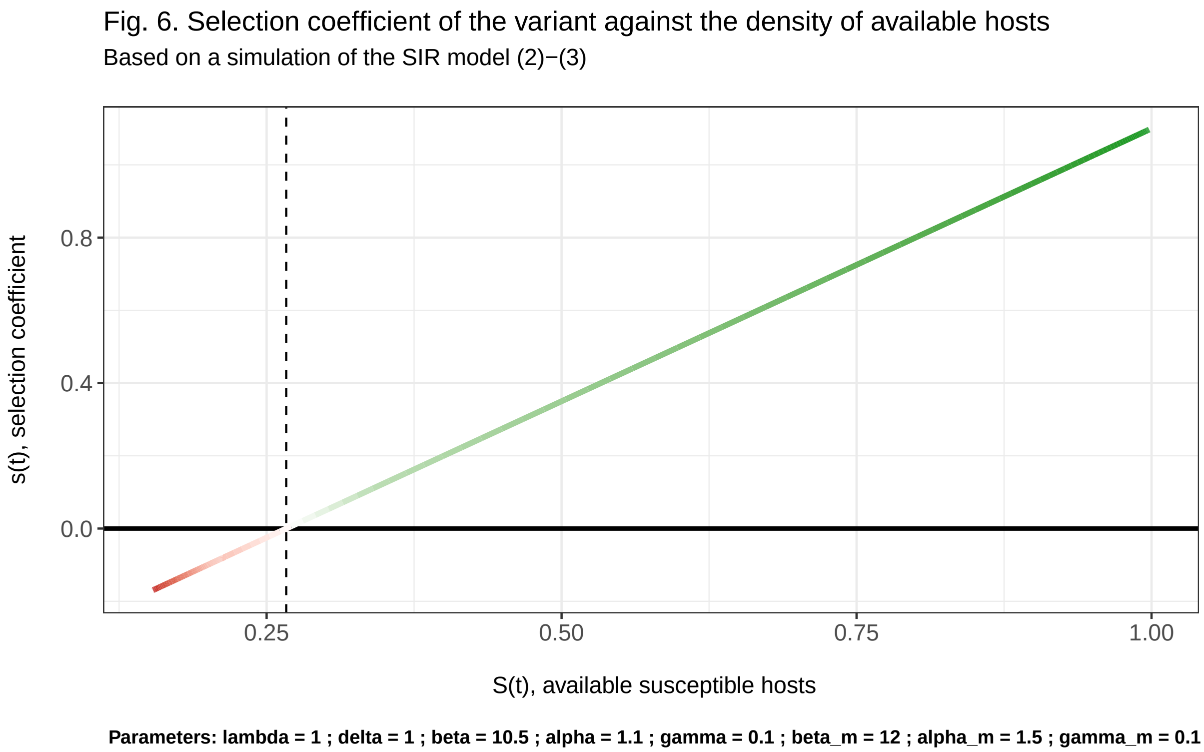
We added here the temporal dynamics of the S compartment (*cf.* **Fig. 5-C**). Note how the dynamics of $S(t)$ mirrors the dynamics of the selection coefficient. A particular value of $S(t)$ is associated with the time point when $s(t) = 0$ (*i.e.*, when the variant reaches its maximum frequency).

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):

$$s(t) = 0 \quad \iff \quad (\beta_m - \beta)\hat{S}(t) + (\alpha + \gamma) - (\alpha_m + \gamma_m) = 0 \quad \iff \quad \hat{S}(t) = \frac{(\alpha_m + \gamma_m) - (\alpha + \gamma)}{\beta_m - \beta}$$

```
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'))
```



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)
}
```

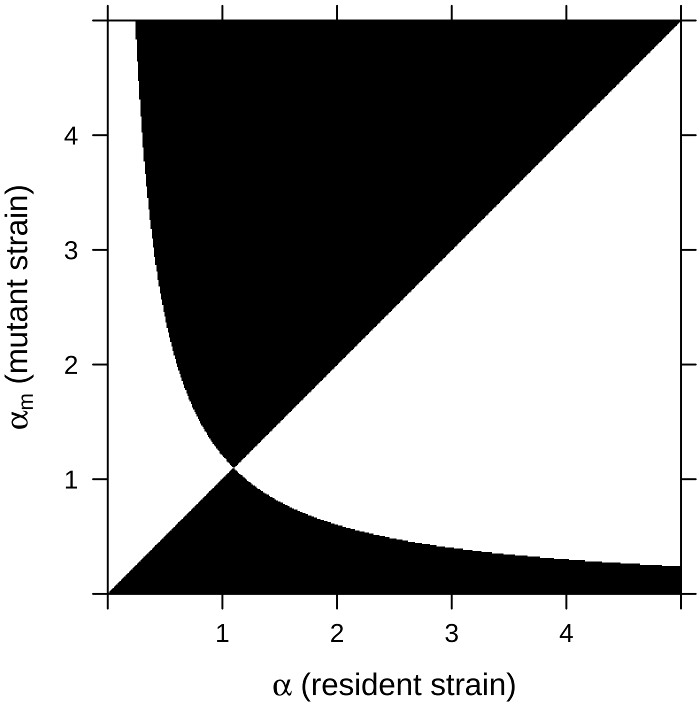
```
# Parameters

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'

# Analytical result
alpha_ESS <- delta + gamma

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(alpha_ESS) %>% as.data.frame
  colnames(tab) <- "$\\alpha^{*} (time^{-1})$"
  rownames(tab) <- c("Numerical approximation", "Analytical solution")
  kable(tab)
}
```

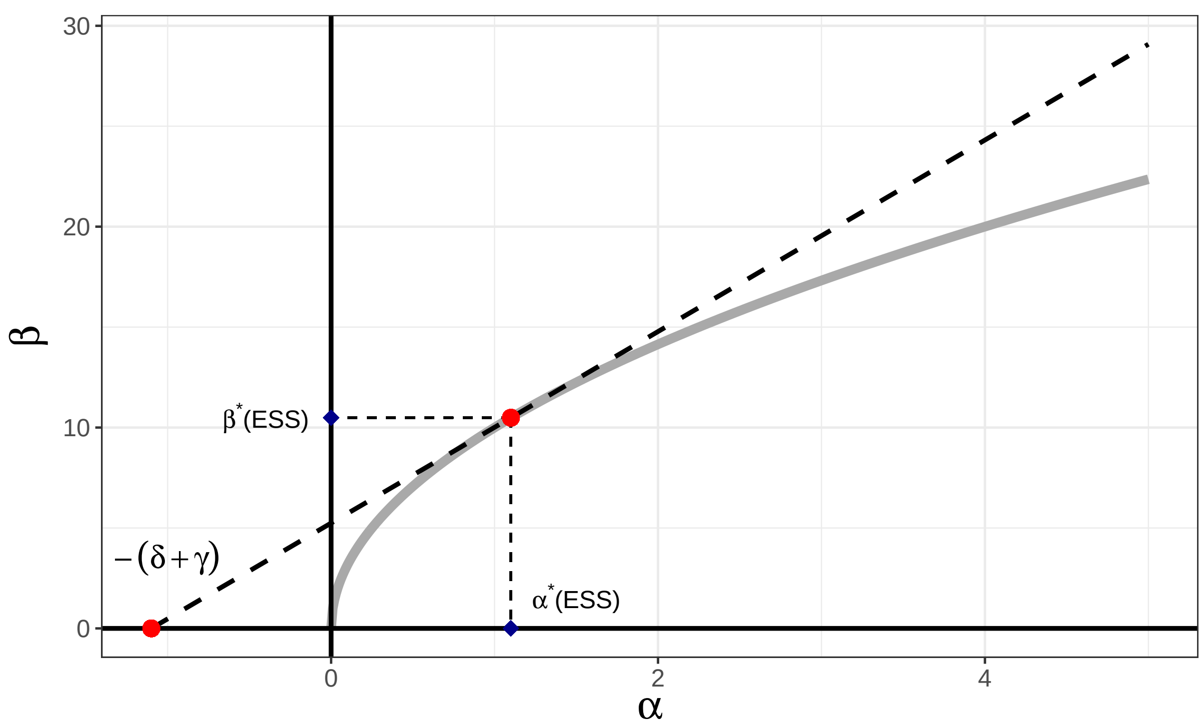
	$\alpha^*(time^{-1})$
Numerical approximation	1.102 +/- 0.005
Analytical solution	1.1

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)

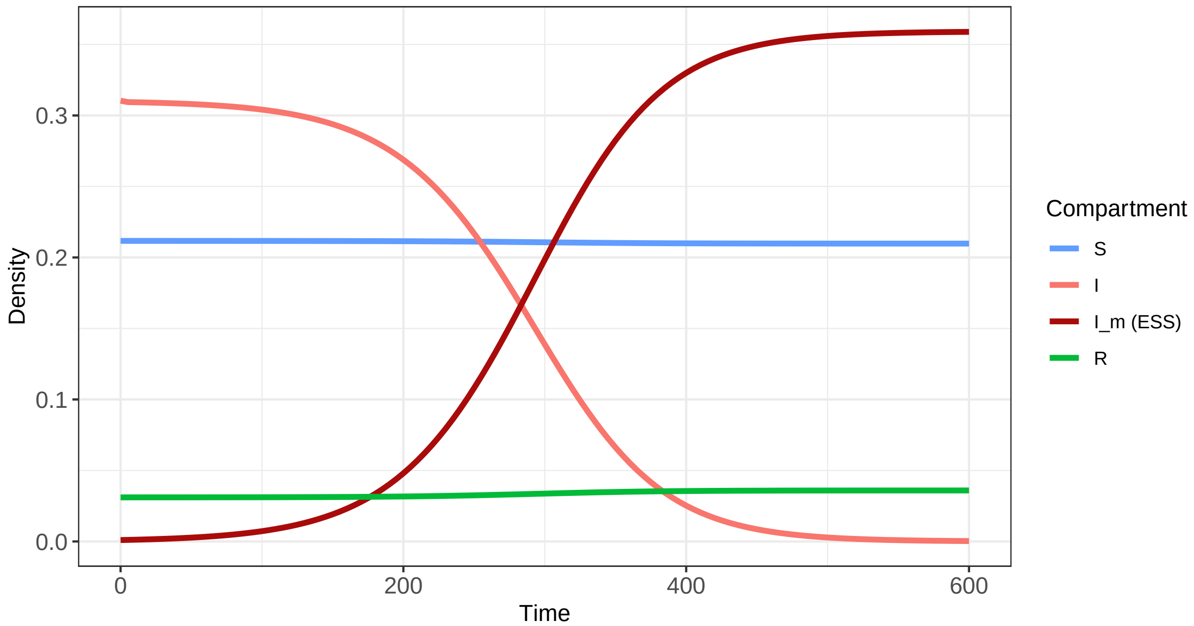
2.3.1 The ESS wins in the long term...

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
```

```r
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

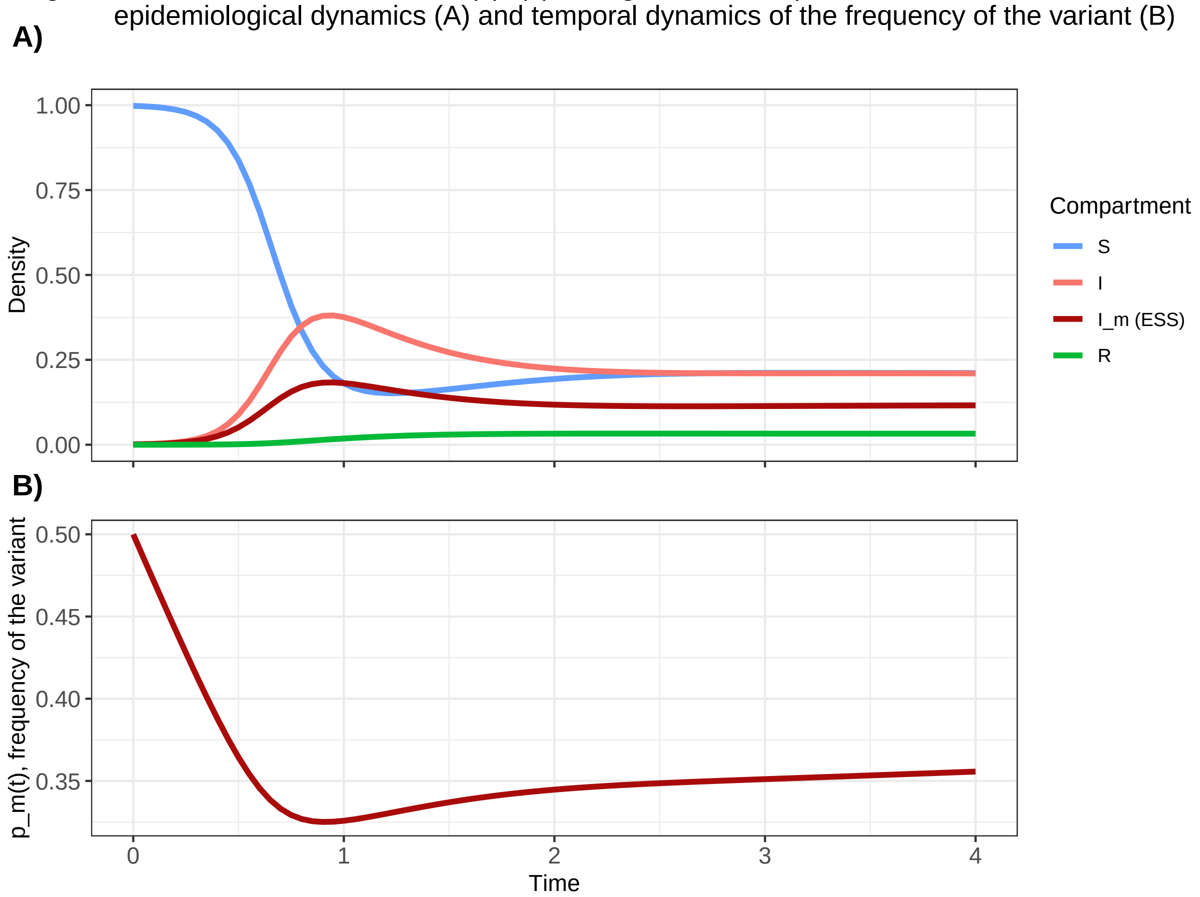
Q8. Starting from the disease-free equilibrium, introduce two pathogen (one at the ESS) in small but equal densities, both following the same trade-off function. Is the ESS pathogen always more selected than the other pathogen? For the second pathogen, try with $\beta < \beta_m$ and $\beta > \beta_m$. 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')),
```

```r
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.

Bibliography

This is the bibliography supporting Chapter one (General introduction) and Chapter five (General discussion).

[1] Simon Benhamou and Valérie Séguinot. 'How to find one's way in the labyrinth of path integration models'. In: *Journal of Theoretical Biology* 174.4 (1995), pp. 463–466. DOI: 10.1006/jtbi.1995.0112 (cited on page vi).

[2] Hongzhou Lu, Charles W Stratton, and Yi-Wei Tang. 'Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle'. In: *Journal of medical virology* 92.4 (2020), p. 401. DOI: 10.1002/jmv.25678 (cited on pages 3, 167).

[3] WHO. 'WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020'. In: (2020) (cited on pages 3, 167).

[4] Raymond E Goldstein. 'Are theoretical results 'Results'?' In: *Elife* 7 (2018), e40018. DOI: 10.7554/eLife.40018 (cited on pages 3, 167).

[5] Daniel Bernoulli. 'An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it'. In: *Histoire de l'Académie royale des sciences avec les mémoires de mathématique, de physique* (1760) (cited on pages 3, 167).

[6] Ronald Ross. 'Inaugural lecture on the possibility of extirpating malaria from certain localities by a new method'. In: *British medical journal* 2.2009 (1899), p. 1. DOI: 10.1136/bmj.2.2009.1 (cited on pages 3, 167).

[7] Ronald Ross. 'The logical basis of the sanitary policy of mosquito reduction'. In: *Science* 22.570 (1905), pp. 689–699. DOI: 10.1126/science.22.570.689 (cited on pages 3, 167).

[8] Ronald Ross. *The prevention of malaria*. John Murray, 1911 (cited on pages 3, 167).

[9] William Ogilvy Kermack and Anderson G McKendrick. 'A contribution to the mathematical theory of epidemics'. In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), pp. 700–721. DOI: 10.1098/rspa.1927.0118 (cited on pages 3, 7, 26, 162, 167).

[10] Henrik Salje, Cécile Tran Kiem, Noémie Lefrancq, Noémie Courtejoie, Paolo Bosetti, Juliette Paireau, Alessio Andronico, Nathanaël Hozé, Jehanne Richet, Claire-Lise Dubost, et al. 'Estimating the burden of SARS-CoV-2 in France'. In: *Science* 369.6500 (2020), pp. 208–211. DOI: 10.1126/science.abc3517 (cited on pages 3, 167).

[11] Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, et al. 'Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand'. In: 16 (2020). DOI: 10.25561/77482 (cited on pages 3, 167).

[12] Juliette Paireau, Alessio Andronico, Nathanaël Hozé, Maylis Layan, Pascal Crepey, Alix Roumagnac, Marc Lavielle, Pierre-Yves Boëlle, and Simon Cauchemez. 'An ensemble model based on early predictors to forecast COVID-19 health care demand in France'. In: *Proceedings of the National Academy of Sciences* 119.18 (2022), e2103302119. DOI: 10.1073/pnas.2103302119 (cited on pages 3, 167).

[13] Nathan D Grubaugh, Mary E Petrone, and Edward C Holmes. 'We shouldn't worry when a virus mutates during disease outbreaks'. In: *Nature microbiology* 5.4 (2020), pp. 529–530. DOI: 10.1038/s41564-020-0690-4 (cited on pages 4, 168).

[14] Jason W Rausch, Adam A Capoferri, Mary Grace Katusiime, Sean C Patro, and Mary F Kearney. 'Low genetic diversity may be an Achilles heel of SARS-CoV-2'. In: *Proceedings of the National Academy of Sciences* 117.40 (2020), pp. 24614–24616. DOI: 10.1073/pnas.2017726117 (cited on pages 4, 168).

[15] Bette Korber, Will M Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, Elena E Giorgi, Tanmoy Bhattacharya, Brian Foley, et al. 'Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus'. In: *Cell* 182.4 (2020), pp. 812–827. DOI: 10.1016/j.cell.2020.06.043 (cited on page 4).

[16] Jessica A Plante, Yang Liu, Jianying Liu, Hongjie Xia, Bryan A Johnson, Kumari G Lokugamage, Xianwen Zhang, Antonio E Muruato, Jing Zou, Camila R Fontes-Garfias, et al. 'Spike mutation D614G alters SARS-CoV-2 fitness'. In: *Nature* 592.7852 (2021), pp. 116–121. DOI: 10.1038/s41586-020-2895-3 (cited on pages 4, 168).

[17] Erik Volz, Verity Hill, John T McCrone, Anna Price, David Jorgensen, Áine O'Toole, Joel Southgate, Robert Johnson, Ben Jackson, Fabricia F Nascimento, et al. 'Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity'. In: *Cell* 184.1 (2021), pp. 64–75. DOI: 10.1016/j.cell.2020.11.020 (cited on pages 4, 168).

[18] Nathan D Grubaugh, William P Hanage, and Angela L Rasmussen. 'Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear'. In: *Cell* 182.4 (2020), pp. 794–795. DOI: 10.1016/j.cell.2020.06.040 (cited on pages 4, 168).

[19] Public Health England. *Investigation of novel SARS-COV-2 variant 202012/01: technical briefing 5*. 2020 (cited on pages 4, 23, 168).

[20] Erik Volz, Swapnil Mishra, Meera Chand, Jeffrey C Barrett, Robert Johnson, Lily Geidelberg, Wes R Hinsley, Daniel J Laydon, Gavin Dabrera, Áine O'Toole, et al. 'Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England'. In: *Nature* 593.7858 (2021), pp. 266–269. DOI: 10.1038/s41586-021-03470-x (cited on pages 4, 168).

[21] Shruti Khare, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo, et al. 'GISAID's role in pandemic response'. In: *China CDC weekly* 3.49 (2021), p. 1049. DOI: 10.46234/ccdcw2021.255 (cited on pages 4, 23, 168).

[22] Sébastien Lion, Akira Sasaki, and Mike Boots. 'Extending eco-evolutionary theory with oligomorphic dynamics'. In: *Ecology Letters* 26 (2023), S22–S46. DOI: 10.1111/ele.14183 (cited on pages 7, 14, 17).

[23] Stefan AH Geritz, E Kisdi, Géza Meszéna, and Johan AJ Metz. 'Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree'. In: *Evolutionary ecology* 12 (1998), pp. 35–57 (cited on pages 7, 14).

[24] Ulf Dieckmann. 'Adaptive dynamics of pathogen-host interactions'. In: *Adaptive Dynamics of Infectious Diseases: In Pursuit of Virulence Management*. Ed. by Ulf Dieckmann, Johan AJ Metz, Maurice W Sabelis, and Karl Sigmund. Cambridge University Press, 2002, pp. 39–59 (cited on pages 7, 14).

[25] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991 (cited on pages 7, 10).

[26] Einav G Levin, Yaniv Lustig, Carmit Cohen, Ronen Fluss, Victoria Indenbaum, Sharon Amit, Ram Doolman, Keren Asraf, Ella Mendelson, Arnona Ziv, et al. 'Waning immune humoral response to BNT162b2 Covid-19 vaccine over 6 months'. In: *New England Journal of Medicine* 385.24 (2021), e84. DOI: 10.1056/NEJMoa2114583 (cited on page 8).

[27] UKHSA. *COVID-19 vaccine surveillance report – Week 16*. 2022 (cited on page 8).

[28] Alessandro M Carabelli, Thomas P Peacock, Lucy G Thorne, William T Harvey, Joseph Hughes, Sharon J Peacock, Wendy S Barclay, Thushan I De Silva, Greg J Towers, and David L Robertson. 'SARS-CoV-2 variant biology: immune escape, transmission and fitness'. In: *Nature Reviews Microbiology* 21.3 (2023), pp. 162–177. DOI: 10.1038/s41579-022-00841-7 (cited on page 8).

[29] Hamish McCallum, Nigel Barlow, and Jim Hone. 'How should pathogen transmission be modelled?' In: *Trends in ecology & evolution* 16.6 (2001), pp. 295–300. DOI: 10.1016/S0169-5347(01)02144-9 (cited on page 8).

[30] Jacco Wallinga and Marc Lipsitch. 'How generation intervals shape the relationship between growth rates and reproductive numbers'. In: *Proceedings of the Royal Society B: Biological Sciences* 274.1609 (2007), pp. 599–604. DOI: 10.1098/rspb.2006.3754 (cited on pages 9, 10, 162).

[31] Raphaël Forien, Guodong Pang, and Étienne Pardoux. 'Estimating the state of the COVID-19 epidemic in France using a model with memory'. In: *Royal Society open science* 8.3 (2021), p. 202327. DOI: 10.1098/rsos.202327 (cited on pages 9, 162).

[32] Mircea T Sofonea, Bastien Reyné, Baptiste Elie, Ramsès Djidjou-Demasse, Christian Selinger, Yannis Michalakis, and Samuel Alizon. 'Memory is key in capturing COVID-19 epidemiological dynamics'. In: *Epidemics* 35 (2021), p. 100459. DOI: 10.1016/j.epidem.2021.100459 (cited on pages 9, 162).

[33] Åke Svensson. 'A note on generation times in epidemic models'. In: *Mathematical biosciences* 208.1 (2007), pp. 300–311. DOI: 10.1016/j.mbs.2006.10.010 (cited on page 9).

[34] Sang Woo Park, David Champredon, Joshua S Weitz, and Jonathan Dushoff. 'A practical generation-interval-based approach to inferring the strength of epidemics from their speed'. In: *Epidemics* 27 (2019), pp. 12–18. DOI: 10.1016/j.epidem.2018.12.002 (cited on pages 9, 10).

[35] Roy M Anderson and Robert M May. 'Coevolution of hosts and parasites'. In: *Parasitology* 85.2 (1982), pp. 411–426. DOI: 10.1017/S0031182000055360 (cited on pages 10, 16).

[36] Odo Diekmann, Johan Andre Peter Heesterbeek, and Johan Anton Jacob Metz. 'On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations'. In: *Journal of mathematical biology* 28 (1990), pp. 365–382 (cited on page 10).

[37] Odo Diekmann and Johan Andre Peter Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Vol. 5. John Wiley & Sons, 2000 (cited on page 10).

[38] Jonathan Dushoff and Sang Woo Park. 'Speed and strength of an epidemic intervention'. In: *Proceedings of the Royal Society B* 288.1947 (2021), p. 20201556. DOI: 10.1098/rspb.2020.1556 (cited on page 10).

[39] Maximilian M Nguyen, Ari S Freedman, Sinan A Ozbay, and Simon A Levin. 'Fundamental bound on epidemic overshoot in the SIR model'. In: *Journal of the Royal Society Interface* 20.209 (2023), p. 20230322. DOI: 10.1098/rsif.2023.0322 (cited on page 11).

[40] Sarah Cobey. 'Modeling infectious disease dynamics'. In: *Science* 368.6492 (2020), pp. 713–714. DOI: 10.1126/science.abb5659 (cited on page 11).

[41] Karline Soetaert, Thomas Petzoldt, and R Woodrow Setzer. 'Solving differential equations in R: package deSolve'. In: *Journal of statistical software* 33 (2010), pp. 1–25. DOI: 10.18637/jss.v033.i09 (cited on page 11).

[42] Selma Gago, Santiago F Elena, Ricardo Flores, and Rafael Sanjuán. 'Extremely high mutation rate of a hammerhead viroid'. In: *Science* 323.5919 (2009), pp. 1308–1308. DOI: 10.1126/science.1169202 (cited on page 12).

[43] Csaba Pal, María D Maciá, Antonio Oliver, Ira Schachar, and Angus Buckling. 'Coevolution with viruses drives the evolution of bacterial mutation rates'. In: *Nature* 450.7172 (2007), pp. 1079–1081. DOI: 10.1038/nature06350 (cited on page 12).

[44] Siobain Duffy, Laura A Shackelton, and Edward C Holmes. 'Rates of evolutionary change in viruses: patterns and determinants'. In: *Nature Reviews Genetics* 9.4 (2008), pp. 267–276. DOI: 10.1038/nrg2323 (cited on page 12).

[45] Everett Clinton Smith, Nicole R Sexton, and Mark R Denison. 'Thinking outside the triangle: replication fidelity of the largest RNA viruses'. In: *Annual review of virology* 1.1 (2014), pp. 111–132. DOI: 10.1146/annurev-virology-031413-085507 (cited on page 12).

[46] Motoo Kimura et al. 'Evolutionary rate at the molecular level'. In: *Nature* 217.5129 (1968), pp. 624–626. DOI: 10.1038/217624a0 (cited on page 12).

[47] Charles Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray, Albemarle Street., 1859 (cited on page 12).

[48] Andrew Gonzalez, Ophélie Ronce, Regis Ferriere, and Michael E Hochberg. 'Evolutionary rescue: an emerging focus at the intersection between ecology and evolution'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1610 (2013), p. 20120404. DOI: 10.1098/rstb.2012.0404 (cited on page 12).

[49] Sylvain Gandon, Michael E Hochberg, Robert D Holt, and Troy Day. 'What limits the evolutionary emergence of pathogens?' In: *Philosophical transactions of the Royal Society B: biological sciences* 368.1610 (2013), p. 20120086. DOI: 10.1098/rstb.2012.0086 (cited on page 12).

[50] Peter V Markov, Mahan Ghafari, Martin Beer, Katrina Lythgoe, Peter Simmonds, Nikolaos I Stilianakis, and Aris Katzourakis. 'The evolution of SARS-CoV-2'. In: *Nature Reviews Microbiology* 21.6 (2023), pp. 361–379. DOI: 10.1038/s41579-023-00878-2 (cited on pages 12, 13).

[51] Robert C Lacy. 'Loss of genetic diversity from managed populations: interacting effects of drift, mutation, immigration, selection, and population subdivision'. In: *Conservation biology* 1.2 (1987), pp. 143–158. DOI: 10.1111/j.1523-1739.1987.tb00023.x (cited on page 12).

[52] Sewall Wright. 'Classification of the factors of evolution.' In: *Cold Spring Harbor Symposia on Quantitative Biology* 20 (1955), pp. 16–24 (cited on page 13).

[53] Troy Day, Sylvain Gandon, Sébastien Lion, and Sarah P Otto. 'On the evolutionary epidemiology of SARS-CoV-2'. In: *Current Biology* 30.15 (2020), R849–R857. DOI: 10.5683/SP2/VKH3LE (cited on pages 13, 18, 21, 158).

[54] Sébastien Lion and Johan AJ Metz. 'Beyond R0 maximisation: on pathogen evolution and environmental dimensions'. In: *Trends in ecology & evolution* 33.6 (2018), pp. 458–473. DOI: 10.1016/j.tree.2018.02.004 (cited on pages 14, 15).

[55] Odo Diekmann. 'A beginner's guide to adaptive dynamics'. In: *Banach Center Publications* 63 (2004), pp. 47–86 (cited on page 15).

[56] Samuel Alizon, Amy Hurford, Nicole Mideo, and Minus Van Baalen. 'Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future'. In: *Journal of evolutionary biology* 22.2 (2009), pp. 245–259. DOI: 10.1111/j.1420-9101.2008.01658.x (cited on pages 15, 16).

[57] Troy Day and Stephen R Proulx. 'A general theory for the evolutionary dynamics of virulence'. In: *The American Naturalist* 163.4 (2004), E40–E63. DOI: 10.1086/382548 (cited on page 16).

[58] Samuel Alizon and Yannis Michalakis. 'Adaptive virulence evolution: the good old fitness-based approach'. In: *Trends in ecology & evolution* 30.5 (2015), pp. 248–254. DOI: 10.1016/j.tree.2015.02.009 (cited on page 16).

[59] Troy Day and Sylvain Gandon. 'Insights from Price's equation into evolutionary epidemiology'. In: *Disease evolution: models, concepts, and data analyses* 71 (2006), pp. 23–44. DOI: 10.1090/dimacs/071/02 (cited on pages 17, 22).

[60] Peter D Taylor and Leo B Jonker. 'Evolutionary stable strategies and game dynamics'. In: *Mathematical biosciences* 40.1-2 (1978), pp. 145–156. DOI: 10.1016/0025-5564(78)90077-9 (cited on page 17).

[61] Peter Schuster and Karl Sigmund. 'Replicator dynamics'. In: *Journal of theoretical biology* 100.3 (1983), pp. 533–538. DOI: 10.1016/0022-5193(83)90445-9 (cited on page 17).

[62] Troy Day and Sylvain Gandon. 'Applying population-genetic models in theoretical evolutionary epidemiology'. In: *Ecology Letters* 10.10 (2007), pp. 876–888. DOI: 10.1111/j.1461-0248.2007.01091.x (cited on page 18).

[63] Sylvain Gandon and Sébastien Lion. 'Targeted vaccination and the speed of SARS-CoV-2 adaptation'. In: *Proceedings of the National Academy of Sciences* 119.3 (2022), e2110666119. DOI: 10.1073/pnas.2110666119 (cited on pages 18, 21, 170).

[64] Luis-Miguel Chevin. 'On measuring selection in experimental evolution'. In: *Biology letters* 7.2 (2011), pp. 210–213. DOI: 10.1098/rsbl.2010.0580 (cited on page 18).

[65] Sarah P Otto, Troy Day, Julien Arino, Caroline Colijn, Jonathan Dushoff, Michael Li, Samir Mechai, Gary Van Domselaar, Jianhong Wu, David JD Earn, et al. 'The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic'. In: *Current Biology* 31.14 (2021), R918–R929. DOI: 10.1016/j.cub.2021.06.049 (cited on pages 18, 158, 159).

[66] Laura Boyle, Sofia Hletko, Jenny Huang, June Lee, Gaurav Pallod, Hwai-Ray Tung, and Richard Durrett. 'Selective sweeps in SARS-CoV-2 variant competition'. In: *Proceedings of the National Academy of Sciences* 119.47 (2022), e2213879119. DOI: 10.1073/pnas.2213879119 (cited on pages 18, 159).

[67] Erik Volz. 'Fitness, growth and transmissibility of SARS-CoV-2 genetic variants'. In: *Nature Reviews Genetics* 24.10 (2023), pp. 724–734. DOI: 10.1038/s41576-023-00610-z (cited on pages 18, 36, 159).

[68] Thomas W Berngruber, Rémy Froissart, Marc Choisy, and Sylvain Gandon. 'Evolution of virulence in emerging epidemics'. In: *PLoS pathogens* 9.3 (2013), e1003209. DOI: 10.1371/journal.ppat.1003209 (cited on pages 18, 35, 158, 164, 171, 172).

[69] Andrei Nikolaevich Tikhonov. 'Systems of differential equations containing small parameters in the derivatives. [In Russian]'. In: *Matematicheskii sbornik* 73.3 (1952), pp. 575–586 (cited on page 20).

[70] Sergio Rinaldi and Marten Scheffer. 'Geometric analysis of ecological models with slow and fast processes'. In: *Ecosystems* 3 (2000), pp. 507–521. DOI: 10.1007/s100210000045 (cited on page 20).

[71] Ferdinand Verhulst. 'Singular perturbation methods for slow–fast dynamics'. In: *Nonlinear Dynamics* 50 (2007), pp. 747–753. DOI: 10.1007/s11071-007-9236-z (cited on pages 20, 21).

[72] Erida Gjini and Sten Madec. 'A slow-fast dynamic decomposition links neutral and non-neutral coexistence in interacting multi-strain pathogens'. In: *Theoretical Ecology* 10 (2017), pp. 129–141. DOI: 10.1007/s12080-016-0320-1 (cited on pages 20, 21).

[73] Hildeberto Jardón-Kojakhmetov, Christian Kuehn, Andrea Pugliese, and Mattia Sensi. 'A geometric analysis of the SIR, SIRS and SIRWS epidemiological models'. In: *Nonlinear Analysis: Real World Applications* 58 (2021), p. 103220. DOI: 10.1016/j.nonrwa.2020.103220 (cited on pages 20, 21).

[74] Troy Day, David A Kennedy, Andrew F Read, and Sylvain Gandon. 'Pathogen evolution during vaccination campaigns'. In: *PLoS biology* 20.9 (2022), e3001804. DOI: 10.1371/journal.pbio.3001804 (cited on page 21).

[75] Sylvain Gandon. 'Why be temperate: lessons from bacteriophage λ'. In: *Trends in microbiology* 24.5 (2016), pp. 356–365. DOI: 10.1016/j.tim.2016.02.008 (cited on page 22).

[76] Sylvain Gandon. 'Evolution of multihost parasites'. In: *Evolution* 58.3 (2004), pp. 455–469. DOI: 10.1111/j.0014-3820.2004.tb01669.x (cited on page 22).

[77] Roland R Regoes, Martin A Nowak, and Sebastian Bonhoeffer. 'Evolution of virulence in a heterogeneous host population'. In: *Evolution* 54.1 (2000), pp. 64–71. DOI: 10.1111/j.0014-3820.2000.tb00008.x (cited on page 22).

[78] Sébastien Lion. 'Class structure, demography, and selection: reproductive-value weighting in nonequilibrium, polymorphic populations'. In: *The American Naturalist* 191.5 (2018), pp. 620–637. DOI: 10.1086/696976 (cited on pages 22, 160, 170).

[79] Odo Diekmann, JAP Heesterbeek, and Michael G Roberts. 'The construction of next-generation matrices for compartmental epidemic models'. In: *Journal of the royal society interface* 7.47 (2010), pp. 873–885. DOI: 10.1098/rsif.2009.0386 (cited on page 22).

[80] Emma B. Hodcroft. *CoVariants: SARS-CoV-2 Mutations and Variants of Interest*. 2021. URL: https://covariants.org/ (cited on page 23).

[81] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. 'Nextstrain: real-time tracking of pathogen evolution'. In: *Bioinformatics* 34.23 (2018), pp. 4121–4123. DOI: 10.1093/bioinformatics/bty407 (cited on page 23).

[82] Ronald A Fisher. 'On an absolute criterion for fitting frequency curves'. In: *Messenger of mathematics* 41 (1912), pp. 155–156 (cited on page 24).

[83] Ronald A Fisher. 'On the mathematical foundations of theoretical statistics'. In: *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222.594-604 (1922), pp. 309–368. DOI: 10.1098/rsta.1922.0009 (cited on page 24).

[84] Hirotugu Akaike. 'A new look at the statistical model identification'. In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705 (cited on page 25).

[85] Franz-Georg Wieland, Adrian L Hauber, Marcus Rosenblatt, Christian Tönsing, and Jens Timmer. 'On structural and practical identifiability'. In: *Current Opinion in Systems Biology* 25 (2021), pp. 60–69. DOI: 10.1016/j.coisb.2021.03.005 (cited on page 26).

[86] N Cunniffe, F Hamelin, A Iggidr, A Rapaport, and G Sallet. 'Identifiability and Observability in Epidemiological Models'. In: (2023) (cited on pages 26, 27).

[87] Claudio Cobelli and Joseph J Distefano III. 'Parameter and structural identifiability concepts and ambiguities: a critical review and analysis'. In: *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 239.1 (1980), R7–R24. DOI: 10.1152/ajpregu.1980.239.1.r7 (cited on page 26).

[88] Andreas Raue, Johan Karlsson, Maria Pia Saccomani, Mats Jirstrand, and Jens Timmer. 'Comparison of approaches for parameter identifiability analysis of biological systems'. In: *Bioinformatics* 30.10 (2014), pp. 1440–1448. DOI: 10.1093/bioinformatics/btu006 (cited on pages 26, 27).

[89] Giuseppina Bellu, Maria Pia Saccomani, Stefania Audoly, and Leontina D'Angiò. 'DAISY: A new software tool to test global identifiability of biological and physiological systems'. In: *Computer methods and programs in biomedicine* 88.1 (2007), pp. 52–61. DOI: 10.1016/j.cmpb.2007.07.002 (cited on page 26).

[90] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. 'Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood'. In: *Bioinformatics* 25.15 (2009), pp. 1923–1929. DOI: 10.1093/bioinformatics/btp358 (cited on pages 26, 27).

[91] F Hamelin, A Iggidr, A Rapaport, G Sallet, and M Souza. 'About the identifiability and observability of the SIR epidemic model with quarantine'. In: *IFAC-PapersOnLine* 56.2 (2023), pp. 4025–4030. DOI: 10.1016/j.ifacol.2023.10.1384 (cited on page 26).

[92] Necibe Tuncer and Trang T Le. 'Structural and practical identifiability analysis of outbreak models'. In: *Mathematical biosciences* 299 (2018), pp. 1–18. DOI: 10.1016/j.mbs.2018.02.004 (cited on pages 26, 27).

[93] John A Nelder and Roger Mead. 'A simplex method for function minimization'. In: *The computer journal* 7.4 (1965), pp. 308–313. DOI: 10.1093/comjnl/7.4.308 (cited on page 28).

[94] Bradley Efron. 'Bootstrap methods: Another look at the jackknife'. In: *The Annals of Statistics* 7.1 (1979), pp. 1–26. DOI: 10.1214/aos/1176344552 (cited on page 33).

[95] Regina Y Liu. 'Bootstrap procedures under some non-iid models'. In: *The annals of statistics* 16.4 (1988), pp. 1696–1708. DOI: 10.1214/aos/1176351062 (cited on page 33).

[96] Patrick Kline and Andres Santos. 'A score based approach to wild bootstrap inference'. In: *Journal of Econometric Methods* 1.1 (2012), pp. 23–41. DOI: 10.1515/2156-6674.1006 (cited on page 33).

[97] Peter Bühlmann. 'Sieve bootstrap for time series'. In: *Bernoulli* (1997), pp. 123–148. DOI: 10.2307/3318584 (cited on page 33).

[98] Gustavo Ulloa, Héctor Allende-Cid, and Héctor Allende. 'Sieve bootstrap prediction intervals for contaminated non-linear processes'. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I 18*. Springer. 2013, pp. 84–91 (cited on page 33).

[99] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, et al. 'A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)'. In: *Nature human behaviour* 5.4 (2021), pp. 529–538. DOI: 10.1038/s41562-021-01079-8 (cited on pages 34, 158, 170).

[100] Shrikanth Sampath, Anwar Khedr, Shahraz Qamar, Aysun Tekin, Romil Singh, Ronya Green, Rahul Kashyap, et al. 'Pandemics throughout the history'. In: *Cureus* 13.9 (2021). DOI: 10.7759/cureus.18136 (cited on page 157).

[101] Kate E Jones, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, and Peter Daszak. 'Global trends in emerging infectious diseases'. In: *Nature* 451.7181 (2008), pp. 990–993. DOI: 10.1038/nature06536 (cited on page 157).

[102] Lars Hufnagel, Dirk Brockmann, and Theo Geisel. 'Forecast and control of epidemics in a globalized world'. In: *Proceedings of the national academy of sciences* 101.42 (2004), pp. 15124–15129. DOI: 10.1073/pnas.0308344101 (cited on page 157).

[103] William E Diehl, Aaron E Lin, Nathan D Grubaugh, Luiz Max Carvalho, Kyusik Kim, Pyae Phyo Kyawe, Sean M McCauley, Elisa Donnard, Alper Kucukural, Patrick McDonel, et al. 'Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic'. In: *Cell* 167.4 (2016), pp. 1088–1098. DOI: 10.1016/j.cell.2016.10.014 (cited on page 157).

[104] Richard A Urbanowicz, C Patrick McClure, Anavaj Sakuntabhai, Amadou A Sall, Gary Kobinger, Marcel A Müller, Edward C Holmes, Félix A Rey, Etienne Simon-Loriere, and Jonathan K Ball. 'Human adaptation of Ebola virus during the West African outbreak'. In: *Cell* 167.4 (2016), pp. 1079–1087. DOI: 10.1016/j.cell.2016.10.013 (cited on page 157).

[105] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. 'Unifying the epidemiological and evolutionary dynamics of pathogens'. In: *science* 303.5656 (2004), pp. 327–332. DOI: 10.1126/science.1090727 (cited on page 159).

[106] Erik M Volz, Katia Koelle, and Trevor Bedford. 'Viral phylodynamics'. In: *PLoS computational biology* 9.3 (2013), e1002947. DOI: 10.1371/journal.pcbi.1002947 (cited on page 159).

[107] Oliver G Pybus and Andrew Rambaut. 'Evolutionary analysis of the dynamics of viral infectious disease'. In: *Nature Reviews Genetics* 10.8 (2009), pp. 540–550. DOI: 10.1038/nrg2583 (cited on page 159).

[108] Stephen W Attwood, Sarah C Hill, David M Aanensen, Thomas R Connor, and Oliver G Pybus. 'Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic'. In: *Nature Reviews Genetics* 23.9 (2022), pp. 547–562. DOI: 10.1038/s41576-022-00483-8 (cited on page 159).

[109] Samuel Alizon. 'Phylodynamique'. In: *Modèles et méthodes pour l'évolution biologique* (2022). DOI: 10.51926/ISTE.9069.ch11 (cited on page 159).

[110] Sébastien Lion and Sylvain Gandon. 'Evolution of class-structured populations in periodic environments'. In: *Evolution* 76.8 (2022), pp. 1674–1688. DOI: 10.1111/evo.14522 (cited on pages 160, 170).

[111] George EP Box. 'Science and statistics'. In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799 (cited on page 161).

[112] François Blanquart, Thomas Berngruber, Marc Choisy, and Sylvain Gandon. 'Evolution of virulence in emerging epidemics: inference from an evolution experiment'. In: *bioRxiv* (2020), pp. 2020–08. DOI: 10.1101/2020.08.19.256917 (cited on page 161).

[113] Daniel T Gillespie. 'A general method for numerically simulating the stochastic time evolution of coupled chemical reactions'. In: *Journal of computational physics* 22.4 (1976), pp. 403–434. DOI: 10.1016/0021-9991(76)90041-3 (cited on page 161).

[114] Daniel T Gillespie. 'Exact stochastic simulation of coupled chemical reactions'. In: *The journal of physical chemistry* 81.25 (1977), pp. 2340–2361. DOI: 10.1021/j100540a008 (cited on page 161).

[115] Daniel T Gillespie. 'Approximate accelerated stochastic simulation of chemically reacting systems'. In: *The Journal of chemical physics* 115.4 (2001), pp. 1716–1733. DOI: 10.1063/1.1378322 (cited on page 161).

[116] Todd L Parsons, Amaury Lambert, Troy Day, and Sylvain Gandon. 'Pathogen evolution in finite populations: slow and steady spreads the best'. In: *Journal of The Royal Society Interface* 15.147 (2018), p. 20180135. DOI: 10.1098/rsif.2018.0135 (cited on page 161).

[117] Troy Day, Todd Parsons, Amaury Lambert, and Sylvain Gandon. 'The Price equation and evolutionary epidemiology'. In: *Philosophical Transactions of the Royal Society B* 375.1797 (2020), p. 20190357. DOI: 10.1098/rstb.2019.0357 (cited on page 161).

[118] Aaron A King, Dao Nguyen, and Edward L Ionides. 'Statistical inference for partially observed Markov processes via the R package pomp'. In: *Journal of Statistical Software* 69.12 (2015), pp. 1–43. DOI: 10.18637/jss.v069.i12 (cited on page 161).

[119] Edward L Ionides, Carles Bretó, and Aaron A King. 'Inference for nonlinear dynamical systems'. In: *Proceedings of the National Academy of Sciences* 103.49 (2006), pp. 18438–18443. DOI: 10.1073/pnas.0603181103 (cited on page 161).

[120] Carles Bretó, Daihai He, Edward L Ionides, and Aaron A King. 'Time series analysis via mechanistic models'. In: *The Annals of Applied Statistics* (2009), pp. 319–348. DOI: 10.1214/08-AOAS201 (cited on page 161).

[121] François Blanquart, Nathanaël Hozé, Benjamin John Cowling, Florence Débarre, and Simon Cauchemez. 'Selection for infectivity profiles in slow and fast epidemics, and the rise of SARS-CoV-2 variants'. In: *Elife* 11 (2022), e75791. DOI: 10.7554/eLife.75791 (cited on page 162).

[122] Sang Woo Park, Benjamin M Bolker, Sebastian Funk, C Jessica E Metcalf, Joshua S Weitz, Bryan T Grenfell, and Jonathan Dushoff. 'The importance of the generation interval in investigating dynamics and control of new SARS-CoV-2 variants'. In: *Journal of The Royal Society Interface* 19.191 (2022), p. 20220173. DOI: 10.1098/rsif.2022.0173 (cited on page 162).

[123] Bastien Reyné, Quentin Richard, Christian Selinger, Mircea T Sofonea, Ramsès Djidjou-Demasse, and Samuel Alizon. 'Non-Markovian modelling highlights the importance of age structure on Covid-19 epidemiological dynamics'. In: *Mathematical Modelling of Natural Phenomena* 17 (2022), p. 7. DOI: 10.1051/mmnp/2022008 (cited on page 162).

[124] William Ogilvy Kermack and Anderson G McKendrick. 'Contributions to the mathematical theory of epidemics. II.—The problem of endemicity'. In: *Proceedings of the Royal Society of London. Series A, containing papers of a mathematical and physical character* 138.834 (1932), pp. 55–83. DOI: 10.1098/rspa.1932.0171 (cited on page 162).

[125] William Ogilvy Kermack and Anderson G McKendrick. 'Contributions to the mathematical theory of epidemics. III.—Further studies of the problem of endemicity'. In: *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 141.843 (1933), pp. 94–122. DOI: 10.1098/rspa.1933.0106 (cited on page 162).

[126] Tjibbe Donker, Alexis Papathanassopoulos, Hiren Ghosh, Raisa Kociurzynski, Marius Felder, Hajo Grundmann, and Sandra Reuter. 'Estimation of SARS-CoV-2 fitness gains from genomic surveillance data without prior lineage classification'. In: *Proceedings of the National Academy of Sciences* 121.25 (2024), e2314262121. DOI: 10.1073/pnas.2314262121 (cited on page 163).

[127] Sunetra Gupta, Martin CJ Maiden, Ian M Feavers, Sean Nee, Robert M May, and Roy M Anderson. 'The maintenance of strain structure in populations of recombining infectious agents'. In: *Nature medicine* 2.4 (1996), pp. 437–442. DOI: 10.1038/nm0496-437 (cited on page 163).

[128] Julia R Gog and Bryan T Grenfell. 'Dynamics and selection of many-strain pathogens'. In: *Proceedings of the National Academy of Sciences* 99.26 (2002), pp. 17209–17214. DOI: 10.1073/pnas.252512799 (cited on page 163).

[129] Troy Day and Sylvain Gandon. 'The evolutionary epidemiology of multilocus drug resistance'. In: *Evolution* 66.5 (2012), pp. 1582–1597. DOI: 10.1111/j.1558-5646.2011.01533.x (cited on page 163).

[130] David V McLeod and Sylvain Gandon. 'Effects of epistasis and recombination between vaccine-escape and virulence alleles on the dynamics of pathogen adaptation'. In: *Nature ecology & evolution* 6.6 (2022), pp. 786–793. DOI: 10.1038/s41559-022-01709-y (cited on page 163).

[131] Sylvain Gandon, Margaret J Mackinnon, Sean Nee, and Andrew F Read. 'Imperfect vaccines and the evolution of pathogen virulence'. In: *Nature* 414.6865 (2001), pp. 751–756. DOI: 10.1038/414751a (cited on page 163).

[132] Alicia Walter and Sébastien Lion. 'Epidemiological and evolutionary consequences of periodicity in treatment coverage'. In: *Proceedings of the Royal Society B* 288.1946 (2021), p. 20203007. DOI: 10.1098/rspb.2020.3007 (cited on page 163).

[133] Claude Hannoun. 'The evolving history of influenza viruses and influenza vaccines'. In: *Expert review of vaccines* 12.9 (2013), pp. 1085–1094. DOI: 10.1586/14760584.2013.824709 (cited on page 163).

[134] Robert S Paton, Christopher E Overton, and Thomas Ward. 'The rapid replacement of the SARS-CoV-2 Delta variant by Omicron (B.1.1.529) in England'. In: *Science Translational Medicine* 14.652 (2022), eabo5395. DOI: 10.1126/scitranslmed.abo5395 (cited on page 163).

[135] Martin Guillemet, Hélène Chabas, Antoine Nicot, François Gatchich, Enrique Ortega-Abboud, Cornelia Buus, Lotte Hindhede, Geneviève M Rousseau, Thomas Bataillon, Sylvain Moineau, et al. 'Competition and coevolution drive the evolution and the diversification of CRISPR immunity'. In: *Nature Ecology & Evolution* 6.10 (2022), pp. 1480–1488. DOI: 10.1038/s41559-022-01841-9 (cited on page 164).

[136] Jianyu Lai, Kristen K Coleman, S-H Sheldon Tai, Jennifer German, Filbert Hong, Barbara Albert, Yi Esparza, Aditya K Srikakulapu, Maria Schanz, Isabel Sierra Maldonado, et al. 'Evolution of SARS-CoV-2 shedding in exhaled breath aerosols'. In: *Clinical Infectious Diseases* 76.5 (2023), pp. 786–794. DOI: 10.1093/cid/ciac846 (cited on page 170).