Reconstruction of the transmission of a virus during an epidemic by statistical learning on genomic data

Maryam Alamil

Aix*Marseille université
Initiative d'excellence

NNT/NL : 2020AIXM0001/001ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université

le 11 décembre 2020 par

Maryam ALAMIL

Reconstruction des transmissions d'un virus au cours d'une épidémie par apprentissage statistique sur données génomiques

Discipline
Mathématique appliquées

Spécialité
Biostatistiques

École doctorale
ED 184 Mathématiques et Informatique

Laboratoire/Partenaires de recherche
INRAE - PACA
Biostatistique et Processus Spatiaux
(UR 546, BioSP)

Biostatistique BIOSP & Processus Spatiaux INRAE

AGENCE NATIONALE DE LA RECHERCHE ANR

AVIGNON UNIVERSITÉ

Composition du jury

Samuel Alizon Directeur de recherche MIVEGEC (CNRS, IRD, UM), Montpellier	Rapporteur
Pierre Nicolas Directeur de recherche MaIAGE, INRAE, Jouy-en-Josas	Rapporteur
Céline Scornavacca Directrice de recherche CNRS, ISEM, Université de Montpellier	Présidente
Virginie Ravigné Chercheuse CIRAD, PVBMT, La Réunion	Examinatrice
Samuel Soubeyrand Directeur de recherche, BioSP, INRAE , Avignon	Directeur de thèse
Gaël Thébaud Chargé de recherche UMR, BGPI, CIRAD, INRAE, Montpellier	Co-directeur de thèse

Je soussigné, Maryam Alamil, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Samuel Soubeyrand et Gaël Thébaud, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisées dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Avignon le 15 octobre 2020

Abstract

Pathogens can cause epidemics of high impact in developing and developed countries. To protect populations against pathogens particularly the viral ones that can cause rapid death and that could mutate to more aggressive variants, it is crucial to predict and control their propagation. To cope with this issue, whats is needed is to better understand how pathogens spread within a host (e.g individual, households, fields) or between hosts. The answer to the question "how do pathogens spread?" may lie in determining who infected whom or who is closely related to whom, that statistically means inferring epidemiological links between hosts.

With the aim of estimating epidemiological links, several empirical and model-based approaches have been developed. Approaches exploiting pathogen sequence data are the most advantageous because they inform which hosts contain pathogen variants that are most closely related to each other. In this thesis, we investigate an alternative approach grounded on statistical learning and based on a semi-parametric pseudo-evolutionary model. This pseudo-model describes transitions between sets of sequences observed from an infected host and its putative sources. And, our approach consists on using this model applied to training data (e.g contact tracing) to learn the structure of epidemiological links and therefore to infer links for the whole dataset. The efficiency of our approach aiming at inferring transmission links of infectious diseases, is assessed by applying it to three different real cases in animal, human and plant epidemics. Then, we applied it to simulated data generated with diverse models for the viral evolution and transmission, performing a sensitivity analysis of the relationship between the accuracy of our approach and the demo-genetic factors that may impact the virus evolution and transmission.

Such innovate approach has the potential to be particularly valuable in the case of a risk of erroneous mechanistic assumptions and sequencing errors, it is adaptable to very different contexts from animal, human and plant epidemics, and it is sufficiently parsimonious to allow handling big data sets in the future. This approach is able to make notable advances in the field of computational biology and quantitative molecular epidemiology. This leads to identify more precisely the epidemiological links, gives better insights into risk factors playing a role in pathogens spread within or between hosts and, consequently, fosters a better understanding of infectious diseases to build robust forward projections and design control policies.

Keywords: pathogen spread, infectious disease, semi-parametric model, pseudo-likelihood, learning, genomic data, within-host pathogen diversity, training data, viral kinetic models, substitutions models, transmission dynamics

Résumé

Afin de prédire et contrôler plus efficacement la propagation des maladies infectieuses, nous devons mieux comprendre comment les agents pathogènes se propagent dans et entre les populations hôtes. La question " comment les agents pathogènes se propagent-ils? " peut être comprise de plusieurs façons. Dans ma thèse, je considère des situations où des unités hôtes, infectées par une maladie infectieuse au cours d'une fenêtre temporelle spécifique, sont identifiées comme infectées et par la suite caractérisées, et la question mentionnée ci-dessus est spécifiée en ces termes : " qui a infecté qui? " ou, plus généralement, " qui est étroitement lié à qui? " dans la dynamique de transmission de la maladie. Les unités d'accueil désignent généralement des individus mais peuvent également désigner des groupes tels que les ménages, les fermes et les parcelles agricoles.
Pour les agents pathogènes à évolution rapide, de nombreuses approches (empiriques ou fondées sur des modèles) utilisent les données de séquence des agents pathogènes pour inférer qui a infecté qui. Ces données renseignent sur les hôtes porteurs de variants de pathogènes étroitement liés les uns aux autres. Chronologiquement, les premières approches évoquées ci-dessus exploitaient principalement une seule séquence pathogène par hôte. Néanmoins, le progrès des techniques de séquençage, révélant le polymorphisme génétique intra-hôte des pathogènes, a favorisé le développement d'approches tenant compte de la génération de la diversité intra-hôte et/ou tirant parti des informations fournies par des ensembles de séquences échantillonnées sur les hôtes. Ma thèse concerne précisément une telle approche visant à estimer des liens de transmission à partir de données de séquençage haut-débit collectées sur plusieurs unités hôtes et reflétant le polymorphisme intra-hôte du virus d'intérêt. L'approche que je propose est essentiellement fondée sur un modèle semi-paramétrique et

pseudo-évolutionniste, une technique d'apprentissage statistique et une quantité limitée de données d'entrainement. Le modèle permet de calculer des mesures de pseudo-vraisemblance des transitions entre des ensembles de séquences observées chez l'unité hôte infectée et chez ses sources putatives. Il est calibré sur les données d'entrainement pour apprendre la structure des liens épidémiologiques réels puis appliqué à l'ensemble de données pour inférer des liens entre toutes les unités hôtes du jeu de données.
Au-delà du développement de l'approche brièvement décrite ci-dessus, je présente son application à des données réelles en santé humaine, animale et végétale (en l'occurrence, à des données concernant Ebola, la grippe porcine, la grippe équine et un potyvirus inféodé aux plantes), ainsi qu'à des données simulées. Les données simulées sont obtenues avec un modèle original que j'ai développé et qui permet la génération de dynamiques démo-génétiques hors équilibre et à variations rapides pour la population virale intra-hôte. Après avoir caractérisé numériquement la capacité de ce modèle à générer une diversité d'agents pathogènes intra-hôte, il est utilisé pour simuler des scénarios démo-génétiques significativement divers, scénarios auxquels l'approche d'estimation des liens épidémiologiques est appliquée. En utilisant ce modèle, j'ai effectué une analyse de sensibilité formelle de la relation entre la performance de notre approche d'inférence et les facteurs démo-génétiques qui peuvent avoir un impact sur l'évolution, la diversité et la transmission du virus.
D'un point de vue général, l'approche proposée ouvre la voie à l'utilisation de l'apprentissage statistique dans la reconstruction des transmissions et des liens épidémiologiques à partir de données génomiques. Elle pourrait contribuer à améliorer la compréhension des facteurs de risque jouant un rôle dans la propagation des agents pathogènes au sein des populations hôtes et, par conséquent, à favoriser une meilleure compréhension de la dynamique des maladies infectieuses ainsi que la conception de projections prévisionnelles et de stratégies de contrôle robustes.

Mots clés: Apprentissage statistique, diversité intra-hôte des pathogènes, données d'apprentissage, données génomiques, dynamique de transmission, épidémiologie moléculaire, maladies infectieuses, modèles cinétiques viraux, modèles de substitution, modèle semi-paramétrique, propagation des agents pathogènes, pseudo-vraisemblance.

Acknowledgment

"Success is for those that presses on when they see a wall, find ways around a road block and consistently push boundaries and conquer territories against all odds." *~ Oscar Bimpong*

Against all odds, against Covid-19 sanitary crisis, I have managed to complete my thesis. Indeed, the work achieved in this thesis could have never been done without the constant support of several people whose good humour and interest in my research allowed me to progress.

I would like to express my heartiest gratitude and sincere thanks to my supervisor Samuel Soubeyrand, who was consistently providing me with valuable guidance and indispensable suggestions throughout my thesis. Sam, despite your busy schedule, you have managed to find the time for answering my questions, for advising and encouraging me forward with great patience and kindness.

I would like to address my warm gratitude for Gaël Thébaud for being a great co-supervisor of this thesis. Thank you for your kindness, for your precious advices and comments. Thank you for sharing your knowledge in molecular biology with such patience and passion.

My most sincere thanks goes to the reporters Samuel Alizon and Pierre Nicolas for both approving to report my thesis and giving their attention to my work. I would also like to thank Céline Scornavacca and Virginie Ravigné for participating in the jury of my thesis.

Many thanks to Karine Berthier, Mélina Ribaud, Claude Bruchou and Joseph Hughes for the precious help they gave me on genetic diversity assessment, transmission inference and sensitivity analysis techniques.

I wish to express my appreciation and respect for BioSP staff. In this unit, I found the family atmosphere in which one feels supported by everyone. I sincerely

thank each of you. Unfortunately I can't mention everyone so I'll just make a few specific mentions. So I begin by thanking my french mom Sylvie Jouslin for her kindness, her listening, her efficiency and her unlimited support, care and love. I would like to particularly thank Loic Houde for his efforts in assisting all the technical troubleshooting we have faced. I know, it was difficult to deal with all the problems, especially during the lockdown, so thank you very much. Special thanks to Emilie, Marine, Mélina, Patrizia, Florian Patout and Florent for their good mood, their sympathy and the serious and funny discussions we had especially during the lunch break. Finally, thanks to the Lebanese colleagues, Candy and Maria, for their support and encouragement throughout the three years of my thesis.

I owe a deep sense of gratitude to the family who took care of me during the worst times of my life, family Jawhar: Marie-claire, Bassel, Joïa and Joey. Joïa, you are the first little baby girl who stole my heart, the second girl will be surely my daughter. My Lyonnaise family: Christine, Hanine, Jessy, Lea, Perla, Rita, Anthony, François, Joe, Joseph and Michel, you showed me the true meaning of friendship, the true colors of caring. I love you and I appreciate every moment with you. Thanks to Sahmout-Guérin Family with whom the week-ends in Avignon were special.

This thesis would not have been successful without my friends. Hanine, my caring neighbor, our friendship has gone beyond all stages and so you became a part of my family, I am greatly thankful for all the joyful moments we have shared throughout these three years. Thank you for always being by my side. You're going back to Lebanon in a month and it will be the hardest goodbye in my life, I will miss you so much. Pamela Alalam, being thankful isn't enough, you were always the one who tolerates my mood swings and tried his best to relief and motivate me. I am sure that our friendship will last for life. Gloria, despite what we have been through, I am convinced that our friendship is hard to be broken and will always be connected by heart. Thank you for the good mornings we have shared together, for listening to me and giving me hope to always head forward. Nicole, my beautiful cousin, thank you for always being by my side. Despite the distance, you were the first person I call every time I had the slightest concern. Sarah, you are the person who makes me laugh, the thing I like the most. Ali, Charbel, Daher, Georges, Hussein and Sarkis, I will always treasure the beautiful moments that we spent together in Avignon. I thank Pamela El Hajj and Patricia for the wonderful adventures we had in Montpellier, Reims and Paris. For all my friends in Lebanon, Anna, Carla, Denise, Diana, Liliane, Nathalie, Sophie and Therese, I keep you in my heart and thoughts.

For the new born babies, Ayla and Micho, you gave me the motivation and the happiness during the dissertation writing period, without knowing that, I love you so much.

Of course, this acknowledgment would not be complete without thanking my parents Therese and François, my sister Issamar and my brother Joseph, who supported and encouraged me throughout my entire studies. Thanks to your loving support and faith, I was able to finish my study. I would also like to thank Massaad Family for their support and patience, and not to forget my aunts, uncles and cousins.

Issamar, Lea, Marie-rose, François and Georges, thanks for your criticism and feedback on the draft versions of this thesis.

My final acknowledgement is addressed to my love, my other half and my backbone, Georges, with whom I look forward to spend the rest of my life. Georges, I just want to tell you that this wouldn't have been possible without you and that the future is ours.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. General overview of my thesis

Fast-evolving pathogens such as RNA viruses can cause epidemics of high impact in developing and developed countries alike. Some of these viruses have major impact not only on human species, but also on animal and plant species. Significant global expenditures for virus prevention have been incurred in previous years. For instance, to date, more than 200 countries are affected by the COVID-19 infecting over 26 millions people and causing the death of at least 864,000 people as reported by the World Health Organization (WHO, September 2020). A recent estimation published by the European Centre for Disease Prevention and Control (ECDC) indicates that Europe experienced 10,705 laboratory-confirmed hospitalised influenza cases during the 2019–2020 influenza season. According to the WHO, rabies and hepatitis E are estimated to cause respectively 55,000 and 57,000 individual deaths per year. The total global expenditure for rabies prevention was assessed to be more than US$1 billion annually. During the 2001 outbreak of foot-and-mouth disease in Great Britain, 6 million animals were culled (Anderson et al., 1996; Haydon et al., 2003). Much like diseases of humans and other animals, plant diseases such as Sharka can give rise to severe consequences while damaging vegetation and crops, reducing yields and raising the treatment and prevention expenditures. The management of sharka disease on Prunus trees remains a challenge and this plant disease induced an overall cost above €10 billion at the global scale over three decades (Cambra et al., 2006; Rimbaud et al., 2015). In order to most effectively predict and control the spread of such

infectious diseases, we need to better understand how pathogens spread within- and between-host populations and what is the role of the environment in the transmissions. The question "How pathogens spread?" being a broad question with several meanings, it is restricted in my PhD work into "Who infected whom?" or "Who is closely related to whom?" in the disease dynamics.

To unravel disease transmission links between hosts and thus to plan and develop effective strategies for controlling infectious diseases, various mathematical models based on epidemiological data have been used (e.g. Cauchemez and Ferguson, 2012; Cauchemez et al., 2006, 2016; Haydon et al., 2003; Heijne et al., 2012; Kao, 2002). Typical epidemiological data exploited in this context, and collected during infectious disease outbreaks, are data on the timing of symptoms, contact tracing and surveillance effort. Recently, for fast-evolving pathogens, several frameworks have been proposed to integrate the analysis of pathogen sequence data in virus transmission studies (Campbell et al., 2018, 2019; Cottam et al., 2008; De Maio et al., 2016, 2018; Didelot et al., 2014, 2017; Hall et al., 2015; Hughes et al., 2012; Jombart et al., 2011, 2014; Leavitt et al., 2020; Morelli et al., 2012; Skums et al., 2018a; Worby et al., 2016, 2017; Wymant et al., 2018; Ypma et al., 2012, 2013b). Indeed, genetic data enable identifying hosts containing pathogen variants that are most closely related to each other and therefore potentially linked with more or less direct transmissions.

Numerous methods have been proposed to reconstruct transmission links from genetic data using one pathogen sequence per infected host, typically the consensus sequence or the majority sequence (Campbell et al., 2018, 2019; Cottam et al., 2008; Jombart et al., 2011, 2014; Morelli et al., 2012; Worby et al., 2016, 2017; Ypma et al., 2012, 2013b). Using a single variant per host limits the amount of information about the viral composition within the host although recent sequencing techniques, such as *deep Sanger sequencing* (DSS) and *high-throughput sequencing* or *next-generation sequencing* (HTS or NGS) open new routes to access a detailed description of the genetic variation that can exist within a host during an infection (Alizon et al., 2011; Gire et al., 2014; Lauck et al., 2012; Murcia et al., 2010, 2012; Nelson and Hughes, 2015; Worby et al., 2014, 2017; Wright et al., 2011). Thanks to these techniques providing a subsample of the pathogen variants in the host at the sampling time, it is now recognized that the virus genetic diversity may vary spatially and temporally during the infection, due to several demographic and genetic factors (e.g mutation, selection and genetic drift processes) acting at the within-host scale (Abel et al., 2015; Alizon et al., 2011; Cuevas et al., 2015; Poirier and Vignuzzi, 2017; Pybus and Rambaut, 2009; Simmons et al., 2012). Vari-

ations in the within-host genetic diversity may then affect between-hosts virus transmissions (Abel et al., 2015; Worby et al., 2014). Offering the possibility to assess within-host pathogen diversity, such sequencing techniques fostered the development of model-based approaches exploiting, in one way or another, the degree of genetic similarity between viral variants present within the different hosts to identify linked hosts and infer transmissions (De Maio et al., 2016, 2018; Didelot et al., 2014, 2017; Hall et al., 2015; Skums et al., 2018a).

To infer epidemiological links between host units, it first seems appropriate and natural to adopt some approaches grounded on a mechanistic vision of transmission and micro-evolutionary processes. Indeed, such approaches are underlined by mechanistic assumptions acting as relevant constraints, which are expected to guide the inference. However, implementing mechanistic approaches taking into account within-host diversity is today relatively complicated and potentially misleading because of the complexity of necessary mechanistic assumptions, in particular if one has to handle sequence data that do not accurately reflect the within-host pathogen population because of sequencing bias or errors. In this context, my research aims at investigating a modelling and statistical approach to quickly and robustly infer epidemiological links of infectious fast-evolving pathogens from deep sequencing data. Thus, I developed a method called SLAFEEL (Statistical Learning Approach For Estimating Epidemiological Links), which is based on a pseudo-mechanistic model and on statistical learning (Friedman et al., 2001; James et al., 2013). The overall concept of SLAFEEL is to learn the epidemiological links structure with a pseudo-evolutionary model applied to training data, and then to use this initial training stage for the inference of the links for the whole data set. For limiting computational burden, the pseudo-evolutionary model concisely describes transitions between sets of sequences sampled at different times from a host unit and its putative sources. Training data are used to replace mechanistic hypotheses for constraining the inference and typically consist of classical contact information such as contact tracing, or proxies of contact information such as geographical distances between host units.

To test SLAFEEL approach, I applied it to real cases in animal, human and plant epidemiology. These case studies concern respectively influenza A viruses sampled from animal populations (Hughes et al., 2012; Murcia et al., 2012), Ebola virus sampled from a human population (Gire et al., 2014) and viruses sampled from wild and cultivated plant populations (Desbiez et al., 2017). Datasets exploited in these studies enable us to deal with diverse epidemiological situations and sequencing procedures and therefore to assess SLAFEEL performance. Re-

sults show that our approach is adaptable to very different contexts and data and achieves an encouraging performance level.

Furthermore, to calibrate and validate the efficiency of SLAFEEL, I applied it to simulated data, performing a global sensitivity analysis of the relationship between SLAFEEL accuracy and the demo-genetic factors that may impact the virus evolution and consequently the transmissions inference. Simulated data are generated under multiple demographic and genetic settings (e.g low/high evolutionary rate of the rate of evolution of viral genetic sequences, low/high fitness differences between viral variants, low/high bottleneck strength, ...), with a stochastic model for the evolution and transmission of populations of sequences.

In recent years, several methods (often implemented in software packages) have emerged to explicitly model the pathogen evolution and transmission (Campbell et al., 2018; De Maio et al., 2016; Didelot et al., 2017; Jombart et al., 2011, 2014; Klinkenberg et al., 2017; Stadler and Bonhoeffer, 2013; Worby and Read, 2015; Worby et al., 2016). These methods differ in their underlying genetic models (e.g phylogenetic (Campbell et al., 2018; Didelot et al., 2017; Klinkenberg et al., 2017; Stadler and Bonhoeffer, 2013) or non-phylogenetic models (Jombart et al., 2011, 2014; Worby and Read, 2015; Worby et al., 2016)) and epidemiological models (e.g compartmental models (Worby and Read, 2015; Worby et al., 2016) or branching process models (Didelot et al., 2014; Jombart et al., 2011; Klinkenberg et al., 2017) as well as in their ability to account for the within-host diversity. Most of the methods accounting for the within-host genetic diversity (Campbell et al., 2018; Didelot et al., 2017; Klinkenberg et al., 2017) are based on within-host phylogenetic tree while describing a linear growing for the within-host pathogen population size. In contrast, the method of Worby and Read (2015) gives the opportunity to generate, in a forward framework implemented in the SEEDY package, the pathogen evolution within each infected host providing, at each time, the within-host viral composition. Worby and Read assumed that the size of the pathogen population converges to an attraction function via the sum of binomial jumps, and the pathogen population varies within a host during an infection due to several demographic and genetic factors such as the inoculum size, mutations, natural selection and random genetic drift. Based on its construction, SEEDY can be used to simulate non-equilibrium pathogen populations, which is of particular interest for testing, in complex situations, the performance of methods for reconstructing transmissions. However, SEEDY does not enable an exact control of the pathogen population size and is not clearly adapted to produce fast changes in the dominant pathogen genotypes within the host. In this respect, using as a

foundation the work by Worby and Read (2015), I developed a versatile within-host pathogen evolution model that allows me to simulate fast changes in the genetic composition of the virus population, control exactly the temporal variation in the population size, and hence provide sequences and frequencies of variants across time under very diverse demo-genetic conditions. Demographic effects are considered first by initiating the infection of a host by single or multiple variants and, second according to a viral demographic kinetic model (e.g., grounded on a set of differential equations) used to quantify the temporal variation of the viral load during an infection (Baccam et al., 2006; Beauchemin and Handel, 2011; Beauchemin et al., 2008; Canini and Perelson, 2014; Handel et al., 2010; Nowak and May, 2000; Pawelek et al., 2012; Saenz et al., 2010; Smith and Perelson, 2011). Genetic effects correspond to the mutation and replication processes subjected to natural selection and random genetic drift. The fluctuations in variant frequencies induced by the two latter phenomena can be reinforced with a shuffling process enabling over-dispersion with respect to classical multinomial draws.

I embedded the within-host evolutionary model into a host-to-host transmission model depending on a contact process. In this stochastic model I assume that: (i) the outbreak starts with one infected host, (ii) all hosts have equal contact probability, (iii) the hosts in contact with an infectious host and the contact times are randomly drawn, (iv) the success of transmission from an infectious host to an exposed host depends on the viral load within the infectious host at the contact time and (v) a subsample of sequences within the infecting host at the transmission time is transmitted to the newly infected host characterizing its initial viral kinetic and composition state.

As briefly mentioned above, a sensitivity analysis of SLAFEEL performance has been carried out to map the performance with respect to the variations of the input demo-genetic factors. The sensitivity analysis allows the identification of factors that exert an influence on the inference accuracy, taking into account the interactions between these factors. Here, I conduct a variance-based sensitivity analysis computing the first-order and the total effect indices of each factor (Saltelli et al., 2000, 2008). The first-order sensitivity index of a factor is a measure of the main effect of this factor on the performance of SLAFEEL. In contrast, the total index of a factor corresponds to a measure that captures its overall influence on the efficiency of SLAFEEL, including its interactions with the other factors. Precisely estimating the sensitivity indices for the stochastic process that we consider requires numerous repetitions of the simulation and inference process that we consider.

On the basis of my work, one of the future challenges is to extend the application range of SLAFEEL with the potential of being able to handle big sequence data corresponding for instance to large sequencing depths, sequence lengths and host numbers, and to take into account the environmental factors that prevent or enhance disease transmission. This should lead to (i) more accurate inferences of transmission links, (ii) a better understanding of risk factors playing a role in disease evolution and transmission, (iii) better forward projections, and (iv) the development of tools to foster the use of big data generated in molecular ecology and epidemiology.

1.2. Research questions

In brief, my work is an attempt to answer the following question:

- How can statistical learning ideas be adapted for inferring epidemiological links between hosts infected by a virus from data reflecting the within-host diversity of the pathogen?

To answer this question, I propose the SLAFEEL approach exploiting partial contact information as learning data used for calibrating an original pseudo-evolutionary model and allowing the inference of epidemiological links. This proposal leads to the following technical question:

- How robust, effective and versatile is this inference approach?

This question is investigated with real data using a cross-validation technique and simulated data using sensitivity analysis. Simulated data are generated with a new flexible stochastic model allowing me to consider specific demo-genetic and sampling settings. The behaviour of this model is characterized by addressing the following question:

- What sort of within-host pathogen diversity can be generated by the stochastic model proposed for simulating data?

Figure 1.1 provides a summary diagram of my thesis objectives.

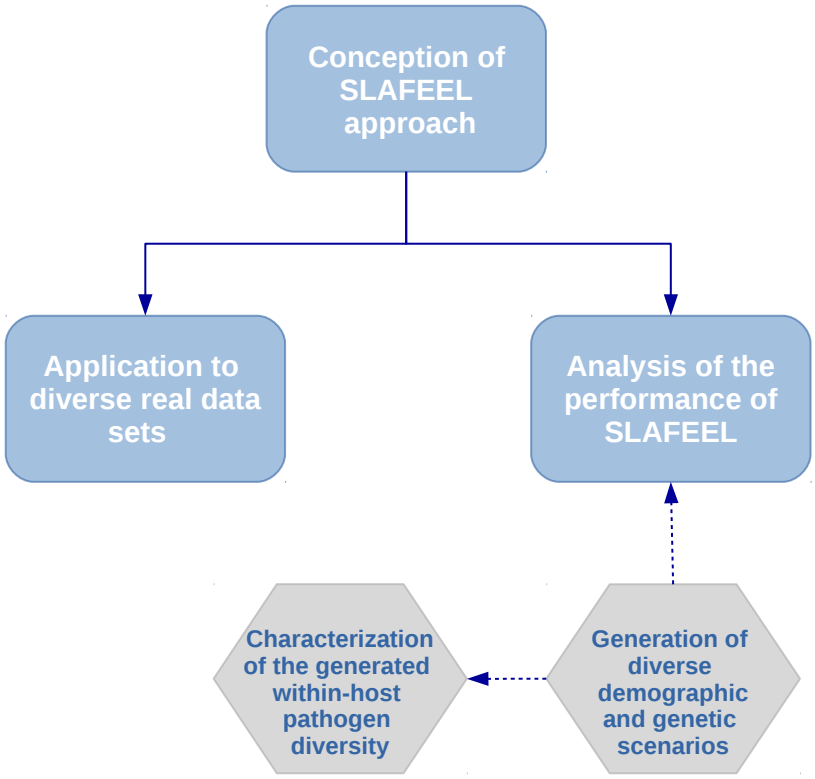


Figure 1.1.: Summary diagram of my thesis objectives.

1.3. Manuscript organization

In order to meet the main objectives mentioned above, this manuscript is organized as follows:

Chapter 2 provides a review of the literature relevant to the aims and objectives of my thesis research and develops a framework for the subsequent chapters. The first section gives a general definition of epidemiology, the causes and the consequences of infectious diseases with a particular focus on virus diseases. The sections that follow respectively address: mathematical modelling in epidemiology, inference of epidemiological links and the assessment of inference approach performance using sensitivity analysis indices.

Chapter 3 presents an original statistical learning framework for inferring transmission links from pathogen sequence data. This chapter is divided into two

parts. The first part details the statistical approach (called SLAFEEL) developed to estimate epidemiological links from deep sequencing data and shows that it is adaptable to different contexts and data from animal, human and plant epidemics. This statistical approach is grounded on statistical learning and based on a mechanistic pseudo-evolutionary model and an associated estimation method. The pseudo-evolutionary model describes transitions between sets of sequences sampled from different hosts and we attach to it a penalized pseudo-likelihood. The adaptation of this approach to different data is performed by adopting specific penalization shapes. In the second part, we apply SLAFEEL to an Equine influenza virus data set. This example allows me to further how SLAFEEL output vary with respect to the tuning choices. I especially investigate this question by adopting different penalization shapes, using therefore alternative parameter estimation methods and assuming diverse temporal assumptions guiding the selection of sources of infection.

Chapter 4 provides an evaluation of the effectiveness of SLAFEEL (developed in Chapter 3) in reconstructing disease outbreaks. This evaluation is performed by applying SLAFEEL to simulated data generated under various demographic and genetic situations. We carried out this work in two phases. The first phase is structured as follows: the proposal of an original demo-genetic model for generating temporal genetic variation in the within-host pathogen population, and the numerical analysis of the model for characterizing its ability to produce within-host diversity. We especially took advantage of the implementation of the demo-genetic model to characterize the viral within-host diversity in fast and non-equilibrium demo genetic dynamics with diverse diversity indices. Hence, we examined how several demo-genetic forces impact the evolution of genetic diversity within a host. The second phase aims at exploring which factors impact the performance of the method for reconstructing epidemiological links of infectious diseases. In this aim, we performed a formal sensitivity analysis of the relationship between SLAFEEL accuracy and the demo-genetic factors that may impact the virus evolution/diversity and transmission.

Chapter 5 gives (i) a global conclusion summarizing my thesis work, (ii) a discussion based on a comparison between my research and other related research and (iii) some directions that may be explored in further research such as incorporating a kernel for better taking into account indirect transmissions, and combining the pseudo-evolutionary model of SLAFEEL with a SEIR model.

Chapter 2

Literature review

Table of contents

2.1. Epidemiology of infectious diseases

2.1.1. Epidemiology definition

The epidemiology has been generally defined as "the study (scientific, systematic) of the distribution (frequency, pattern) and determinants (causes, risk factors) of health states or events (not just diseases) related to health in given populations (neighborhood, school, city, state, country, global), and the application of this study to fight against health problems" (John, 2001; MacMahon et al., 1960; Porta, 2014). In epidemiology, populations of individuals are studied with three principal goals:

- describing the health phenomena occurred within a population according to the characteristics of this population;
- identifying the risk factors that may lead certain health dangers to be more active and efficient in some individual groups than in others;
- assessing the effectiveness of implemented public health interventions.

In this thesis, we focus on molecular epidemiology with the intention of tracing the development of an infectious disease within a host population, which could help us, over a longer term, in understanding how such diseases can spread within a host population, unravelling the risk factors playing a role in disease transmission, characterizing the structure of the host population and the pathogen evolution and, consequently, designing control strategies (Foxman and Riley, 2001; Porta, 2014; Schulte and Perera, 1998; Wang et al., 2015).

2.1.2. Pathogens and infectious diseases

A disease is known as an interruption, cessation or disorder of structure, systems or functions in a human, an animal or a plant. Diseases can be classified into two groups: communicable and non-communicable diseases. Non-communicable diseases can last for a long time and result from genetic, physiological, environmental or behavioral factors or from a mixing of these factors. Cardiovascular diseases (heart and stroke), diabetes, cancers and chronic respiratory diseases (such as chronic obstructive pulmonary disease or asthma) are among the principal non-communicable diseases affecting humans. In animals, there are, for example, some diseases that affect pets such as diabetes, cancer, liver disease

and endocrine disorders. As well, different non-communicable diseases affect the plants representing physiological disorders, which refer to metabolic disturbances, growth retardation or developmental abnormalities resulting often from environmental causes or lack of nutrition. Host-to-host communicable diseases often relate to genetic disorders and infectious diseases. In this thesis, we are interested in infectious diseases that can spread through a host (e.g. individuals, households, agriculture fields and premises) population. An infectious disease is a disease caused by the transmission of a pathogen to a susceptible host. A pathogen is a micro-organism which may be a bacterium, virus, fungus, viroid, algae, prion or protozoan (Alberts et al., 2002).

2.1.3. Transmission modes

Infectious diseases can spread in various ways. Some disease transmissions might occur by direct host-to-host contacts, while other diseases may be transmitted indirectly, e.g. via insect vectors and food products. For instance, diseases can be carried by some intermediate vectors (organisms such as microbes or parasites) to susceptible hosts. For example, malaria, dengue, west Nile are transmitted to humans through mosquitoes. In plants, several mosaic viruses and sharka virus are transmitted by aphids. As well, mosquitoes carry the Yellow fever disease and transmit it to susceptible animals or humans. Hepatitis E can be transmitted via contaminated food and water. However, other disease infections can be airborne (e.g. SARS and influenza) or sexually transmitted (e.g. HIV/AIDS and Herpes). Indeed, SARS and influenza can be transmitted through air and HIV/AIDS and Herpes through breastfeeding, contaminated blood and semen or during birth.

2.1.4. Epidemic

My thesis focuses only on infectious diseases caused by viruses considered to be the most harmful microorganisms that evolve rapidly and cause worldwide epidemics of high impact in humans, animals and plants (Burke, 1997). An epidemic refers to an increase above the expected level of a disease within a given short period of time in a defined host population. In other words, we speak of an epidemic when there is a significant increase in the incidence and prevalence rates within a population in a given place at a given time. The incidence and prevalence rates are the essential criteria used in epidemiology to characterize a disease while determining the speed and frequency of its occurrence (Williams and Wright,

1998). The incidence rate is a measure of the frequency with which a disease is manifested over a specified time period. It represents the rate of new cases of a disease observed from a population at risk within a given period (Rothman et al., 2008; Williams and Wright, 1998). The prevalence rate is the proportion of hosts in a population who are affected by a disease at a specified point in time or over a specified time period (Rothman et al., 2008; Williams and Wright, 1998). Thus, prevalence includes all existing cases (new and preexisting) within the population at the specified time, whereas incidence includes new cases only.

Viral epidemics have major impact not only on human species, but also on animal and plant species while involving serious socio-economic consequences (Mandary et al., 2019).

2.1.5. Consequences

Human, animal and plant epidemics have caused many deaths and high yield losses in agriculture and continue to have serious consequences nowadays. Furthermore, significant global expenditures for virus prevention have been incurred causing considerable economic losses. For instance, to date, more than 200 countries are affected by the COVID-19 infecting over 32.7 million people and causing the death of at least 991,000 people as reported by the World Health Organization (WHO, 27 September 2020). Figure 2.1 illustrates the evolution of the weekly number of registered COVID-19 cases according to the region, and that of the weekly number of global deaths from December 30, 2019 (considered as the start of the pandemic), to September 27, 2020. This figure reflects the severity of COVID-19 outbreak currently causing the deaths of more than 36,000 people per week.

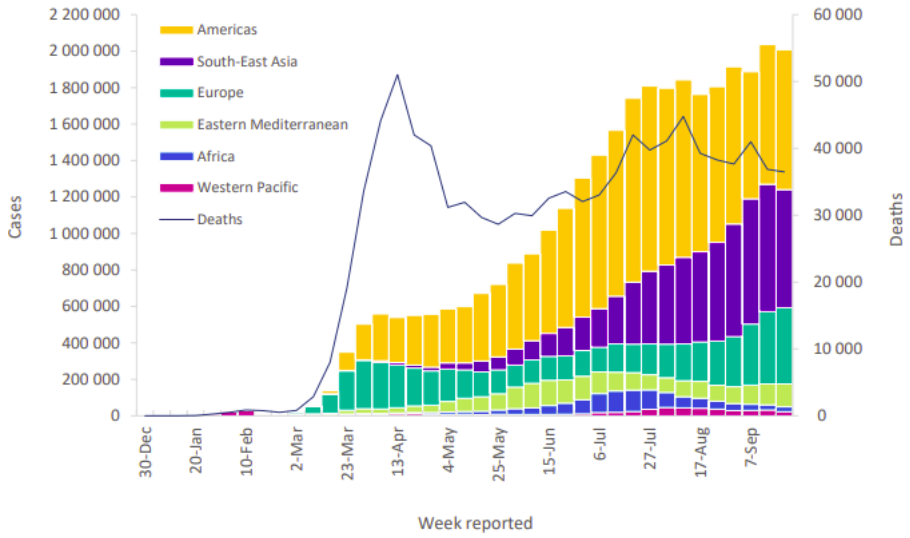


Figure 2.1.: Number of registered COVID-19 cases reported weekly by the WHO in different world regions, and global deaths, from December 30, 2019, to September 27, 2020, published in the weekly report of the WHO[1].

In addition, a recent estimation published by the European Centre for Disease Prevention and Control (ECDC) indicates that Europe experienced 10,705 laboratory-confirmed hospitalized influenza cases during the 2019–2020 influenza season. According to the WHO, rabies and hepatitis E are estimated to cause respectively 55,000 and 57,000 individual deaths per year. The total global expenditure for rabies prevention was assessed to be more than US$1 billion annually. During the 2001 outbreak of foot-and-mouth disease in Great Britain, 6 million animals were culled (Haydon et al., 2003). Similarly to diseases of humans and other animals, plant diseases incite severe consequences while damaging vegetation and crops, reducing yields and raising the treatment and prevention expenditures. For example, the management of sharka disease on Prunus trees remains a challenge and this plant disease induced an overall cost above €10 billion at the global scale over three decades (Cambra et al., 2006; Rimbaud et al., 2015).

[1] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200928-weekly-epi-update.pdf?sfvrsn=9e354665_6

2.1.6. Prevention and control

In order to fight such infectious diseases, limit their propagation and mitigate their potential consequences, various strategic plans have succeeded in scoring important gains at epidemiological, medical, economical and sociological levels. For instance, the obvious way to prevent the virus from spreading to uninfected areas is to reduce contacts (through quarantine). This has precisely been applied at large scale in order to limit the spread of COVID-19. A wide variety of prevention tools have been proposed while depending on disease characteristics. The two prevention tools frequently used to limit and control the disease transmission are drugs and vaccines (Bryan, 2020; Das et al., 2010; Fauci, 2006; Gubbins and Gilligan, 1999; Hall et al., 2004; Lakhani, 1992; Salt et al., 1998). A common practice is the use of antibiotics for animal and human diseases (Das et al., 2010; Fauci, 2006) and agrochemicals or phytosanitary products for plant diseases (Gubbins and Gilligan, 1999; Hall et al., 2004). From the One-Health and Eco-Health perspectives, antibiotics need to be used more prudently in treating human and animal diseases to limit the risks of antibiotic resistance (Allen et al., 2013; Casewell et al., 2003). Likewise, the use of plant-health products should be reduced to minimize their impacts on humans and environment (Frische et al., 2018).

Mathematical modeling of disease outbreak could help in the improvement of the application design of all these types of prophylaxis by taking into account their eventual negative feedback. Indeed, modeling disease outbreak can lead to a better understanding of how such diseases evolve within and between hosts and the risk factors impacting the disease spread, and such knowledge can be exploited to better control disease propagation with appropriate prophylaxis (Foxman and Riley, 2001; Porta, 2014; Schulte and Perera, 1998; Wang et al., 2015).

2.2. Mathematical models of infectious diseases

2.2.1. Mathematical modeling and models in epidemiology

Mathematical modeling is the art of translating our real-life problems into mathematical language based on numerical and theoretical analysis providing answers and insights helpful to understand, solve, or prevent the repetition of these problems (Huppert and Katriel, 2013; Neumaier, 2004). Mathematical modeling relies on a precise language in which each term or assumption expresses make explicit an idea or a vision related to the studied phenomenon. Once a mathematical

model is implemented, mathematical analysis, eventually combined with numerical experiments grounded on simulations, helps us examining how this model is behaving and drawing out the consequences of the formulated assumptions. Thus, the model enables us to thoroughly understand the studied phenomenon, predict the consequences of the problem and also study how these predictions change when the entities settled out in the model vary.

Therefore, the propagation of an infectious disease in a host population could be mathematically modeled while implementing a model describing the transmission of the pathogen between hosts. Strategies for dealing with the disease emergence include focusing special attention on behavioural, biological and/or environmental determinants that promote the propagation of pathogens. Understanding how such determinants interact provides insights into disease dynamics. A mathematical model of infectious disease might take into account the effect of such determinants exploiting sequencing, observations or laboratory experimental results providing, for example, the timing of symptoms, the patterns of contacts among susceptible and infectious hosts, the pathogen sequences within each host, the duration of infectiousness and/or the the latency period (the time from infection to infectiousness). Formulating some or all of these factors in a model allows us to identify the transmission routes, to predict the number of infected cases during an epidemic as well as to plot the entire epidemic curve illustrating the expected number of infected cases within a population at each point in time.

Several mathematical models are used in epidemiology to better understand the causes of a human, animal or plant disease and the risk factors playing a role in disease propagation, evaluate the efficiency of pathogen propagation, predict its current and future courses and design strategies of controlling it (Anderson et al., 1996; Balcan et al., 2009; Bartlett, 1949; Diekmann et al., 1995; Fenichel et al., 2011; Ferguson et al., 2001; Hamer, 1906; Herbeck et al., 2014; Kendall, 1956; Pethybridge and Madden, 2003; Ross, 1911). In this section, we present a brief review of mathematical models in epidemiology used to describe an epidemic and its spread.

2.2.1.1. Brief history

An early study of infectious disease data was traced back to the 17th century and the work of John Graunt in his 1662 book "Natural and Political Observations made upon the Bills of Mortality" (Graunt, 1939). In this study, Graunt analyzed

the different causes of death in London parishes between 1592 and 1603 and gave a method to predict the risks of dying from divers disease. The study was based on data coming from weekly records, called "The Bills of Mortality", containing the numbers and causes of deaths.

What is usually considered as the starting point of infectious disease mathematical modelling is the work of Daniel Bernoulli (Bernoulli, 1760; Dietz and Heesterbeek, 2002) who aimed at assessing the effectiveness of variolation techniques against smallpox. The inference of the temporal and spatial pattern of cholera cases by John Snow (Johnson, 2006; Snow, 1855) is another early valuable work contributing to detect and then better understand microbial disease epidemics. This study was carried out during the cholera epidemic in 1855 in London, identifying the Broad Street water pump as the infection source. An alike understanding of the propagation of typhoid was achieved by Budd (1873). These studies were followed by the study of Farr (1840) who investigated the statistical returns in order to underscore the laws behind the rise and fall of epidemics.

However, the development of methods have been only really picked up in the 20th century with the work of public health physicians (Hamer, 1906; Kermack and McKendrick, 1927; Ross, 1911, 1916; Ross and Hudson, 1917). Their work are considered as the foundations of mathematical epidemiology, setting out the principle of homogeneous mixing (called also the mass-action principle) —by which the spread of infection depends on the current numbers of susceptible and infectious individuals in the population— and the classical deterministic system of equations defining the SIR epidemic model. According to some reviews of the literature (Becker, 1979; Dietz, 1967, 1988; Dietz and Schenzle, 1985; Hethcote, 1994; Hethcote and Levin, 1989; Hethcote et al., 1981; Schwager et al., 1989; Wickwire, 1977) following the publication of Kermack and McKendrick (1927) showing threshold results determining whether a disease outbreak could occur or not, an overwhelming increase in mathematical modelling was noticed particularly in the biological/epidemiological sciences. With a focus on diseases like chickenpox, cancer, rabies, malaria, HIV, smallpox and diphteria (Anderson and May, 1982; Anderson et al., 1992; Bailey et al., 1982; Becker, 1989; Daley and Gani, 2001; Hethcote, 2000; Hethcote and Van Ark, 1991; Isham and Medley, 1996; Longini Jr and Halloran, 2005; Schwager et al., 1989; Usher, 1994), the mathematical models have addressed various aspects such as spatial spread, vaccination, disease vectors, quarantine, acquired and passive immunity, chemotherapy, stages of infection and age structure. The either deterministic or stochastic nature of models form a common classification in the mathematical modeling of infectious diseases.

The following parts of this section are designed to review the main models developed to describe the transmission process with respect to the deterministic and stochastic streams.

2.2.1.2. Deterministic compartmental models

Compartmental models are widely used to assess the contagion probability during an outbreak (Anderson and May, 1982; Hethcote, 2000). These models split up the population into epidemiological classes. Four compartments are often used: S, E, I and R designating susceptible, exposed, infectious and recovered sub-populations, respectively. The S compartment is essential, since there must initially be individuals vulnerable to infection. If a susceptible individual is exposed to the disease, he does not necessarily become able to produce the virus and therefore transmit it immediately. In other words, such an individual does not belong directly to the infectious compartment I but he belongs to a compartment denoted by E for exposed individuals. The incorporation of the E compartment in the epidemic model depends on the disease. If the disease takes time to make the individual infectious, it is required to distinguish between exposed and infectious individuals. After an individual has been infected, the disease can end providing to the individual an immunization against a possible reinfection. Such individual is assigned to the recovered compartment R. Depending on the nature of diseases and pathogens, various mathematical models can be established with ordinary differential equations in terms of these compartments such as SI, SIS, SIR, SIRS, SEI, SEIS, SEIR and SEIRS. In what follows, we present the classical compartment model (SIR) with and without vital dynamics (Allen, 2017; Beckley et al., 2013; Bloomfield, 2009; Kermack and McKendrick, 1927).

2.2.1.2.1. SIR model without vital dynamics: The SIR model is composed of three compartments: the susceptible (S), the infected (I) and the recovered (R), with the temporal functions $S(t)$, $I(t)$ and $R(t)$ providing their respective sizes in the population across time, denoted by t. In the standard SIR model, the birth and death of individuals are not taken into account, there are only infection and recovery. This model is schematically represented in Figure 2.2.

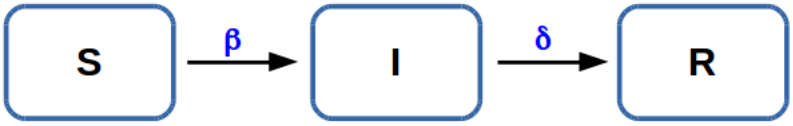


Figure 2.2.: Flow chart for the SIR model without vital dynamics.

The evolution of the compartment sizes is described by the following set of ordinary differential equations:

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta}{N} SI, \\ \dfrac{dI}{dt} = \dfrac{\beta}{N} SI - \delta I, \\ \dfrac{dR}{dt} = \delta I, \\ N = S + I + R. \end{cases} \tag{2.1}$$

where dS/dt, dI/dt and dR/dt quantify the rates of evolution of the sub-population sizes $S(t)$, $I(t)$ and $R(t)$. β is the transmission rate representing the average number of individuals infected by one infectious individual per time unit assuming that all of the individuals in contact with this infectious individual are susceptible to be infected. Thus, a high β is referred to a highly infectious disease. δ is the recovery rate, so that $1/\delta$ is the average time period during which the infectious individual transmits the disease.

2.2.1.2.2. SIR model with vital dynamics: To make it more realistic, the SIR model was established with vital dynamics characterizing the population by a death rate μ and birth rate λ (see Figure 2.3). Thus, the set of differential equations (2.1) was modified as follows:

$$\begin{cases} \dfrac{dS}{dt} = \lambda - \mu S - \dfrac{\beta}{N} SI, \\ \dfrac{dI}{dt} = \dfrac{\beta}{N} SI - \delta I - \mu I, \\ \dfrac{dR}{dt} = \delta I - \mu R, \\ N = S + I + R. \end{cases} \tag{2.2}$$

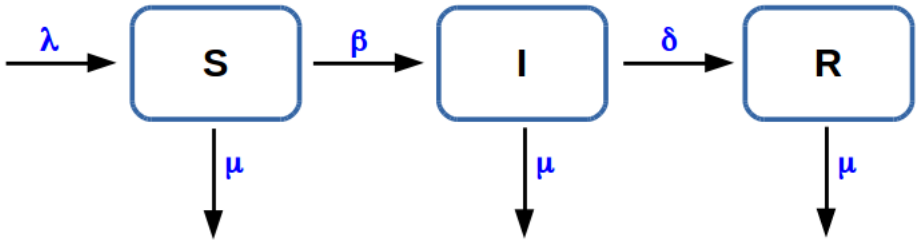


Figure 2.3.: Flow chart for the SIR model with vital dynamics.

2.2.1.2.3. Basic reproductive number R_0: The basic reproductive number R_0 is a key threshold outcome of such epidemic models. Indeed, this number quantifies the transmission of pathogens while providing valuable information about the capacity of disease propagation and the effect of control mechanisms (Diekmann and Heesterbeek, 2000; Grassly and Fraser, 2008; Murray, 1989). R_0 is defined as the average number of people infected by a single infected individual over the disease infectious period, within a completely susceptible population (Diekmann and Heesterbeek, 2000; Huppert and Katriel, 2013; Van den Driessche and Watmough, 2002). For instance, for the two models represented above (SIR without and with vital dynamics), R_0 can be determined respectively by:

$$R_0 = \frac{\beta}{\delta} \qquad \text{and} \qquad R_0 = \frac{\beta}{\mu + \delta}. \tag{2.3}$$

Based on this reproductive number, a disease can be characterized according to its potential to cause an epidemic. If $R_0 > 1$, the disease pathogen is intensely

transmitted and an epidemic will outbreak. Consequently, setting up prophylaxis allowing the reduction of R_0 below one is an important challenge for the control of infectious diseases (this is typically what is expected with lockdown and other sanitary measures currently applied for hampering the spread of COVID-19; Roques et al. 2020b & Roques et al. 2020a).

2.2.1.3. Stochastic models

A deterministic model does not take into account the random effects in the disease spread mechanism and is generally based on a system of differential equations or difference equations. The *solutions* of such a model are entirely determined by, typically, the initial conditions and the parameter values.

Galton, Watson and Steffensen showed, already a long time ago, that the spread of diseases can be viewed as a random process (Galton, 1894; Steffensen, 1933, 1930; Watson and Galton, 1875). Their viewpoint has been translated into what is called *branching process*. In fact, they stated that an outbreak begins with a very small number of infecting hosts and the transmission of infection is a stochastic event depending on the factor of contact between hosts of the population. From this viewpoint, deterministic models are inappropriate for small populations, but are valuable and effective within large populations. Indeed, within a large population, the random effects that could be taken into account are reduced through the law of large numbers, and the output resulting from the deterministic model is close to the average output obtained from a large number of trials of the stochastic counterpart of the model.

A stochastic model takes into account the uncertain events in the mechanism it describes via random effects. The origin of these uncertain events (and, therefore, of the random effects) is potentially related to several factors corresponding to, e.g., environmental, genetic and demographic forces. Because of the numerous uncertainties attached to the role of such forces, stochastic models play an important role in disease transmission modeling (Brauer, 2017). One of the most frequently used stochastic models is the chain model developed by Reed and Frost (Abbey, 1952; Wilson and Burke, 1942). This model is a version of the standard stochastic SIR epidemic model that was first discussed by M'Kendrick (1925).

Like the deterministic SIR model without vital dynamics described in section 2.2.1.2.1, the Reed-Frost model describes the disease transmission within an homogeneous uniformly mixing population of size N partitioned into the three compartments S, I and R (Britton, 2010; Greenwood and Gordillo, 2009). $S(t)$,

$I(t)$ and $R(t)$ respectively represent the numbers of susceptible, infected and recovered individuals at time t, while assuming that at the beginning of outbreak (at $t = 0$), these quantities are initialized by $S(0) = N - M$, $I(0) = M$ and $R(0) = 0$. The dynamics of this model are based on the fact that the transmission of the disease depends on the contact factor. In this way, the disease is transmitted from the infectious individual to the closest susceptible individuals. Contacts between infectious and other individuals are randomly drawn in time (often with a Poisson process) at constant rate λ. Before recovery, infected individuals remain infectious for a given period of time η.

Explicitly, the epidemic starts at time $t = 0$ with M infected individuals, evolves according to the process defined above and ends with the extinction of the sub-population of infected individuals at time T. At the final stage of the epidemic (i.e at time T), the numbers of susceptible, infected and recovered individuals are respectively defined by $S(T) = N - R(T)$, $I(T) = 0$ and $R(T) = M + Z$ where Z is the total number of the individuals infected during the outbreak (over $(0, T]$).

The Markov-process-based SIR epidemic model is similar to the process of Reed-Frost but supposes the randomness of the infectious period (Bailey et al., 1975). In such a Markov process, the infectious periods are supposed to be independent and drawn from an exponential distribution with mean equal to $1/\lambda$ (i.e., λ is the intensity parameter). The Reed-Frost and Markov models have been used widely as basic stochastic epidemic models and many extensions have been formulated (Allen, 2017; Britton, 2010; Daley and Gani, 2001; Greenwood and Gordillo, 2009).

2.2.2. Aggregated models at the population level

A wide range of stochastic models was developed to model the transmission dynamics at the population-level often assuming that individuals are identical and homogeneously mixing (Andersson and Britton, 2012; Ball et al., 2009; Barbour and Mollison, 1990; Ross et al., 2010). Most of these models take the form of the SIR stochastic model mentioned above (or its variants) incorporating for example the demography features of the population (Andersson and Britton, 2012; Bartlett, 1956; Kelatlhegile, 2012; Nåsell, 1999), the seasonal periodicity of infection (Dietz, 1976; Greenhalgh and Moneim, 2003; Keeling et al., 2001; Lin et al., 2015; Nåsell, 2002) or the spatial structure of the population (Bailey et al., 1975; Ball et al., 1997; Milner and Zhao, 2008; Murray et al., 1986). Demography models take

into account the variation of the population size during an outbreak due to the births, deaths and migrations of individuals. The rates of births and deaths are classically assumed to be constant in time. Periodicity models assume that the rate of effective contacts between infected and susceptible individuals depend on the fitness of pathogen during a given season. The spatial models explicitly describe the spatial diffusion of the pathogen.

2.2.3. Host-to-host transmission models

Individual-based models consist of tracking each individual in the population separately while allowing for heterogeneous behavior in relation to social mixing. The heterogeneous behavior of each individual can be defined by an extensive set of relevant characteristics such as age, gender, immunity status, locality, household composition, genetic status and overall health status. Taking into account that each individual can differently get the disease or infect another individual, such models consist of:

- building a social network estimating the possible contacts between individuals;
- dividing the epidemic process into two sub-models called: within-host disease progression and between-host disease transmission. The first sub-model describes the evolution of the disease within each individual. The second describes the pathogen transmission from one host to another.

A large number of individual-based models have been constructed to describe disease dynamics while depending on the heterogeneity source. Individual age is considered as one from the obvious source of variation between individuals (Lui et al., 1988; van Hoek et al., 2012). Another source of heterogeneity is the sex of individuals. Such source is often used to model the dynamics of sexually transmitted disease (Blythe and Castillo-Chavez, 1989; Castillo-Chavez, 2013; Castillo-Chavez et al., 1996). In addition, several epidemic models consist of dividing the population according to the host immunological status (e.g. herd-immunity, vaccine-immunity and passive-immunity; Andreasen, 2003; Andreasen et al., 1997). Other types of individual-based epidemic models are the mixing heterogeneity models such as household and network models (Ball and Neal, 2008; Ball et al., 1997; Chao et al., 2010; Grefenstette et al., 2013). The mixing models are the models assuming different contact rates according to the type of individuals.

For example, in the household epidemic models, where the individuals are divided into small groups called households, the contact rate between a pair of individuals from the same household is different from that between pairs of individuals from different households. We can also cite the models with heterogeneous demography traits, which assume that hosts are characterized by different birth and death rates (Hoppenstaedt, 1975; Kelatlhegile, 2012; Nåsell, 1999).

Likewise, several individual-based models considered the genetic structure of individuals and pathogens as a source of heterogeneity within the host population (Anderson et al., 1992; Gilchrist and Sasaki, 2002; Koelle et al., 2006). To model the host-pathogen co-evolution, the first models developed were focused on the evolution of pathogen virulence and individual resistance in a quantitative manner (Frank, 1994; Sasaki and Godfray, 1999).

The process of pathogen evolution has been neglected in most of the models listed above, whereas mutations and other factors can vary the fitness of the pathogen during an outbreak (Britton et al., 2015) or can be used as a source of information about epidemiological processes (Morelli et al., 2012). That means that the epidemic dynamics may be influenced by the pathogen evolution. Pathogen evolution may be objectively ignored, at least over short period, when the pathogen evolves slowly (e.g for smallpox and measles). However, there are many fast-evolving pathogens (e.g influenza, dengue, ebola, rabies, sharka and mosaic viruses), for which failure to consider evolutionary process may lead to biased conclusions or for which the observation of the evolutionary process may help in unravelling the determinants of transmission dynamics.

Recently, Grenfell et al. (2004) propose an original joint representation of evolutionary phylogenies and epidemic dynamics within a same framework called phylodynamic. Since that time, inspired by the principle of phylodynamic or approaching concepts, several methods (often implemented in software packages) have emerged to model the viral pathogen evolution and transmission (Campbell et al., 2018; Cottam et al., 2008; De Maio et al., 2016; Didelot et al., 2017; Jombart et al., 2011, 2014; Klinkenberg et al., 2017; Mollentze et al., 2014; Morelli et al., 2012; Stadler and Bonhoeffer, 2013; Worby and Read, 2015; Worby et al., 2016).

Packages	Within-host pathogen population simulation	Outbreak simulation	Better to simulate
Outbreaker Jombart et al. (2014)	• Single genotype per host, • No specification for pathogen dynamic, • Do not explicitly model the evolution of pathogen sequences, • Do not model the uninfected population	• SIR stochastic model, • Mutation of sequence at transmission time, • Fixed sampling time, • Fixed infectious period.	Outbreaks where the pathogen does not evolve rapidly (so not for RNA viruses), i.e the within and between host genetic diversity and therefore exploiting a single genotype per host is sufficient
Seedy Worby and Read (2015)	• Multiple genotypes per host, • Multiple samples per host, • Model explicitly the evolution of pathogen sequences, • Specifications for pathogen dynamics	• SIR stochastic model, • Variable transmission bottleneck size, • stochastic infection and recovery generation.	Small outbreaks where pathogen evolves rapidly within the host while leading to high within-host genetic diversity
Transphylo Didelot et al. (2017)	• Single genotype per host, • Do not model the evolution of pathogen sequences, • Specifications for pathogen dynamics	• Monte-Carlo Markov Chain, • Simulated phylogeny tree used as input to account for within-host diversity, • Varying infectiousness level and accounting for unsampled cases, • Complete transmission bottleneck (each host is infected by one single variant).	Outbreaks where hosts can be well enough characterized by a single genotype and where the probability of an observed transmission tree (used to construct the phylogeny tree) can be accurately computed.
Phybreak Klinkenberg et al. (2017)	• Single genotype per host, • Model the within-host evolution combining a coalescence model with Jukes-Cantor substitution model, • Specifications for pathogen dynamics	• Monte-Carlo Markov Chain, • Model simultaneously the phlogenetic and transmission trees, • All cases are sampled, • Requires sampling transmission data, • Complete transmission bottleneck.	Outbreaks where sequences and sampling data are available for all hosts.

Table 2.1.: Comparison of methods to transmission modelling from genetic data.

These methods differ in their underlying genetic models (e.g phylogenetic (Campbell et al., 2018; Didelot et al., 2017; Klinkenberg et al., 2017; Stadler and Bonhoeffer, 2013) or non-phylogenetic models (Jombart et al., 2011, 2014; Worby and Read, 2015; Worby et al., 2016)) and epidemiological models (e.g compartmental models (Mollentze et al., 2014; Morelli et al., 2012; Worby and Read, 2015; Worby et al., 2016) or branching process models (Didelot et al., 2014; Jombart et al., 2011; Klinkenberg et al., 2017), as well as in their ability to account for the within-host pathogen diversity/evolution. In Table 2.1, we review the methods implemented in four software packages and often used in recent studies to simulate outbreaks and validate new developed approaches.

2.2.4. Within-host models

Within a host, the pathogen is subjected to different mechanisms such as replication, mutation (nucleotide substitutions), natural selection and random genetic drift. Such mechanisms may cause changes in the viral load and the viral composition. Most of the methods modelling the pathogen evolution and transmission (mentioned in the previous section) accounting for the within-host genetic diversity/evolution (Klinkenberg et al., 2017; Didelot et al.,2017; Campbell et al., 2018) are based on within-host phylogenetic tree while describing a linear growth for the within-host pathogen population size. In contrast, the method of Worby and Read (2015) gives the opportunity to generate, in a forward setting, the pathogen evolution within each infected host providing at each time the within-host viral composition. Worby and Read assume that the size of the pathogen population converges to an attraction function via the sum of binomial jumps, and the pathogen population varies within a host during an infection due to several demographic and genetic factors such as the inoculum size, mutations, natural selection and random genetic drift. In what follows, we first present the within-host pathogen evolution model developed by Worby and Read, upon which we have relied in this thesis to develop the within-host evolution model represented in Chapter 4. Then, we review various kinetic models that can be used to model the viral growth and decay within a host and therefore, quantify the temporal variation in the viral load.

2.2.4.1. A demo-genetic model for the within-host pathogen evolution

The within-host pathogen evolution model developed by Worby and Read (2015) provides the within-host viral composition at each generation (time unit). The viral composition is defined as a set of different genotypes (genetic sequences) and the frequencies of these genotypes.

In their model, Worby and Read suppose that the within-host pathogen population size tends from one to a given equilibrium population size N_{eq} (N_{eq} can eventually vary with time; in this case N_{eq} is viewed as an attraction function; see the details below), with a probability of mortality per generation, at time t, equal to $\frac{N(t)}{2N_{eq}}$ where $N(t)$ is the population size at time t.

Let $F(t)$ denote the vector of genotype frequencies at time t:

$$F(t) = (f_i(t); i = 1, ..., n(t))$$

with $n(t)$ the number of genotypes at time t. The population size $N(t)$, at time t, is the sum of the frequencies of all genotypes:

$$N(t) = \sum_{i=1}^{n(t)} f_i(t).$$

The frequency of each genotype $f_i(t+1)$ at time $t+1$ is obtained from a binomial distribution conditional on $f_i(t)$ and $N(t)$. Indeed, the authors consider that:

$$f_i(t+1) = 2Y_i(t),$$

where

$$Y_i(t)|f_i(t), N(t) \sim \text{Binomial}(f_i(t), 1-\lambda(t)),$$

$\lambda(t)$ is the probability of mortality, at time t, defined as follows:

$$\lambda(t) = \min\left(1, \frac{1}{2} + \frac{1}{2} \times \frac{N(t) - s(t)}{s(t)}\right) = \min\left(1, \frac{N(t)}{2s(t)}\right),$$

and $s(t)$ is the attraction function of $N(t)$. Indeed, $N(t)$ tends towards $s(t)$ since:

$$\begin{cases} \mathbb{E}[f_i(t+1)|f_i(t), N(t)] = f_i(t) & \text{if } N(t) = s(t) \\ \mathbb{E}[f_i(t+1)|f_i(t), N(t)] < f_i(t) & \text{if } N(t) > s(t) \\ \mathbb{E}[f_i(t+1)|f_i(t), N(t)] > f_i(t) & \text{if } N(t) < s(t). \end{cases}$$

Two specifications for $s(t)$ were considered by (Worby and Read, 2015):

$$\begin{cases} s(t) = N_{eq}, \\ s(t) = N_{max} \times \dfrac{(\cos(2 \times \pi \times \frac{t}{n_{gen}} - \pi) + 1)}{2}, \end{cases}$$

where N_{max} is the maximum population size and n_{gen} is the number of generations during a host infection.

The number of mutations at time t, $M(t)$, is drawn with a binomial distribution:

$$M(t) \sim \text{Binomial}(N(t), \mu).$$

The mutated genotypes are selected by a draw without replacement from the set of genotypes at time t with probabilities proportional to the frequencies at time t. The positions of the mutated nucleotides are selected uniformly and randomly among the L positions (L is the size of the genome) under the Jukes-Cantor model.

2.2.4.2. Kinetic models

The pathogen population size within a host varies with time due to several factors such as natural selection and random genetic drift. The temporal variations in the viral load V within a host, can be described in a stochastic manner, e.g. with the demographic part of the model of Worby and Read (2015). It can also be described with models taken from the abundant literature about deterministic kinetic models forming a rich class of versatile models (Baccam et al., 2006; Beauchemin and Handel, 2011; Beauchemin et al., 2008; Canini and Perelson, 2014; Handel et al., 2010; Nowak and May, 2000; Pawelek et al., 2012; Saenz et al., 2010; Smith and Perelson, 2011). These kinetic models, representing demographic effects, are grounded on systems of ordinary differential equations governing basically the numbers of susceptible target cells, infected cells and virions. Most of these models have been used to predict the course of influenza infection within a host under various situations especially where the host represents a delay in the viral production after the infection, has an innate immune response or is treated with therapeutic antiviral agents.

Here, we represent the simplest classical kinetic model of viral dynamics, which includes susceptible target cells (S), infected cells (I) and virions (V) (see Figure 2.4). It is described by the following differential equations:

$$\begin{cases} \dfrac{dS}{dt} = -\beta SV, \\ \dfrac{dI}{dt} = \beta SV - \delta I, \\ \dfrac{dV}{dt} = pI - cV, \end{cases} \tag{2.4}$$

where the infected cells, I, are generated by the interaction between the virus, V, and susceptible target cells, S, at rate β. These cells are lost at rate δ and produce virus at rate p. The virions, V, are assumed to be cleared with constant rate c.

Figure 2.5 shows how the numbers of target cells, infected cells and virions governed by the simple kinetic model evolve within a host during the infectious period (10 generations).

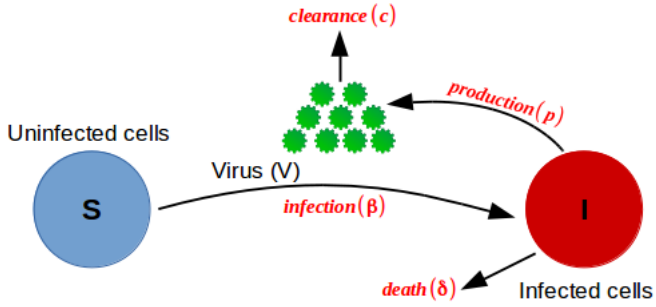


Figure 2.4.: Schematic representation of the simple kinetic model defined by the system of differential equations (2.4).

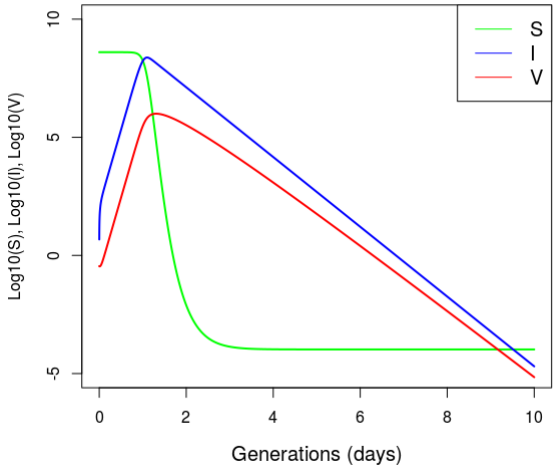


Figure 2.5.: Changes in within-host quantities of susceptible cells (green line), infected cells (blue line) and virions (red line) predicted by the simple kinetic model defined by the ssystem of differential equations (2.4).

2.3. Inference of epidemiological links

With the aim of understanding how pathogens spread within a host population, a particular field in epidemiological modeling contributed to the reconstruction of transmission pathways by addressing the question "who infected whom?" (or, more generally, "who is closely related to whom?" in the disease transmission dynamics). Using data collected from numerous host units (individuals, households, agriculture fields, ...) carrying an infectious disease, answering the above-question consists in inferring the transmission chains linking the observed infected hosts. These chains form a graph in which nodes represent the infected hosts and edges illustrate the transmission success with the possibility of adding spatio-temporal information (Kendall et al., 2018).

Inferring epidemiological links during an outbreak inform underlying epidemiological processes (i.e., provide insights into transmission dynamics and risk factors) and, as a consequence, help in designing control strategies. However, estimating the transmission routes with high accuracy requires detailed information at the individual level, and the mathematical approaches that can be used to

achieve this objective obviously depend on the available data (Cori et al., 2017). In the literature, we find several approaches designed to infer the epidemiological links in infectious disease outbreak exploiting different types of data, and we review them in the following sections.

2.3.1. Epidemiological investigation

The most frequent approaches developed to unravel transmission links within a host population are those analyzing detailed epidemiological data (Albrich and Harbarth, 2008; Aldrin et al., 2011; Assiri et al., 2013; Cauchemez and Ferguson, 2012; Cauchemez et al., 2006, 2011, 2016; Chun et al., 2020; Faye et al., 2015; Ferguson et al., 2001; Haydon et al., 2003; Heijne et al., 2012; Kao, 2002; Kucharski et al., 2020; Leo et al., 2003; Makintubee et al., 1987; Mollentze et al., 2014; Mossong et al., 2008; Shen et al., 2004; Snitkin et al., 2012; Wallinga and Lipsitch, 2007; Wallinga and Teunis, 2004; Xu et al., 2020; Yang et al., 2020). The transmission chains inferred with these approach were used, for example, for the estimation of the number of secondary infected cases, the assessment of the outbreak intensity and the estimation of the basic reproductive number R_0.

Basically, the epidemiological data that can be exploited refer to temporal, contact, spatial and demographic data. The temporal data are the most frequently available data and provide information on events in the outbreak such as the number of recovery cases, the number of dead cases, timing of symptoms and hospitalization (Cauchemez et al., 2006; Faye et al., 2015; Leo et al., 2003; Snitkin et al., 2012; Yang et al., 2020).

Contact tracing data are considered as the backbone of the understanding of many outbreaks. Such data consist of identifying symptomatic cases that were in contact with infected hosts that could have been their sources of infection. These data, if they are collected in real time, can help to limit the spread of the pathogen while giving information about individuals in contact with infected hosts and prompting these individuals for a rapid isolation. This type of data provided most of the information exploited to reconstruct some transmission pathways during the 2002-2003 SARS outbreak (Leo et al., 2003; Shen et al., 2004), the 2009 H1N1 pandemic influenza (Cauchemez et al., 2011), the 2012 Middle East respiratory syndrom epidemic (Assiri et al., 2013), the 2013-2016 Ebola crisis (Faye et al., 2015) and, presently, the COVID-19 pandemic (Xu et al., 2020; Yang et al., 2020). Thus, Xu et al. (2020) was able to reconstruct 1407 transmission pairs for COVID-19 in Mainland China using information about the potential contacts with the infected

hosts and the social relationships between them.

Moreover, some mathematical approaches exploit spatial data such as the locations of hosts, or the geographical distances between hosts to infer the epidemiological links across an outbreak (Aldrin et al., 2011; Ferguson et al., 2001; Mollentze et al., 2014; Morelli et al., 2012). Such data have been particularly used to reconstruct outbreaks between individuals, households, chiefdoms, crop fields and farms. For instance, to infer the transmission links of the 2001 foot-and-mouth disease epidemic in Great Britain, Ferguson et al. used farm location data, as did Aldrin et al. for inferring epidemiological links of infectious salmon anaemia between Norwegian salmon farms. Mollentze et al. used location data of dogs, jackals as well as wild and livestock animals in order to estimate the long-distance transmission routes of the rabies virus.

Stratifying a population by demographic data such as sex, race, age or occupation, is generally used in order to distinguish transmission risks within and between these stratified-groups (Albrich and Harbarth, 2008; Farrington et al., 2001; Heijne et al., 2012; Xu et al., 2020). For example, Heijne et al. quantified transmission of Norovirus while differentiating 4 transmission routes patient to patient, patient to healthcare worker, healthcare worker to patient, and healthcare worker to healthcare worker. The study of Heijne et al. showed that the mainstream recognized transmission route was from patient to patient, followed by patient to healthcare worker. Likewise, Xu et al. estimated the risk of COVID infection stratifying the population by age and sex. The examination of the age-stratified and sex-specific hazard of infection indicated that the risk of infection is more significant within households for young, elderly and female people.

2.3.2. Pathogen sequence data analysis

For fast-evolving pathogens, several frameworks have been proposed to integrate the analysis of pathogen sequence data in virus transmission studies (Campbell et al., 2018, 2019; Cottam et al., 2008; De Maio et al., 2016, 2018; Didelot et al., 2014, 2017; Hall et al., 2015; Hayama et al., 2019; Hughes et al., 2012; Jombart et al., 2011, 2014; Leavitt et al., 2020; Mollentze et al., 2014; Morelli et al., 2012; Skums et al., 2018a; Soubeyrand, 2016; Worby et al., 2016, 2017; Wymant et al., 2018; Ypma et al., 2012, 2013b). These frameworks are grounded on different principles varying from those based on statistical metrics to those based on a mechanistic modelling of pathogen evolution and transmission. The following paragraphs highlights some of these frameworks.

One of the most fundamental statistical methods developed to infer transmission links from pathogen sequence data is based on comparative tools. These comparative tools mainly identify the specific variants shared by different hosts or evaluating the genetic distance between two pathogen sequence samples observed from two different hosts (Eyre et al., 2013; Hughes et al., 2012; Murcia et al., 2012; Walker et al., 2013, 2014). This distance is often defined by the number of single nucleotide polymorphisms (SNPs) between isolates. Therefore, the greater the similarity between the two pathogen sequence samples observed from two different hosts, the more likely the transmission event between the two hosts.

In parallel to these pairwise approaches, there are several approaches inferring transmission events from genetic sequence data based on constructing phylogeny, phylogeography and some birth-death processes (Cottam et al., 2008; De Maio et al., 2018; Didelot et al., 2014, 2017; Hall et al., 2015; Kenah et al., 2016; Leitner and Romero-Severson, 2018; Lemey et al., 2010; Pybus et al., 2012; Rasmussen et al., 2011; Stadler and Bonhoeffer, 2013; Wymant et al., 2018). A phylogenetic tree describes the inferred evolutionary relationships between pathogens sampled from infected hosts and can provide the times of lineage divergence. In the phylogenetic tree, external nodes represent sampled pathogens and internal nodes represent the most recent common ancestor of its descendants (Pybus and Rambaut, 2009; Worby et al., 2016; Ypma et al., 2013a). Exploiting or reconstructing phylogeography tree allows to frame the spatial diffusion process (Lemey et al., 2010). A birth-death process is a Markov chain process that models the current size of a population, where each individual can "give birth" to another individual or "die" (Feller, 2008; Karlin, 2014). In finite populations, such models can be used to study infectious disease dynamics quantifying the number of infected individuals (Andersson and Britton, 2012; Bailey, 1990) or modelling quantities of interest in an evolutionary setting such as coalescence (Crawford et al., 2018). Such evolutionary models or processes are useful to guide the estimation of who-infected-whom and often used as a *priori* for approaches adopting a Bayesian framework.

Additional approaches for inferring who infected whom or who is closely related to whom can be based on models combining minimal genetic distances between intra-host viral populations and properties of social networks relevant to pathogen spread (Skums et al., 2018a) or on joint models of epidemiological dynamics and evolutionary processes (De Maio et al., 2016; Jombart et al., 2014; Lau et al., 2015; Mollentze et al., 2014; Morelli et al., 2012; Soubeyrand, 2016; Worby and Read, 2015; Worby et al., 2016; Ypma et al., 2012, 2013a).

Many of the model-based approaches mentioned above have been proposed to infer transmission links exploiting one pathogen sequence per infected host, typically the consensus sequence or the majority sequences (e.g., Campbell et al., 2018, 2019; Cottam et al., 2008; Jombart et al., 2011, 2014; Morelli et al., 2012; Ypma et al., 2012, 2013b). Using a single variant per host limits the amount of information about the viral composition within the host although recent sequencing techniques, such as *deep Sanger sequencing* (DSS) and *high-throughput sequencing* or *next-generation sequencing* (HTS or NGS) open new routes to access a detailed description of the genetic heterogeneity that can exist within a host during an infection (Alizon et al., 2011; Gire et al., 2014; Lauck et al., 2012; Murcia et al., 2010, 2012; Nelson and Hughes, 2015; Worby et al., 2014, 2017; Wright et al., 2011). Thanks to these techniques providing a subsample of the pathogen variants in the host at the sampling time, it is now recognized that the virus genetic diversity may vary spatially and temporally during the infection, due to several demographic and genetic factors (e.g mutation, selection and genetic drift processes) acting at the within-host scale (Abel et al., 2015; Alizon et al., 2011; Cuevas et al., 2015; Poirier and Vignuzzi, 2017; Pybus and Rambaut, 2009; Simmons et al., 2012). Variations in the within-host genetic diversity may then affect between-hosts virus transmissions (Abel et al., 2015; Worby et al., 2014). Offering the possibility to assess within-host pathogen diversity, such sequencing techniques fostered the development of model-based approaches exploiting, in one way or another, the degree of genetic similarity between viral variants present within the different hosts to identify linked hosts and infer transmissions (De Maio et al., 2016, 2018; Didelot et al., 2014, 2017; Hall et al., 2015; Skums et al., 2018a).

2.3.2.1. Visualization of within-host genetic diversity

Murcia et al. (2010, 2012) and Hughes et al. (2012) studied the within-host genetic diversity through the use of clonal amplicon Sanger sequencing. These studies illustrated how the exploitation of multiple sequences per host provided more information for reconstructing transmission pathways between hosts.

In this section, we present the work of Hughes et al. (2012) examining the dynamic of EIV (Equine influenza virus) genetic diversity within the horses infected during the 2003 Equine influenza outbreak. This examination was based on the analysis of multiple viral sequences sampled from infected horses at different times. Each infected horse is characterized by a dominant variant and several minor variants. Figure 2.6 illustrates the genetic diversity within four horses ac-

cording to sequences having nucleotides G or A on the site 230. Each network in this figure was generated from all sequences collected from the horse at given times and the size of each circle is proportional to the sequence frequency within the host at the given time. The color indicates the yard and the day the sample was taken. Sequences with A230 (nucleotide A at site 230) are circled with a thick line. Black dots on the branch indicate the number of mutations differentiating two sequences.

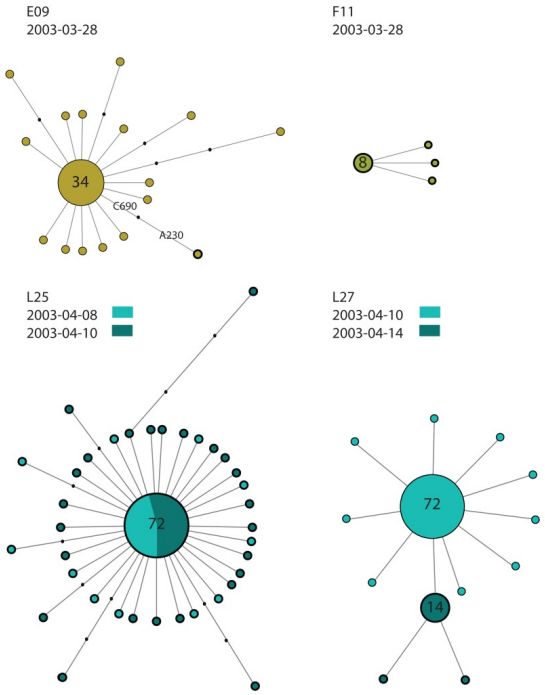


Figure 2.6.: Median joining networks illustrating the intra-host viral diversity of four representative horses. Networks were generated from all the sequences from an individual horse and the size of the circle is relative to the sequence frequency. The color indicates the yard and day the sample was taken from. Sequences with A230 are circled with a thick line. Note that a single clone has A230 in horse E09. Black dots on the branch indicate the number of mutation differentiating two sequences. This figure was extracted from Hughes et al. (2012).

Based on all the data collected by Hughes et al., F11 was the first horse with A230 as the dominant variant (on March 28) and all the clones sequenced from this horse carry the mutation. On the same day, horse E09 presented A230 linked to

C690, which can be explained by the fact that A230 could have been present in the viral population before March 28. L25 and L27 horses were sampled twice during the outbreak, the horse L25 showed sequences with A230 as the dominant variant on both sampled days, while L27 initially showed sequences with G230 then A230 four days later, illustrating the heterogeneity of the viral dynamics within a host. This change in the dominant variant could be due to substitution/mutation or mixed infection.

This case study shows that, within each host, different variants can be observed at a sampling time, and different consensus sequences can be observed at different sampling times. This suggests the within-host evolution as well as the within-host genetic diversity should be taken into account when reconstructing the transmission links of an outbreak because they may lead to improve this reconstruction.

2.3.2.2. Various frameworks

This section is dedicated to review some approaches developed to infer epidemiological links from pathogen sequences data, and hence complements Section 2.3.2. We first provide a comparison and summary of their features in Table 2.2. The cited approaches differ by the type and amount of exploited data and their ability to account for the within-host genetic diversity and evolution. Based on the comparison between the cited approaches, we classify these approaches into three main categories and then give an example of each category. The approaches of the first category do not account for the within-host evolution and diversity. They infer transmission links based on temporal data exploiting a single pathogen sequence per host. The last two categories account for the within-host evolution and diversity but exploit, respectively, a single pathogen sequence per host and multiple pathogen sequences per host. Those exploiting one pathogen sequence per host are often based on phylogeny. Such approaches consider that the viral population within a host correspond to a phylogenetic subtree where the pathogen sequences are modelled as tips.

Approaches	Genetic data	Temporal data	Contact data	Spatial data	Phylogeny	Within-host evolution	Multiple sequences
Cottam et al. 2008	✔	✔	✗	✗	✔	✗	✗
Aldrin et al. 2011	✔	✗	✗	✔	✗	✗	✗
Jombart et al. 2011	✔	✔	✗	✗	✗	✗	✗
Ypma et al. 2012	✔	✔	✗	✔	✗	✗	✗
Morelli et al. 2012	✔	✔	✗	✔	✗	✗	✗
Ypma et al. 2013b	✔	✔	✗	✔	✔	✔	✔
Stadler and Bonhoeffer 2013	✔	✔	✗	✗	✔	✗	✗
Jombart et al. 2014	✔	✔	✗	✗	✗	✗	✗
Didelot et al. 2014	✔	✔	✔	✔	✔	✔	✔
Gavryushkina et al. 2014	✔	✔	✗	✗	✔	✗	✗
Mollentze et al. 2014	✔	✔	✗	✔	✗	✗	✗
Hall et al. 2015	✔	✔	✗	✔	✔	✔	✔
Lau et al. 2015	✔	✔	✗	✔	✗	✗	✗
De Maio et al. 2016	✔	✔	✗	✗	✔	✔	✔
Worby et al. 2016	✔	✔	✗	✗	✗	✔	✔
Soubeyrand 2016	✔	✔	✗	✔	✗	✗	✗
Didelot et al. 2017	✔	✔	✗	✗	✔	✔	✔
Klinkenberg et al. 2017	✔	✔	✗	✗	✔	✔	✔
Worby et al. 2017	✔	✔	✗	✗	✗	✔	✔
De Maio et al. 2018	✔	✔	✗	✗	✔	✔	✔
Skums et al. 2018a	✔	✔	✗	✗	✗	✔	✔
Wymant et al. 2018	✔	✗	✗	✗	✔	✔	✔
Campbell et al. 2018	✔	✔	✗	✗	✔	✔	✗
Campbell et al. 2019	✔	✔	✔	✗	✗	✗	✗
Hayama et al. 2019	✔	✔	✗	✔	✗	✗	✗

Table 2.2.: Comparison of some approaches developed to reconstruct outbreaks. These approaches differ by the type of exploited data and their ability to account for the within-host genetic diversity and evolution. Each row represents an approach for inferring transmission links, and each column represents a feature of the approach. ✔ means that the feature is allowed, while ✗ means that the feature is not included. ✔ means that in the mentioned study, pathogen sequences are modelled as tips in the within host phylogenetic mini-trees, thus the within-host genetic diversity is assessed by the within-host effective population size.

2.3.2.2.1. No within-host evolution or diversity, Morelli et al. (2012): Morelli et al. developed a Bayesian approach to infer transmission links and infection dates by combining genetic, temporal and spatial data. This approach consists of estimating the transmission links while estimating transmission and temporal parameters without taking into account the evolution within a host and therefore the within-host diversity of genetic data. It is based on a joint posterior distribution grounded on pathogen sequence data (single sequence per host) and several epidemiological data (detection time, veterinary assessment of the duration of infection up to detection, culling time, spatial locations). Morelli et al. incorporated the genetic information through a probability distribution for the number of substitutions between sequences over the time that separates them while computing the dates of infection and the probability of observing these sequences for a given transmission tree.

This approach takes into account the delay between infection and infection detection (latency duration), as well as the difference between observed and transmitted pathogen genetic sequences. Morelli et al. assumed that the infection potential of each host depends on its spatial location, time of the end of latency phase (start of infectious period), time of the end of infection and two transmission parameters. The first transmission parameter assesses the infection strength of each infectious host while the second parameter assesses the decrease of the infection potential with the geographical distance.

Morelli et al. built a joint posterior distribution of the transmission tree, infection times, latency duration, periods from infectiousness to detection, and latency and transmission parameters, given the data. The latency parameters are the expected value and the variance of the latency duration. Available data are observed pathogen sequences, observation times, host locations, removal times and observed periods from infectiousness to detection. This distribution is defined by a product of five terms representing: conditional pseudo-distribution of observed sequences, conditional distribution of pathogen observation times, joint distribution of transmissions and infection times, distributions of latency period and detection duration, and prior distribution of transmission and latency parameters.

Assuming that each host is characterized by one pathogen sequence and that the substitution rate per day per nucleotide is constant, the conditional distribution of observed sequences is derived from the probability distribution of the number of different nucleotides between two sequences during the evolutionary period separating them. The conditional distribution of pathogen observation

times is derived from the spatio-temporal transmission probabilistic model proposed by the authors.

Ignoring the presence of unsampled hosts and assuming that: (a) an external source can infect only one host and the other hosts are infected by hosts that are in the data set; (b) each host can transmit the disease up to the culling time; (c) the strength of infection is the same for all hosts; (d) the risk that a susceptible host will become infected by a given infectious host decreases exponentially with the geographical distance between the hosts, the joint distribution of transmissions and infection times is established based on an exponential kernel for geographical distances as the product of two terms. The first term is the probability that a host has not been infected until the time of infection by previously infected host. The second term is the probability density that a host has been infected by other host at the time of infection. Gamma distributions are used for the latency durations and detection durations and independent exponential priors are associated to the latency and transmission parameters.

Morelli et al. showed the potential value of their approach in inferring the epidemiological links between farms for the 2001 and 2007 Foot-and-Mouth Disease Virus (FMDV) outbreaks. Applying their approach to 2007 FMDV outbreak data, the authors were able to identify the interface between the two phases of the outbreak. Several limitations can be pointed out: the ignorance about eventual unobserved cases, the assumption that any infected host carries a single genotype at any time (even if the pathogen evolves across time), the approximation that is made in the genetic part of the likelihood due to the fact that the transmitted genotype is not reconstructed; An attempt to solve the latter issue within a similar modeling framework was proposed by Soubeyrand (2016).

2.3.2.2.2. Within-host evolution and diversity while exploiting a single pathogen sequence per host, Ypma et al. (2013b): Ypma et al. (2013b) developed an approach exploiting both epidemiological and genetic data aiming at inferring simultaneously the transmission tree and the phylogenetic tree, while supposing that these two trees are different. Their inference approach is based on the posterior probability of the transmission tree, the phylogenetic tree and the within-host dynamics. This posterior probability depends on epidemiological and mutational parameters. The included within-host dynamics must be specified according to genetic data and consist of giving the genetic diversity within a host at the specified time. The within-host genetic diversity is measured by the product of the pathogen generation time and the effective pathogen population size

(corresponding to the number of lineages in the phylogenetic subtree) within the host at the specified time. The *joint* posterior probability is defined as the product of four terms representing the likelihood for the transmission tree (independent of sequence data), the likelihood for the phylogenetic tree, the likelihood for the mutational parameters and the prior information on parameters, transmission tree and within-host dynamics.

Ypma et al. considered that the transmission tree delineates the phylogenetic subtrees and the coalescent times are consistent with the epidemiological process. Each subtree is established at the proposed infection time, with a single pathogen sequence. The likelihood component for the phylogenetic tree is considered to be obtained under a coalescent process, considering that the number of lineages of a subtree (within a host) can decrease or increase by one due to a coalescent event or an incoming lineage, respectively. Thus, it is defined as the product of the two likelihoods: the likelihood that the lineage coalesced and the likelihood that the lineage did not coalesce. The mutation model is based on the fact that the ancestry of sampled pathogen is known and the mutation rate is constant for all the infected hosts. The mutation process triggers with each infection and implements the Felsenstein's pruning algorithm in order to take into account the multiple mutations occurring at the same locus. Assuming that the incubation period is gamma distributed, the likelihood component for the transmission tree depends on the probability of the gamma distribution for the length of the incubation period (the difference between the start of the infectious period and the start of infection) and the infectiousness of the infecting host relative to the total infectiousness of all hosts.

The advantage of this approach is that it estimates the phylogenetic and transmission trees simultaneously under the same epidemiological conditions. This might reduce the loss of information containing in the phylogeny tree that can occur when considering a sequential estimation procedure for transmission tree and phylogenetic tree. A limitation of this approach is that it requires the specification of a model describing the within-host dynamics, which is not obvious for all pathogens. As well, this approach does not take into account the unobserved and non sampled hosts. In addition, inferring the transmission links independently from pathogen sequence data can create uncertainty in the reconstruction of an outbreak.

2.3.2.2.3. Within-host evolution and diversity while exploiting multiple pathogen sequences, De Maio et al. (2016): De Maio et al. developed a

Structured COalescent Transmission Tree Inference approach (SCOTTI) to infer the epidemiological links within an outbreak combining genetic data from sampled hosts with epidemiological dating information (exposure of hosts to the outbreak). SCOTTI enables the inference of host-to-host transmission taking into account the within-host genetic variation (allowing the exploitation of multiple sequence samples per host), the multiple infections of the same host and the non-sampled hosts. This Bayesian approach consists of modeling the hosts as containers of different pathogen populations, and the transmissions between these populations as migration events. Instead of constructing the phylogenetic tree between hosts, SCOTTI adopts an approximation of the structured coalescent, which is exploited to infer the transmission routes.

Broadly speaking, SCOTTI jointly infers the transmission links between hosts and the structured coalescent, while taking into account the unobserved cases and using genetic and epidemiological data. The structured coalescent is a statistical model developed by the authors in order to describe the genealogy of hosts (sampled from a structured population) whose pathogen populations have important deviations resulting from important migrations. The inclusion of epidemiological data was performed by introducing the set of host exposure times into the joint posterior distribution. If two hosts are exposed to the infection at the same time, they are supposed to have the same migration / transmission rate. De Maio et al. assumed that each host is characterized by multiple pathogen sequences at the sampling time , allowing for multiple samples at different times from any host.

To infer transmission pathways in a Bayesian framework, De Maio et al. built a joint posterior distribution that mainly depends on four parameters: (1) the evolutionary parameter representing the molecular evolution process, (2) the set of hosts exposure times allowing to incorporate epidemiological / temporal knowledge, (3) a bifurcating tree elucidating the phylogenetic relationship between sampled hosts, and (4) the history of migration events. Furthermore, this posterior distribution is made up of three terms concerning the pathogen evolution and transmission. The first term represents the likelihood of the pathogen sequences given the genealogy and the nucleotide substitution model. This likelihood is calculated with Felsenstein's pruning algorithm (Felsenstein, 1981) assuming that the sequences evolve during the outbreak according to a continuous Markov chain. The second term is dedicated to the probability density of the migration history. The last term is the joint prior distribution on the parameters of the evolution and migration models.

SCOTTI is subjected to several limitations. One of these limitations is that this

approach does not model the transmission bottlenecks. This may be beneficial in some situations where the size of transmission inoculum is large (Hughes et al., 2012) but not in other situations where not modelling the bottleneck transmission may result in biases in transmissions inference. A second limitation is that SCOTTI does not estimate the infection times, meaning that the important epidemiological insights such as the infection period of each host are not determined. Furthermore, in order to include the epidemiological data into the posterior distribution, SCOTTI requires a predefined exposure interval for each host in order to stake out its infectious period whereas such data are frequently unavailable.

Chapter 3

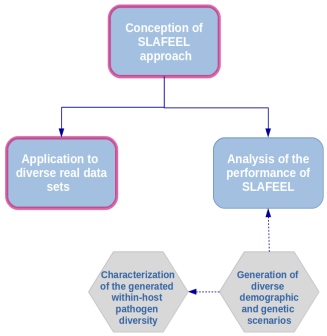
Development of a statistical approach for estimating transmissions of infectious diseases and application to real data sets

This chapter introduces a published article Alamil et al. (2019) followed by an application of SLAFEEL to Equine influenza data.

Table of contents

3.1. Graphical summary

Chapter tasks

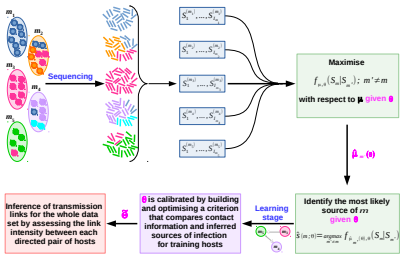


Answer to:

- How to get insight into the transmission dynamics of an infectious disease?
- How to develop a statistical learning approach adapted to infer epidemiological links between hosts infected by a virus from pathogen sequence data reflecting the within-host diversity of pathogen?
- How robust, effective and versatile is the developed inference approach?

Methodological ingredients

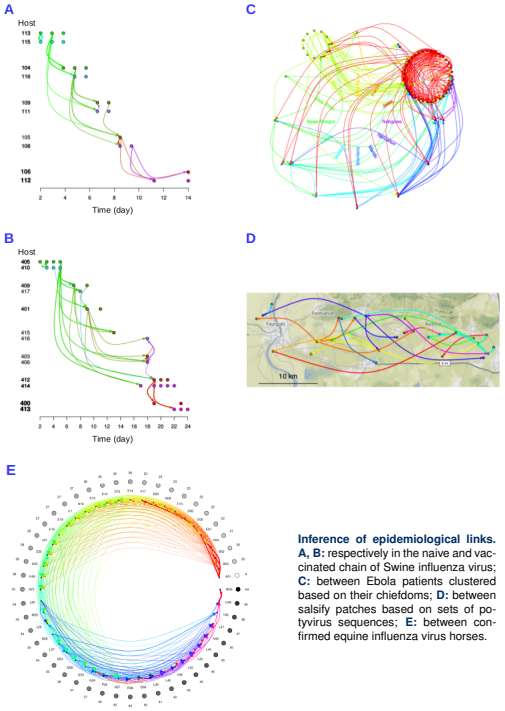
Graphical representation of the developed statistical learning approach



Real epidemic genomic Data

- **Animal epidemics:**
  - Swine influenza virus:
    - Numbers of sampled hosts: 10 in the naive chain and 13 in the vaccinated chain;
    - One sample per host;
    - Mean sequencing depth $\sim$ 41 & 58 sequence fragments per host, respectively;
    - Fragment length: 939 nucleotide bases.
  - Equine influenza virus:
    - Numbers of sampled hosts: 48 horses;
    - One sequence fragment sample per host;
    - Mean sequencing depth $\sim$ 9 sequence fragments per host;
    - Fragment length: 903 nucleotide bases.
- **Human epidemic (Ebola virus):**
  - Numbers of sampled hosts: 58 patients;
  - 31 sequence fragment samples per host;
  - Mean sequencing depth $\sim$ 14300 on average over the 31 sequence fragments;
  - Fragment length: 885 nucleotide bases on average over the 31 available sequence fragments.
- **Plant epidemic (Potyvirus of wild salsify):**
  - Numbers of sampled hosts: 27 patches;
  - One sequence fragment sample from each of the 189 sampled plants;
  - Mean sequencing depth $\sim$ 1550 sequence fragments per host;
  - Fragment length: 438 nucleotide bases.

Main results



Inference of epidemiological links. A, B: respectively in the naive and vaccinated chain of Swine influenza virus; **C:** between Ebola patients clustered based on their chiefdoms; **D:** between salsify patches based on sets of potyvirus sequences; **E:** between confirmed equine influenza virus horses.

Conclusion & perspectives

- **This approach is:**
  - **Adaptable** to very different contexts and data from animal, human and plant epidemics.
  - **Valuable** in non-standard situations where classical mechanistic assumptions may be erroneous and when sequencing and variant calling may be noisy.
  - **Sensitive** to the penalization shapes associated to the hypotheses considered for the penalization, as well as to the temporal constraint used to select the putative sources of infected hosts.
- **Further works will aim to:**
  - **Examine** in depth the sensitivity of this approach to the penalization shapes and temporal constraints.
  - **Test** how the performance of `SLAFEEL` scale up with big data (e.g, using multiple samples per host, exploiting sequences of large size, ...)
  - **Allow** multiple infections.
  - **Combine** `SLAFEEL` to a SEIR model.

3.2. Article

Note: there is a typographical error in Equations (4.4) and (4.5). The parameter θ must be superscript.

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Alamil M, Hughes J, Berthier K, Desbiez C, Thébaud G, Soubeyrand S. 2019 Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. *Phil. Trans. R. Soc. B* **374**: 20180258.
http://dx.doi.org/10.1098/rstb.2018.0258

Accepted: 14 March 2019

One contribution of 15 to a theme issue 'Modelling infectious disease outbreaks in humans, animals and plants: approaches and important themes'.

Subject Areas:
computational biology, ecology, health and disease and epidemiology, microbiology

Keywords:
contact information, infectious disease, pathogen spread, training data, transmission trees, within-host pathogen diversity

Author for correspondence:
S. Soubeyrand
e-mail: samuel.soubeyrand@inra.fr

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4450355.

THE ROYAL SOCIETY PUBLISHING

Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases

M. Alamil[1], J. Hughes[2], K. Berthier[3], C. Desbiez[3], G. Thébaud[4] and S. Soubeyrand[1]

[1]BioSP, INRA, 84914 Avignon, France
[2]MRC-University of Glasgow Centre for Virus Research, Glasgow G61 1QH, UK
[3]Pathologie Végétale, INRA, 84140 Montfavet, France
[4]BGPI, INRA, Univ. Montpellier, SupAgro, Cirad, 34398 Montpellier, France

SS, 0000-0003-2447-3067

Pathogen sequence data have been exploited to infer who infected whom, by using empirical and model-based approaches. Most of these approaches exploit one pathogen sequence per infected host (e.g. individual, household, field). However, modern sequencing techniques can reveal the polymorphic nature of within-host populations of pathogens. Thus, these techniques provide a subsample of the pathogen variants that were present in the host at the sampling time. Such data are expected to give more insight on epidemiological links than a single sequence per host. In general, a mechanistic viewpoint to transmission and micro-evolution has been followed to infer epidemiological links from these data. Here, we investigate an alternative approach grounded on statistical learning. The idea consists of learning the structure of epidemiological links with a pseudo-evolutionary model applied to training data obtained from contact tracing, for example, and using this initial stage to infer links for the whole dataset. Such an approach has the potential to be particularly valuable in the case of a risk of erroneous mechanistic assumptions, it is sufficiently parsimonious to allow the handling of big datasets in the future, and it is versatile enough to be applied to very different contexts from animal, human and plant epidemiology.

This article is part of the theme issue 'Modelling infectious disease outbreaks in humans, animals and plants: approaches and important themes'. This issue is linked with the subsequent theme issue 'Modelling infectious disease outbreaks in humans, animals and plants: epidemic forecasting and control'.

1. Introduction

In order to most effectively predict and control the spread of infectious diseases, we need to better understand how pathogens spread within and between host populations and assess the role of the environment in the transmissions. The question *how do pathogens spread?* can be understood in many ways. Here, we consider the case where we observe numerous host units infected by an endemic or epidemic infectious disease, and the question of *how do pathogens spread?* translates into *who infected whom?* or *who is closely related to whom?* in the disease transmission dynamics. Host units typically designate individuals but can also designate groups such as households, premises and agricultural fields.

For fast-evolving pathogens, numerous approaches exploiting pathogen sequence data have been developed with the aim of inferring who infected whom or who is closely related to whom. These approaches are grounded on a wide variety of principles, from those based on statistical metrics to those

based on a mechanistic modelling of pathogen transmission and micro-evolution. For instance, transmission links can be inferred by identifying specific variants shared by different hosts or minimizing differences in single nucleotide polymorphisms (SNP) [1–3], by combining minimal genetic distances between intra-host viral populations and properties of social networks relevant to pathogen spread [4], by applying methods based on phylogeny, phylogeography and some forms of birth–death processes [5–14], or by using methods based on joint models of epidemiological dynamics and evolutionary processes [15–21]. Initially, model-based approaches mostly exploited a single pathogen sequence per host. Nevertheless, the progress of sequencing techniques revealing the within-host genetic polymorphism of pathogens fostered the development of model-based approaches accounting for the generation of within-host diversity and/or leveraging the information provided by sets of sequences sampled from hosts [4–7,9,14,20].

Approaches based on a mechanistic vision of transmission and micro-evolutionary processes are the most obvious direction to follow for inferring epidemiological links between host units. Indeed, mechanistic assumptions underlying these approaches act as relevant constraints, which are expected to guide the inference. However, statistical learning techniques [22] adapted to the inference of epidemiological links should also be developed, in particular (i) when mechanistic assumptions could be inadequate and, therefore, misleading, (ii) when sequence data do not accurately reflect the within-host pathogen population because of sequencing bias or errors and (iii) when a fast method is required to tackle big datasets in terms of number of hosts, sequencing depth and sequence length.

Here, we propose a statistical learning approach for estimating epidemiological links from deep sequencing data (called SLAFEEL), which is based on a parsimonious semi-parametric pseudo-evolutionary model. This model is designed as a regression function where the response variable is the set of sequences $\mathbf{S}$ observed from a recipient host unit and the explanatory variable is the set of sequences $\mathbf{S}_0$ observed from a putative source. The coefficients of the regression are weights measuring how much each sequence in $\mathbf{S}_0$ contributes to explaining each sequence in $\mathbf{S}$. These weights account for the gain and loss of virus variants during within-host evolution and their loss during between-host transmission. The model is semi-parametric because it depends both on parameters and on a kernel smoother (a tool from non-parametric statistics), which accounts for unsampled sequences in the source of infection, the evolution of new viral variants and potential sequencing errors. The model is pseudo-evolutionary because, even if it does not explicitly model evolutionary processes, it contains terms that macroscopically reflect these processes. From this model, we built a penalized pseudo-likelihood, which is used for selecting who infected whom (or who is closely related to whom). Two hypotheses (H1 and H2) were considered for the penalization. H1: The penalization assesses whether the contributions of sequences in $\mathbf{S}_0$ to explain sequences in $\mathbf{S}$ are homogeneous (two penalization shapes were introduced in this case: H1-normal and H1-χ^2). H2: The penalization assesses whether the distance between sequences in $\mathbf{S}$ and their contributing sequences in $\mathbf{S}_0$ is consistent with some known features, e.g. with an expected value for this distance (one penalization shape was introduced in this case: H2-normal). In both cases, a penalization parameter measures the strength of the penalization, and this parameter is calibrated with training data. In the epidemiological contexts tackled in this study, training data consist of contact tracing (who has been in contact with whom) or geographical distances between host units (that can be viewed as a contact proxy). Contact information has to be available only for a subset of hosts, hereafter called *training hosts*. Finally, for each putative donor–recipient pair, our method provides a link intensity measuring whether the set $\mathbf{S}_0$ collected from the putative donor likely explains the set $\mathbf{S}$ collected from the recipient. In addition, the link intensities can enable an assessment of the uncertainty of the reconstruction of donor–recipient links.

In what follows, we pave the way for this statistical learning approach aiming at inferring transmissions of infectious diseases (caused by fast-evolving pathogens) from deep sequencing data, and we apply it to three real cases in animal, human and plant epidemiology. The animal case study concerns swine influenza virus (SIV) and here serves as a test study since the transmission chain is partly known. The human case study, dealing with Ebola, is a particularly challenging situation since little diversity is observed in the pathogen population and limited contact tracing information is available. The plant case study concerns a potyvirus of wild salsify transmitted by aphids where the host unit is the meadow. In this latter application, we are more interested in estimating who is closely related to whom than who infected whom. The generic nature of SLAFEEL allows dealing with diverse epidemiological situations and sequencing procedures, as illustrated by the three case studies and in §3 of this article.

2. Results

(a) Tracing experimental swine influenza outbreaks

The first dataset was generated from an experimentally controlled transmission chain of SIV in pigs with different immunological histories (naive and vaccinated; [2]). For each chain, pairs of pigs were successively settled in an experimental enclosure, with a temporal overlap between the arrival of the new pair and the departure of the preceding pair to allow the virus to be transmitted. Thus, the infection pathways are partly known and will be used to assess the efficiency of SLAFEEL. For each pig, the virus population was sampled on a daily basis, and multiple clones of the hemagglutinin gene were sequenced using a capillary approach (Sanger sequencing). The naive chain consisted of five pairs of pigs from which 21 samples of the viral populations were collected with multiple time points for eight pigs. The vaccinated chain consisted of seven groups of pigs from which 29 samples of the viral populations were collected with multiple time points for seven pigs. Further details about the SIV dataset are provided in electronic supplementary material, table S1.

Transmission chains were inferred for the two experimental outbreaks with SLAFEEL. The penalization was calibrated for each outbreak with contact information from two *training* hosts, which were either the two pigs of the last group of the outbreak or a pig from the third group and a pig from the fourth group. The training hosts and the hosts with which they have been in contact, including the host in the same group, are detailed in electronic supplementary material, table S2. For this application, we chose the H1-normal penalization (see §4b) that led to higher consistency between contact information and inferred transmissions.

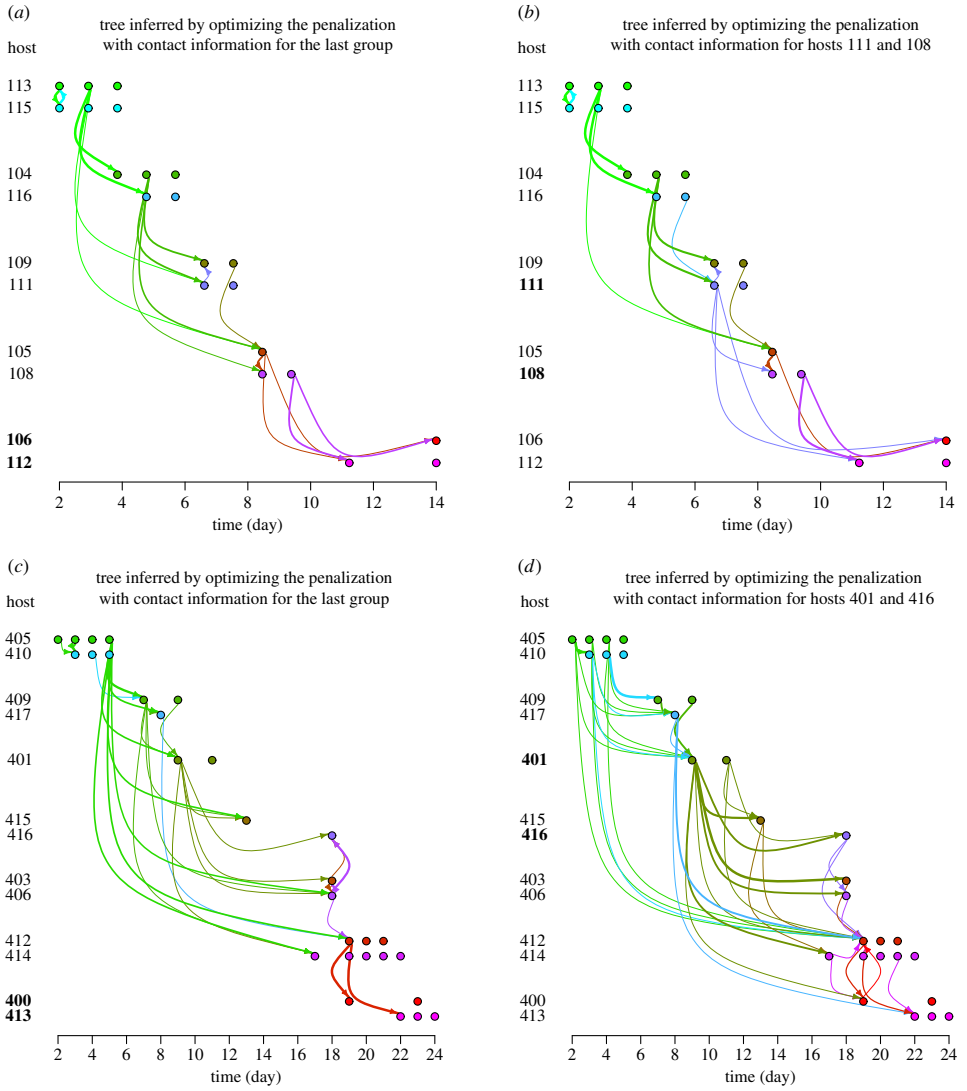


Figure 1. Transmissions inferred in the naive and vaccinated chains with two different pairs of training hosts for calibrating the penalization. Panel (*a*) corresponds to the naive chain using pair 106–112 as training hosts (i.e. the last group of the chain); (*b*) naive chain, pair 111–108; (*c*) vaccinated chain, pair 400–413 (i.e. the last group of the chain); (*d*) vaccinated chain, pair 401–416. Training hosts are written in bold. The thickness of each arrow is proportional to the intensity of the corresponding inferred link. (Online version in colour.)

For each host, the *response* set of sequences was the first sample collected from this host, and the potential *explanatory* sets of sequences were every sample collected earlier or at the same time from all the other hosts.

Figure 1 shows transmissions inferred with SLAFEEL for the naive and vaccinated chains. For the naive chain, we observe rather consistent estimations with the two pairs of training hosts, even if we observe variation in secondary links with low intensities displayed with thin arrows (the link intensity measures the likelihood of the link; see §4c). By contrast, for the vaccinated chain, the training hosts have an impact on the inference. Indeed, the use of training hosts in the last group leads to the identification of many indirect links as transmissions, whereas the use of training hosts in the middle of the chain reduces this shortcoming (even if the sources for hosts 403, 406, 412 and 414 remain inadequately inferred). Electronic supplementary material, figure S1 shows how this uncertainty is also reduced by adding a third training host to the last group. Using more contact information allows a finer calibration of the penalization (electronic supplementary material, figure S2) and, consequently, a more accurate resolution of transmissions. Moreover, the advantage of introducing a penalization is clearly illustrated by electronic supplementary material, figure S3, which displays transmissions estimated without penalization: for the naive chain, host 113 is erroneously identified as the source of infection of numerous hosts.

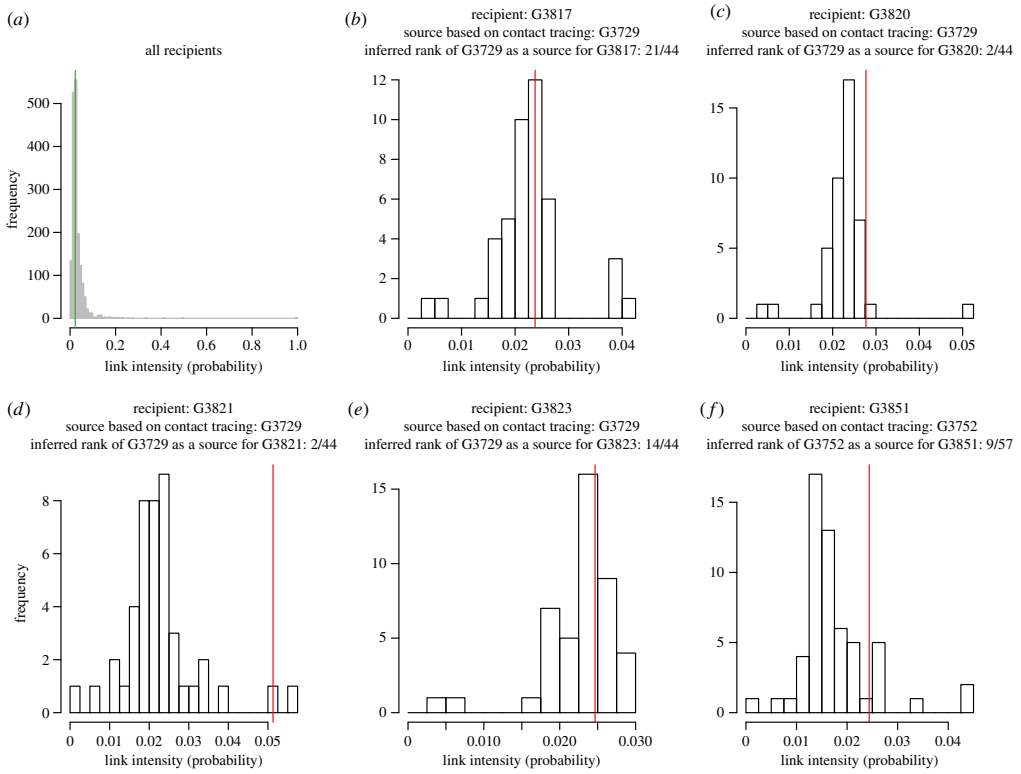


Figure 2. Estimated intensities of links for all recipients (*a*; vertical line: median intensity) and for each recipient in the training set of hosts (*b*–*f*; vertical lines: intensity for the source identified with contact tracing). This figure was obtained from the combined analysis of 31 sequence fragments and with cross-validation. Analogous figures obtained without cross-validation and with half of the fragments are given in electronic supplementary material, figures S6–S8. The second half of fragments led to approximately the same results for training hosts. Note in addition that using only one fragment for inferring transmissions led to particularly stochastic outputs. (Online version in colour.)

(b) Inferring Ebola epidemiological links despite low pathogen diversity

In this section, we analyse the dataset generated during the 2014 Ebola virus disease (EVD) outbreak in Sierra Leone [23]. We were able to include in our analysis 58 confirmed EVD patients, from which within-host populations of the virus were collected and sequenced. This number of patients represents nearly 50% of the EVD patients diagnosed in Sierra Leone from late May to mid-June. Viral populations were sequenced using the Nextera library construction method and Illumina sequencing and the haplotypes were estimated in a sliding window of 1000 bases every 500 bases using Predict-Haplo [24].

More details about the Ebola dataset are provided in electronic supplementary material, table S1. Here, we simply highlight the rather low pathogen diversity that was observed: on average, 16.1 haplotypes per fragment of 1000 bases were identified for the 58 patients included in the analysis (s.d. = 8.0), and 1.37 haplotypes per fragment of 1000 bases per patient (s.d. = 0.64).

Epidemiological links between patients were inferred by calibrating the penalization with contact tracing published in [25]. We were able to use five donor–recipient *training* pairs identified with contact tracing (see electronic supplementary material, table S2), four of them having the same putative donor. For this application, we chose the H2-normal penalization (see §4b), which led to higher consistency between contact information and inferred transmissions in a situation where observed pathogen populations show relatively low levels of diversity. Several samples were available for some of the patients collected at different time points [23]. These samples were merged in our analysis to increase the within-host sequence diversity. In addition, we applied the statistical learning approach separately for 31 partly overlapping fragments of 1000 nucleotides, and we aggregated the results for reconstructing the epidemiological links. For each host, potential sources were inferred among patients observed earlier than or at the same time as the target host (point discussed in §3).

Because of the reduced pathogen diversity, the inferred intensities of epidemiological links are generally quite low (figure 2*a*) and multiple sources for any host are plausible (except those at the earlier time points of sampling for which only a few potential sources are allowed). Thus, source identification is quite uncertain. Figure 2*b*–*f* shows the distributions of the link intensities with plausible sources for the five recipients in the training data, and give the ranks of their sources identified with contact tracing. The intensities and ranks were inferred with a leave-one-out cross-validation approach (i.e. the host of interest in each panel is removed from the training data when one infers its source and the

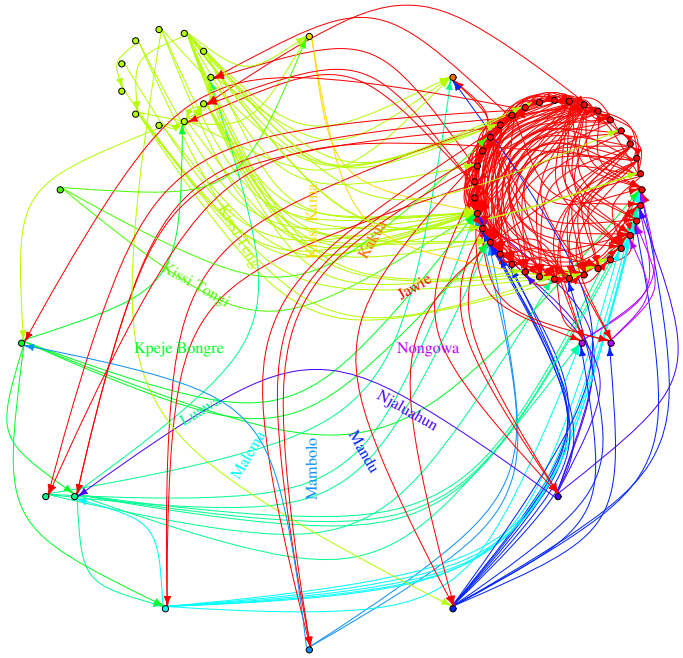


Figure 3. Most likely epidemiological links cumulating to 20% probability for each recipient (i.e. for each recipient, potential donors were ranked with respect to link intensity, and the subset of donors with higher ranks for whom the sum of link intensities reached 0.2 were displayed on the graph). (Online version in colour.)

rank of its donor based on contact-tracing). The donors identified with contact tracing are well ranked for patients G3820, G3821, G3823 and G3851, but not for G3817. The pathogen population collected from the latter patient is actually quite different from the population observed in its putative donor G3729 (see electronic supplementary material, table S3, and the Ebola phylogeny built from the consensus sequences [26]). Thus, the epidemiological link between G3817 and G3729 could be revisited by focusing on patients who are more closely connected to G3817 than G3729 (see electronic supplementary material, tables S4–S8). Figure 3 displays the most likely epidemiological links cumulating to 20% of probability for each recipient (see figure caption). Patients are clustered based on their chiefdoms, whose locations are provided in electronic supplementary material, figure S4. The Jawie chiefdom seems to be an interface between Kissi Teng and Kissi Tongi chiefdoms on the one hand and most of the other chiefdoms on the other hand. Based on temporal data (electronic supplementary material, figure S5), the Kissi Teng and Kissi Tongi chiefdoms include mostly early cases and, therefore, individuals in Jawie chiefdom may have played the role of a relay in the outbreak.

(c) Assessing epidemiological links at the metapopulation scale

This dataset was generated from a wild plant species (*Tragopogon pratensis*, hereafter called wild salsify), which is a reservoir for a potyvirus closely related to the endive necrotic mosaic virus (ENMV; [27]). Within-host virus variants were sequenced from 189 infected host plants sampled in 2014 in a 40×10 km region of south-eastern France. High-throughput sequencing was applied on viral PCR amplicons (final length: 438 bp of the capsid gene) using the Illumina technology [28]. Sequence data were merged at the scale of the patch (i.e. meadows, agricultural fields or urbanized areas) with the aim of assessing epidemiological links between a subset of the metapopulation formed by the potyvirus (the 189 sampled plants were distributed in 27 patches). Further details about this dataset are provided in electronic supplementary material, table S1.

Epidemiological links between sampled patches were inferred by calibrating the penalization with information on inter-patch distances, assuming that, on average, geographically close host patches are infected by similar viral variants (isolation-by-distance process). Here, the H1-χ^2 penalization (see §4b) was chosen because it led to a lower average distance between connected patches (see criterion (4.7), §4c).

Figure 4 shows the inferred links between sampled patches. Here, all the optimal values for the penalization parameter (shown in electronic supplementary material, figure S9) led to the same set of links and, therefore, no secondary arrows are displayed (electronic supplementary material, figure S10 shows links inferred without penalization). Even if most links are relatively short compared to the mean distance between sampled patches (see electronic supplementary material, figure S11), there is a non-negligible proportion of long links that could be the signature of the long-distance dispersal ability of the aphid to transmit the virus. Additionally, common environmental conditions and host demography and genetics at the scale of the study area may partly explain the inferred long-distance links. Indeed, environmental conditions constrain host local abundance

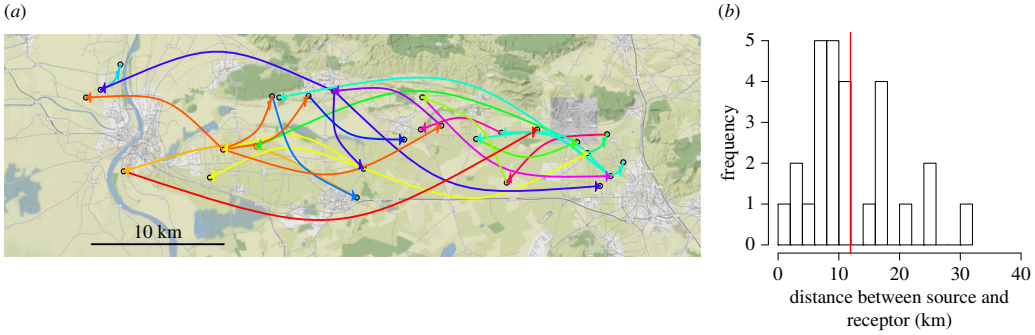


Figure 4. Links inferred between salsify patches based on sampled sets of potyvirus sequences (*a*; links from the same source have the same colour) and distribution of link distances (*b*; the vertical red line gives the mean distance). (Online version in colour.)

and, therefore, genetic drift impacts on the levels of diversity and differentiation within and between local pathogen populations. Spatial variation in host genetics may also shape the spatial structure of pathogen populations by selecting different variants regardless of the distance between host patches [29,30].

(d) Benchmarking SLAFEEL

We first compared SLAFEEL and BadTrIP [5] for influenza data to assess the ability of both methods to identify infection pathways that are partly known. Electronic supplementary material, figure S12, gives details about the application of BadTrIP and shows inferred transmission trees. Whatever training hosts were used, SLAFEEL generally performed better than BadTrIP with respect to the proportion of correct source identifications (that focuses on the most likely inferred source) and the average Jeffreys discrepancy (that compares the probabilities for any recipient host to be linked with any putative source) as presented in electronic supplementary material, table S10.

Second, we compared the transmissions inferred with SLAFEEL from the Ebola data and those obtained in [5] with BadTrIP. Here, we assessed the consistency of both estimations (since potential infection pathways are not known, unlike in the influenza case study). The most likely sources are the same for 8% of recipient hosts (electronic supplementary material, table S10) and the most likely sources inferred with SLAFEEL are among the 10 most likely sources identified with BadTrIP for almost 50% of recipients (electronic supplementary material, figure S13). These rather low percentages may be explained by the low pathogen diversity in this study, leading to generally quite low inferred link intensities with SLAFEEL and, to a lesser extent, with BadTrIP (see electronic supplementary material, figure B in [5]). They may also be explained by the assumptions made and the constraints imposed in [5], where information from sampling dates, nucleotide frequencies and sequencing coverage was used, and where the introduction date (removal date) of each host was specified as its sampling date minus (plus) 21 days, thus allowing each host to be infected at most 21 days before being sampled, and to infect others at most 21 days after being sampled.

Finally, we simulated 1000 datasets with the SEEDY package (simulation of evolutionary and epidemiological dynamics; [20]) by using parameter values chosen by Worby and Read to generate their 4th figure (mean epidemic size: 26.6 infected hosts (s.d. = 2.3); 10 virus genomes sampled per host). The SEEDY package allows not only the generation of datasets, but also a very fast inference of transmissions given infection times, the mutation rate, the equilibrium viral population size within host and the transmission bottleneck size, which are generally not known in practice. Thus, we used SEEDY-based inferences of transmissions as a benchmark, and assessed how SLAFEEL compares with SEEDY in identifying the true source for each recipient of each of the 1000 simulated outbreaks. For the application of SLAFEEL to each simulated outbreak, we randomly drew four training hosts whose sources were supposed to be known, and we chose the H1-normal penalization. On average, the most likely inferred source was correct for 39% [20–61%] of recipients with SEEDY and 36% [17–60%] with SLAFEEL (electronic supplementary material, figure S14). Therefore, in this simulation setting, SLAFEEL performs almost as well as SEEDY.

3. Discussion

We introduced an exploratory approach, called SLAFEEL, for quantitatively investigating epidemiological links between host units from deep sequencing data. This versatile approach, grounded on statistical learning, is adaptable to diverse contexts and data. Here, we applied it to analyse virus dynamics in humans, animals and plants at different spatial scales (e.g. individuals and fields) using data obtained with different sequencing techniques and showing different levels of pathogen diversity. The relatively broad applicability of SLAFEEL implies that, in some contexts, links have to be interpreted in a conservative way: typically, in the salsify potyvirus application, we did not infer *who infected whom* but *who is closely related to whom*. Using the pseudo-evolutionary model and the associated inference approach for estimating epidemiological links should be particularly valuable in non-standard situations where classical mechanistic assumptions may be erroneous and when sequencing and variant calling issues may be misleading. The key property underlying our procedure is the combination of a learning stage and a penalization that can be used to constrain what is a link. This is expected to help in

appropriately dealing with sequencing errors because such errors should be accounted for non-training hosts as they are for training hosts. Nevertheless, as discussed below, the impact of sequencing errors on inference accuracy should be formally assessed in simulation studies.

The training stage can use classical information such as contact tracing data [25], but also contact proxies such as geographical distances between host units, connectivities via air masses for airborne pathogens [31] and social connections [4,32]. To get a contact proxy, one could also infer some transmissions with a (generally more time-consuming) mechanistic approach from a subset of observed cases and use the estimated transmissions as training data in our approach applied to the whole dataset. Thus, the mechanistic approach and SLAFEEL would be complementary. Whatever the way that contact information (or proxies) are gathered, it can be conjectured that the closer the relationship between contact information and epidemiological links, the more informative the training stage. Moreover, the possibility of using very diverse types of contact information in the learning stage of SLAFEEL reinforces its broad relevance to human, animal and plant diseases.

When geographical proximity is used for calibrating the penalization (like in the potyvirus application), short-distance links may be favoured, and the inferred distribution of distances between linked host units hence has to be interpreted with caution. However, in our procedure, geographical proximity is only used after a genetic-based selection of possible configurations: basically, the penalized pseudo-likelihood function (only based on virus sequence data) allows us to eliminate genetically unlikely configurations; then, in the learning stage, spatial information is used to select the most likely configurations within the set of genetically likely configurations, building on the following grounds: among two equally genetically likely configurations, the one showing links at shorter distances is more likely (because of the very classical assumption that 'dispersal is more probable at short distance than at long distance'). Thus, inferring only short-distance links can be interpreted as: 'short distance dispersal is sufficient to explain the genetic spatial pattern of the pathogen'. By contrast, inferring both (i) a mixture of short- and long-distance links and (ii) unlinked nearby host units (like in the potyvirus application) suggests that isolation by distance does not hold at the study scale, and that the assumption 'dispersal is more probable at short distance than at long distance' is perturbed by other drivers (e.g. host genetics), which significantly impact the genetic spatial pattern of the pathogen. Finally, while our analysis in the potyvirus application leads to interpretable results, cross-validation or data-splitting (into training and prediction data) could be applied in further studies to strengthen the analysis conclusions when geographical proximity is used as contact information.

The main objective of this article was to present how statistical learning can be applied for inferring transmissions (or epidemiological links from a conservative perspective) and to examine if such an approach has the potential to be efficient. Results obtained for swine influenza (where the transmission pathways are partly known) and for outbreaks simulated with SEEDY [20] are encouraging. However, further research is required to make the method robust and able to pass a battery of simulation tests such as the one designed for assessing the performance of BadTrIP [5]. The following questions should be specifically investigated using simulations. How does the efficiency and speed of the method scale up with big data? How does the method perform at various sequencing depths (considering a single haplotype for each host as a special case)? How does the method perform in the presence of contamination and sequencing errors (PHYLOSCANNER [14] explicitly handles such issues)? What is the sensitivity of the method to the haplotype reconstruction tool (e.g. comparing Predict-Haplo that we used for the Ebola data with SAVAGE [33] and MLEHaplo [34])? How is SLAFEEL accuracy improved with increasing training information? How can we exploit negative training information (i.e. infected hosts that are known to not have been in contact with certain infected hosts)? How does the method perform in the presence of severe bottlenecks during transmissions, in comparison with approaches exploiting phylogenetic signals that are particularly adapted to such situations [9]?

Before testing SLAFEEL in the latter range of simulation settings, further research should especially focus on the penalization function. Here, we introduced three shapes corresponding to different hypotheses (see §4b), but the penalization could be tuned by considering other hypotheses, which could help circumvent the current limitations of our approach. For instance, the penalization could be improved to take into account (i) the timing, thus constraining the set of likely sources for each host based on observation times and possibly additional temporal information like data on infectious periods [17], (ii) fixed sub-clonal haplotypes (including haplotypes with stop codons) by forcing the selection algorithm to pair host units sharing such haplotypes [1,35] and (iii) sample sizes to avoid biases induced by different levels of observed diversity. Specific penalizations could also be designed to better infer the direction of epidemiological links when temporal data do not discriminate sufficiently. For example, the signature of the link direction could be identified in the genetic training data and incorporated into the penalization function. Other limitations are more difficult to tackle, e.g. de novo mutations at the same site (homoplasy), recombinations, insufficient sequencing depth and lack of sequence diversity, which can lead to uncertainty in the inferences. However, the advantage of our statistical learning approach is that the uncertainty can be objectively assessed on training data. The uncertainty (and potential bias) can even be assessed using cross-validation to prevent over-fitting. The assessment of uncertainty and bias in the inference of links is also an objective way to select the penalization shape. However, we must warn that, if training data are not representative of the whole population, learning model parameters from training data may induce errors in the selection of the penalization and, ultimately, in the reconstruction of epidemiological links (such misleading training data would be analogous to misleading assumptions in mechanistic approaches).

Another important perspective is the implementation of an efficient computer code. The R code that we developed (available at https://doi.org/10.5281/zenodo.1410438) allowed us to test different model specifications, to exploit genetic data from multiple sequence fragments and to perform cross-validation in a limited time-span (e.g. a SLAFEEL run for the swine influenza case study or for a sequence fragment in the Ebola case study took approximately 10–20 minutes with a laptop computer, whereas

BadTrIP takes several days; see caption of electronic supplementary material, figure S12 and [5]). However, implementing further improvements in the code should allow us (i) to include multiple infections in transmission scenarios where an *explanatory* set of sequences would consist of a weighted mixture of several samples collected from several putative sources, (ii) to select a penalization shape among a large library of functions, and (iii) to tackle big data (e.g. large numbers of cases and sequence fragments). Concerning point (iii), our approach based on a simplified representation of dependencies between observations via a statistical regression model is a commonly used approach to handle big data [36].

4. Methods

To infer transmissions of a virus (or, more generally, epidemiological links) within a host population, we built a pseudo-evolutionary model that concisely describes transitions between sets of sequences sampled from different host units, and used this model to select probable source–recipient pairs. In what follows, we provide the outline of our method in one of its simplest forms (see also electronic supplementary material, figure S15), then we technically describe it in its general form by presenting first the model and second the inference.

(a) Outline of the SLAFEEL approach

Let us consider one of the possible source–recipient pairs. For each virus sequence collected from the recipient, we compute the genetic distance (namely, the number of different nucleotides) to each sequence collected from the source, and we identify the nearest sequence(s). By applying this procedure to all sequences from the recipient, we can compute the contribution of each sequence from the source to *explain* the viral population observed from the recipient. This contribution relates to the number of times that this sequence from the source is identified as the nearest sequence (see the exact definition in §4b). Then, a parametric kernel function, derived from the Jukes–Cantor micro-evolutionary process and embedded in a pseudo-likelihood, is used to assess how much each sequence from the recipient is *explained* by its nearest sequence(s) from the source. Moreover, a parametric penalization function is used to assess how likely sequences from the source have been *uniformly* subsampled to generate sequences from the recipient (this is assessed based on the contributions calculated above). Thus, for each possible source–recipient pair, we compute a penalized pseudo-likelihood parameterized by the kernel parameter μ and the penalization parameter θ. The penalized pseudo-likelihood will be high for a putative source–recipient pair if (i) all sequences from the recipient have genetic neighbours in the source and (ii) sequences from the source equally contribute in expectation to the set of sequences collected from the recipient. Note that condition (ii) depends on the rationale underlying the form chosen for the penalization function (here, the penalization is grounded on a uniform subsampling hypothesis).

The balance between the pseudo-likelihood and the penalization is tuned in two steps. First, we estimate μ, for each source–recipient pair and each θ value in a set Θ of candidate values, by maximizing the penalized pseudo-likelihood with respect to μ; then, for each recipient and each θ value, the source leading to the maximum penalized pseudo-likelihood is identified as the most likely source given θ. Second, adopting a learning approach, we calibrate the penalization by selecting the θ values leading to the maximum proportion of *training hosts* for which the most likely sources conditional on θ are consistent with contact information. The link intensity between a given recipient and a possible source is measured by the proportion of selected θ values for which the source has been identified as the most likely source.

The dual form of the penalized pseudo-likelihood and the learning stage are essential to distinguish 'A infected B', 'B infected A' and 'C infected B' when only the former statement is true. Indeed, the pseudo-likelihood tends to impose that each sequence from the recipient must have a neighbour sequence in its source, which should exclude 'C infected B'; the penalization tends to impose that the set of sequences from the recipient has been generated by a subsample of the set of sequences from the source (if the penalization has been built in this way), which should exclude 'B infected A'; the learning stage is expected to determine the adequate relative weights of the pseudo-likelihood and the penalization for obtaining satisfactory inference of epidemiological links. The learning stage can even be exploited to design an adequate penalization form (one should prefer a penalization form leading to higher inference accuracy for training hosts).

(b) Pseudo-evolutionary model for the evolution and transmission of populations of sequences

The method outlined above is grounded on a pseudo-evolutionary model, which concisely describes transitions between sets of sequences sampled from different host units. The general form of the pseudo-evolutionary model is given by the following penalized pseudo-likelihood for the transition from an explanatory set of I sequences $S_1^{(0)}, \ldots, S_I^{(0)}$ to a response set of J sequences $S_1, \ldots, S_J$ (haplotype copies are explicitly incorporated in these sets of sequences):

$$f(S_1, \ldots, S_J \mid S_1^{(0)}, \ldots, S_I^{(0)}) = P(W) \prod_{j=1}^{J} \left(\frac{\sum_{i=1}^{I} w_{ij} K(d(g(S_j), g(S_i^{(0)})))}{\sum_{i=1}^{I} w_{ij}} \right), \tag{4.1}$$

where each term in the product represents the pseudo-probability of obtaining the response sequence S_j given the explanatory sequences $S_1^{(0)}, \ldots, S_I^{(0)}$ and the values of $w_{1j}, \ldots, w_{Ij}$; g is a transformation of sequences (e.g. aiming at reducing the dimension of the space of viral sequences); K is a kernel function and d is a pseudo-distance function introduced to account for unsampled sequences in the source of infection, the evolution of new viral variants and possible sequencing errors; w_{ij} are weights accounting for the loss of virus variants during within-host evolution and between-host transmission; W is the $(I \times J)$-matrix of weights whose element (i, j) is w_{ij}; and $P(W)$ is a penalty for the weight matrix W potentially allowing the incorporation of knowledge on virus evolution and transmission (e.g. on the strength of the transmission bottleneck).

In this article, we focus on a simple semi-parametric version of (4.1) where (i) each sequence S_j is only explained by the closest sequence(s) $S_i^{(0)}$ in terms of the number of different nucleotides and (ii) the penalization measures the discrepancy from a null hypothesis to be specified. Thus, the pseudo-evolutionary model given by equation (4.1) reduces to:

$$f_{\mu,\theta}(S_1, \ldots, S_J \mid S_1^{(0)}, \ldots, S_I^{(0)}) = P_\theta(W) \prod_{j=1}^{J} \left(\frac{\sum_{i=1}^{I} w_{ij} K_\mu\{d(S_j, S_i^{(0)}); \Delta_{ij}\}}{\sum_{i=1}^{I} w_{ij}} \right), \tag{4.2}$$

where $d(\cdot, \cdot)$ gives the number of different nucleotides between two sequences; $w_{ij} = 1/n_j$ for indices i corresponding to sequences $S_i^{(0)}$ minimally distant from sequence S_j, i.e. such that $d(S_j, S_i^{(0)}) = \min\{d(S_j, S_{i'}^{(0)}) : i' = 1, \ldots, I\}$, the number of such sequences being denoted n_j, $w_{ij} = 0$ otherwise (therefore, $\sum_{i=1}^{I} w_{ij} = 1$); Δ_{ij} is the duration separating the two sequences

S_j and $S_i^{(0)}$; $K_\mu(\cdot\,;\Delta)$ is the probability distribution function (p.d.f.) of the binomial law with size L (i.e. sequence length) and success probability $3(1-\exp(-4\mu\Delta))/4$, corresponding to the Jukes–Cantor micro-evolutionary process over a duration Δ and with a substitution parameter μ; and $P_\theta(W)$ is a parametric penalization measuring the likelihood of the contributions of explanatory sequences $S_1^{(0)}, \ldots, S_I^{(0)}$ (measured by $\sum_{j=1}^J w_{ij}$, $i = 1, \ldots, I$) to the response set of sequences $S_1, \ldots, S_J$. If $\sum_{j=1}^J w_{ij} = 0$, then sequence $S_i^{(0)}$ does not contribute to explaining the sequences collected from the recipient and, therefore, may be considered as lost during within-host evolution or between-host transmission.

We consider the three following shapes for P_θ. The H1-normal shape measures the discrepancy between $\sum_{j=1}^J w_{ij}$ and its expected value J/I under the uniform (but not necessarily independent) sampling hypothesis by

$$P_\theta(W) = \prod_{i=1}^{I} \Phi\left(\sum_{j=1}^{J} w_{ij};\ \frac{J}{I},\ \theta\frac{J}{I}\left(1-\frac{1}{I}\right)\right), \tag{4.3}$$

where $\Phi(\cdot\,; a, b^2)$ is the p.d.f. of the normal law with mean a and variance b^2, and $\theta(J/I)(1-1/I)$ is proportional to the multinomial variance up to the over-dispersion parameter $\theta > 0$. The uniform sampling hypothesis amounts to assuming that explanatory sequences have equal chances to contribute to the set of response sequences. With J response sequences, there are J draws of an explanatory sequence (one for each response sequence) among I explanatory sequences. Thus, under the uniform sampling hypothesis, the total contribution $\sum_{j=1}^J w_{ij}$ of the explanatory sequence $S_i^{(0)}$ has expectation J/I.

The H1-χ^2 shape measures the discrepancy between $\sum_{j=1}^J w_{ij}$ and its expected value J/I by

$$P_\theta(W) = \theta\chi^2\left(\sum_{i=1}^{I}\frac{\left(\sum_{j=1}^J w_{ij} - J/I\right)^2}{J/I};\ I-1\right), \tag{4.4}$$

where $\chi^2(\cdot\,; I-1)$ is the p.d.f. of the χ^2 law with $I-1$ degrees of freedom, and $\theta > 0$ measures the influence of the penalization.

The H2-normal shape can be used when estimates of the mean and standard deviation of the distance between any sequence collected from any recipient host and the closest sequence collected from its source, say $\bar{d}_{\text{obs}}$ and σ^2_{obs}, are available (these estimates can be obtained from contact tracing data). The H2-normal shape measures how likely it is that this mean distance for the host unit of interest is drawn from the normal distribution with mean $\bar{d}_{\text{obs}}$ and variance σ^2_{obs}:

$$P_\theta(W) = \theta\prod_{j=1}^{J}\Phi\left(\sum_{i=1}^{I} w_{ij} d(S_j, S_i^{(0)});\ \bar{d}_{\text{obs}},\ \sigma^2_{\text{obs}}\right), \tag{4.5}$$

where $\theta > 0$ measures the influence of the penalization.

Thereafter and whatever the penalization shape, θ is called the penalization parameter.

(c) Estimation and calibration of parameters, and inference of transmissions

Consider M sets of sequences $\mathbf{S}_1, \ldots, \mathbf{S}_M$ collected from M host units. In a first step, for each set of sequences $\mathbf{S}_m$ and each value of θ in a finite set Θ to be specified, the penalized pseudo-likelihoods $f_{\mu,\theta}(\mathbf{S}_m|\mathbf{S}_{m'})$, for $m' \neq m$, are maximized with respect to μ (let $\hat{\mu}_{m'}(\theta)$ denote the maximizer, i.e. the estimate, of μ). The most likely source for host unit m given θ, say $\hat{s}(m;\theta)$, is then the host unit m' leading to the highest value of $f_{\hat{\mu}_{m'}(\theta),\theta}(\mathbf{S}_m \mid \mathbf{S}_{m'})$:

$$\hat{s}(m;\theta) = \underset{m'\neq m}{\operatorname{argmax}}\ f_{\hat{\mu}_{m'}(\theta),\theta}(\mathbf{S}_m \mid \mathbf{S}_{m'}).$$

In a second step, the penalization parameter θ is calibrated by building and optimizing a criterion that compares contact information and inferred sources of infection $\hat{s}(m;\theta)$, for m in a set $\mathcal{M} \subset \{1, \ldots, M\}$ of *training* hosts (this procedure can also be used in practice to select a penalization shape among a set of candidate functions as those proposed in equations (4.3)–(4.5)). Driven by the applications in this study, we introduce the two following criteria. First, consider the case where contact information consists of tracing contacts for hosts $m \in \mathcal{M}$. We define the criterion to be maximized as the proportion of inferred transmissions that are consistent with contact tracing:

$$\tilde{\Theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\mathbf{1}(\hat{s}(m;\theta)\in\mathcal{C}_m), \tag{4.6}$$

where $|\mathcal{M}|$ is the number of elements in $\mathcal{M}$; $\mathbf{1}(E) = 1$ if event E is true, zero otherwise; and $\mathcal{C}_m$ is the set of hosts in $\{1, \ldots, M\}$ that have been in contact with m. Second, consider the case where contact information consists of the geographical distances between hosts in the training set $\mathcal{M} \subset \{1, \ldots, M\}$. We define the criterion to be minimized as the average distance between the training hosts and their inferred sources (if the sources are in the training set):

$$\tilde{\Theta} = \underset{\theta\in\Theta}{\operatorname{argmin}}\frac{\sum_{m\in\mathcal{M}}\delta(m,\hat{s}(m;\theta))\mathbf{1}\{\hat{s}(m;\theta)\in\mathcal{M}\}}{\sum_{m\in\mathcal{M}}\mathbf{1}\{\hat{s}(m;\theta)\in\mathcal{M}\}}, \tag{4.7}$$

where $\delta(m,\hat{s}(m;\theta))$ is the geographical distance between host m and its suspected source $\hat{s}(m;\theta)$. Note that, in both cases, $\tilde{\Theta}$ may be a set of values (and not only a single value) if the criterion is optimal for several θ in Θ. This was the case in the applications that we tackled, since criteria in (4.6) and (4.7) have values in very limited discrete sets (e.g. $\{0, 1/5, 2/5, 3/5, 4/5, 1\}$ in the Ebola application). Thus, in each application, $\tilde{\Theta}$ was obtained by computing the criterion on a regular grid of θ values and by retaining only values maximizing the criterion. We observed that small variations in θ did not impact the criterion value, as well as link intensities defined below in (4.8), and the mesh size of the grid was tuned accordingly. In further applications, the grid search could be improved in two directions: first, one could use an iterative numerical algorithm for the optimization; second, one could replace the maximum/minimum rule by a quantile rule (i.e. using a tolerance threshold).

In a third step, we assess the intensity of the link between m and m' in $\{1, \ldots, M\}$ by the proportion of values of θ in $\tilde{\Theta}$ for which $\hat{s}(m;\theta)$ coincides with m':

$$\frac{1}{|\tilde{\Theta}|}\sum_{\theta\in\tilde{\Theta}}\mathbf{1}\{\hat{s}(m;\theta) = m'\}, \tag{4.8}$$

where $|\tilde{\Theta}|$ is the number of elements in $\tilde{\Theta}$. This intensity of the link between two host units is used to infer who infected whom or, from a more conservative perspective, who is the most related with whom. When several sequence fragments are available (like in the Ebola case study), the link intensity defined in equation (4.8) is computed for each fragment, and then averaged to obtain the overall link intensity. Future work could explore alternatives to the average (e.g. robust mean and median) for assessing link intensities from several fragments.

Model and inference specifications that were used for the three case studies are summarized in electronic supplementary material, table S9.

Data accessibility. Data available at: http://doi.org/10.5281/zenodo.2543673 [37].

Authors' contributions. S.S., J.H. and G.T. conceived the methodology, M.A. and S.S. implemented it and analysed data. J.H. and K.B. prepared data. S.S. led the writing of the manuscript. M.A., J.H., K.B., C.D., G.T. and S.S. all contributed to interpretation of results and preparation of drafts, and gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. This work was funded by an ANR grant (SMITID project; ANR-16-CE35-0006). J.H. is funded by the Medical Research Council (MC_UU_12014/12). Field and laboratory work for the plant virus was funded by the Division for Plant Health and Environment (SPE) of INRA through the AAP-SPE-2014 framework.

Acknowledgements. We thank Nicola De Maio for providing us with the transmissions inferred with BadTrIP from Ebola data and presented in [5].

References

1. Hughes J *et al.* 2012 Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS Pathog.* **8**, e1003081. (doi:10.1371/journal.ppat.1003081)
2. Murcia PR *et al.* 2012 Evolution of an Eurasian avian-like influenza virus in naive and vaccinated pigs. *PLoS Pathog.* **8**, e1002730. (doi:10.1371/journal.ppat.1002730)
3. Walker TM *et al.* 2013 Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146. (doi:10.1016/S1473-3099(12)70277-3)
4. Skums P *et al.* 2017 QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics* **34**, 163–170. (doi:10.1093/bioinformatics/btx402)
5. De Maio N, Worby CJ, Wilson DJ, Stoesser N. 2018 Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Comput. Biol.* **14**, e1006117. (doi:10.1371/journal.pcbi.1006117)
6. Didelot X, Gardy J, Colijn C. 2014 Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **31**, 1869–1879. (doi:10.1093/molbev/msu121)
7. Didelot X, Fraser C, Gardy J, Colijn C. 2017 Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007. (doi:10.1093/molbev/msw275)
8. Hall M, Woolhouse M, Rambaut A. 2015 Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS Comput. Biol.* **11**, e1004613. (doi:10.1371/journal.pcbi.1004613)
9. Leitner T, Romero-Severson E. 2018 Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. *Nat. Microbiol.* **3**, 983. (doi:10.1038/s41564-018-0204-9)
10. Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010 Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885. (doi:10.1093/molbev/msq067)
11. Pybus OG *et al.* 2012 Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl Acad. Sci. USA* **109**, 15 066–15 071. (doi:10.1073/pnas.1206598109)
12. Rasmussen DA, Ratmann O, Koelle K. 2011 Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol.* **7**, e1002136. (doi:10.1371/journal.pcbi.1002136)
13. Stadler T, Bonhoeffer S. 2013 Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. R. Soc. B* **368**, 20120198. (doi:10.1098/rstb.2012.0198)
14. Wymant C *et al.* 2017 PHYLOSCANNER: inferring transmission from within- and between-host pathogen genetic diversity. *Mol. Biol. Evol.* **35**, 719–733. (doi:10.1093/molbev/msx304)
15. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. 2014 Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10**, e1003457. (doi:10.1371/journal.pcbi.1003457)
16. Lau MS, Marion G, Streftaris G, Gibson G. 2015 A systematic Bayesian integration of epidemiological and genetic data. *PLoS Comput. Biol.* **11**, e1004633. (doi:10.1371/journal.pcbi.1004633)
17. Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S. 2012 A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Computat. Biol.* **8**, e1002768. (doi:10.1371/journal.pcbi.1002768)
18. Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, Soubeyrand S. 2014 A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B* **281**, 20133251. (doi:10.1098/rspb.2013.3251)
19. Soubeyrand S. 2016 Construction of semi-Markov genetic-space-time SEIR models and inference. *Journal de la Société Française de Statistique* **157**, 129–152.
20. Worby CJ, Read TD. 2015 'SEEDY' (simulation of evolutionary and epidemiological dynamics): an R package to follow accumulation of within-host mutation in pathogens. *PLoS ONE* **10**, e0129745. (doi:10.1371/journal.pone.0129745)
21. Ypma RJF, Jonges M, Bataille A, Stegeman A, Koch G, van Boven M, Koopmans M, van Ballegooijen WM, Wallinga J. 2013 Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *J. Infect. Dis.* **207**, 730–735. (doi:10.1093/infdis/jis757)
22. James G, Witten D, Hastie T, Tibshirani R 2013 *An introduction to statistical learning with applications in R*. New York, NY: Springer.
23. Gire SK *et al.* 2014 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372. (doi:10.1126/science.1259657)
24. Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V. 2014 HIV haplotype inference using a propagating Dirichlet process mixture model. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **11**, 182–191. (doi:10.1109/TCBB.2013.145)
25. Senga M *et al.* 2017 Contact tracing performance during the Ebola virus disease outbreak in Kenema district, Sierra Leone. *Phil. Trans. R. Soc. B* **372**, 20160300. (doi:10.1098/rstb.2016.0300)
26. Dudas G *et al.* 2017 Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315. (doi:10.1038/nature22040)
27. Desbiez C *et al.* 2017 Molecular and biological characterization of two potyviruses infecting lettuce in southeastern France. *Plant Pathol.* **66**, 970–979. (doi:10.1111/ppa.2017.66.issue-6)
28. Piry S, Wipf-Scheibel C, Martin JF, Galan M, Berthier K. 2017 High throughput amplicon sequencing to assess within- and between-host genetic diversity in plant viruses. bioRxiv. p. 168773. (doi:10.1101/168773)
29. Papaïx J, Burdon JJ, Lannou C, Thrall PH. 2014 Evolution of pathogen specialisation in a host metapopulation: joint effects of host and pathogen dispersal. *PLoS Comput. Biol.* **10**, e1003633. (doi:10.1371/journal.pcbi.1003633)
30. Papaïx J, Burdon JJ, Zhan J, Thrall PH. 2015 Crop pathogen emergence and evolution in agro-ecological landscapes. *Evol. Appl.* **8**, 385–402. (doi:10.1111/eva.12251)
31. Leyronas C, Morris CE, Choufany M, Soubeyrand S. 2018 Assessing the aerial interconnectivity of distant reservoirs of *Sclerotinia sclerotiorum*. *Front. Microbiol.* **9**, 2257. (doi:10.3389/fmicb.2018.02257)
32. Keeling MJ, Eames KT. 2005 Networks and epidemic models. *J. R. Soc. Interface* **2**, 295–307. (doi:10.1098/rsif.2005.0051)
33. Baaijens JA, El Aabidine AZ, Rivals E, Schönhuth A. 2017 De novo assembly of viral quasispecies using overlap graphs. *Genome Res.* **27**, 835–848. (doi:10.1101/gr.215038.116)
34. Malhotra R, Wu MMS, Rodrigo A, Poss M, Acharya R. 2015 Maximum likelihood de novo reconstruction of viral populations using paired end sequencing data. preprint arXiv:1502.04239.
35. Emmett KJ, Lee A, Khiabanian H, Rabadan R. 2015 High-resolution genomic surveillance of 2014 ebolavirus using shared subclonal variants. *PLoS Curr.* **7**, 1–17. (doi:10.1371/currents.outbreaks.c7fd7946ba606c982668a96bcba43c90)
36. Pfeiffer DU, Stevens KB. 2015 Spatial and temporal epidemiological analysis in the big data era. *Prev. Vet. Med.* **122**, 213–220. (doi:10.1016/j.prevetmed.2015.05.012)
37. Alamil M, Hughes J, Berthier K, Desbiez C, Thébaud G, Soubeyrand S. 2019 Data from: SLAFEEL: R scripts and reformatted data analyzed by Alamil *et al.* (2019) (Version 1.5). Zenodo. (http://doi.org/10.5281/zenodo.2543673)

3.3. Application to Equine influenza virus data set

Varying the penalization shape and the temporal constraints in SLAFEEL: Illustration with Equine Influenza virus data

Note: This is a preliminary work presented in a rather concise way. In particular, concerning the description of the methodological background, we only highlight the changes with respect to the SLAFEEL method described in the previous chapter.

1 Introduction

Using sequence data alone to infer transmission routes during an outbreak is rather frequent, but it may leads to a lack of accuracy in the inference (Didelot et al., 2014; Jombart et al., 2014; Hall et al., 2016; De Maio et al., 2016; Suchard et al., 2018; Campbell et al., 2019). In this respect, the above-mentioned authors stressed the importance of incorporating epidemiological data (e.g., timing of symptoms, timing of sampling, contact tracing) providing a prior knowledge about the dynamics of the hosts and the pathogen and constraining the inference. Combining different types of data is expected to improve the reconstruction of transmission links, although it is computationally and methodologically challenging (Suchard et al., 2018). Indeed, it requires to build and evaluate a unified likelihood for both epidemiological and genetic data. Therefore, even integrating different types of data may lead to inaccurate inference, if the likelihood is inadequate.

Here we explore how the reconstruction of epidemiological links with SLAFEEL is influenced by the choices of the penalization function and some temporal constraints derived from epidemiological data. In this aim, we apply SLAFEEL to real data on Equine influenza virus (EIV) by adopting different penalization shapes, and by making different temporal assumptions for selecting the putative sources of infection for each host.

In what follows, we first present an analytical way to estimate the evolutionary parameter μ introduced in the pseudo-evolutionary model underlying SLAFEEL and we introduce some flexibilities in the inference of infection sources with a tolerance parameter. Then, we present different shapes for the penalization function and

diverse temporal assumptions used to select the putative sources of an infected host. Afterwards, we provide preliminary results of the application of `SLAFEEL` to equine influenza data by evaluating the impact of the penalization shapes and the temporal constraints on the putative sources.

2 Methodological ingredients

2.1 Estimation of the evolutionary parameter μ

`SLAFEEL` is based on a pseudo-evolutionary model, from which we derive a penalized pseudo-likelihood depending mainly on a penalization function and a kernel smoothing function. This model is associated to a method for the estimation and calibration of parameters and subsequently the inference of transmissions. In the initial version of `SLAFEEL`, Alamil et al. (2019) estimate numerically the evolutionary parameter μ for each recipient-source pair and each value of the penalization parameter, by using a Nelder-Mead algorithm to maximize the penalized pseudo-likelihood. If the use of a numerical algorithm may be necessary in the general case described by Equation (4.1) in Alamil et al. (2019), one can derive a closed form of the estimate of μ in the particular case described by Equation (4.2), where the kernel smoothing applies only to pairs of sequences from the source and the recipient with minimal genetic distance. Below, we present this formal estimation of μ.

Consider M sets of sequences $\mathbf{S}_1, ..., \mathbf{S}_M$ collected from M host units. For each infected host m (from which the set of sequences $\mathbf{S}_m$ is observed) and any value of the penalization parameter θ, the penalized pseudo-likelihood $f_{\mu,\theta}(\mathbf{S}_m|\mathbf{S}_{m'})$, for $m' \neq m$, is maximized with respect to μ. Let $\hat{\mu}_{m'}(\theta)$ denote the estimate of μ given m' and θ. Using approximately the same notation as Alamil et al. (2019), it can be shown that $\hat{\mu}_{m'}(\theta)$ satisfies:

$$\hat{\mu}_{m'}(\theta) = -\frac{1}{4\Delta_{m'}} \log\left(1 - \frac{4}{3}\hat{p}\right), \tag{1}$$

where $\Delta_{m'}$ is a measure of the evolutionary duration between $\mathbf{S}_m$ and $\mathbf{S}_{m'}$ (which can be fixed to 1 as explained by Alamil et al., 2019),

$$\hat{p} = \text{logit}^{-1}\left(\log\left(\frac{\sum_{j=1}^{J} \tilde{d}_j}{\sum_{j=1}^{J}(L - \tilde{d}_j)}\right)\right),$$

$\tilde{d}_j$ is the minimum of the genetic distances (measured by the number of different nucleotides) between the j-th sequence in $\mathbf{S}_m$ and every sequences in $\mathbf{S}_{m'}$, J is the number of sequences in $\mathbf{S}_m$ and L is the size of the observed genome fragments. Note that, to obtain Equation (1), we assume that the penalization does not depend on μ.

2.2 Identification of the source with a tolerance threshold

In the initial version of `SLAFEEL`, for any value of m and θ, once μ has been estimated for every putative sources m', the most likely source(s) of m is (are) the one(s) maximizing the penalized pseudo-likelihood:

$$\hat{s}(m,\theta) = \underset{m' \neq m}{\operatorname{argmax}} f_{\hat{\mu}_{m'(\theta)},\theta}(\mathbf{S}_m|\mathbf{S}_{m'}), \tag{2}$$

where $\hat{s}(m,\theta)$ is the label of a unique host if only one m' maximizes the penalized pseudo-likelihood, and is a set of labels if several putative sources maximize it. The latter case can arise quite frequently as soon as the observed inter-host pathogen diversity is not very large. In addition, we may observe several putative sources for which $f_{\hat{\mu}_{m'(\theta)},\theta}(\mathbf{S}_m|\mathbf{S}_{m'})$ does not reach the maximum value $f^{\max}(m,\theta) = \max_{m' \neq m} f_{\hat{\mu}_{m'(\theta)},\theta}(\mathbf{S}_m|\mathbf{S}_{m'})$ but is very close from this value.

To avoid to exclude a putative source that could be the real source just because of a small difference between $f_{\hat{\mu}_{m'(\theta)},\theta}(\mathbf{S}_m|\mathbf{S}_{m'})$ and $f^{\max}(m,\theta)$, we relax the maximization problem (2) by introducing a tolerance value $\eta \geq 0$. Hence, the most likely source(s) for host unit m given θ, is (are) the host unit(s) m' leading to a value of $f_{\hat{\mu}_{m'(\theta)},\theta}(\mathbf{S}_m|\mathbf{S}_{m'})$ greater than the difference between the maximum pseudo-likelihood reached among all the putative sources $f^{\max}(m,\theta)$ and η:

$$\hat{s}(m,\theta) = \{m' \in \{1,\ldots,M\} : m' \neq m, f^{\max}(m;\theta) - f_{\hat{\mu}_{m'(\theta)},\theta}(\mathbf{S}_m|\mathbf{S}_{m'}) \leq \eta\}.$$

Note that the putative sources m' may not be searched in the entire set $\{1,\ldots,M\} - \{m\}$ as described in the following section.

2.3 Temporal constraints

For each infected host m, one defines a subset of $\{1,\ldots,M\} - \{m\}$ consisting of the putative sources. This subset can depend on temporal constraints grounded on observed timing data. Here, we propose three different temporal constraints established according to the available temporal data, namely the time of the observation of the infection of each host and an eventual reconstruction of other timing information. We assume that the putative infecting hosts (i.e., sources) are the set of hosts that are either:

1. observed earlier than or at the same time as the recipient host (this option was the one considered by Alamil et al., 2019);

2. observed earlier than or observed up to 2 days after the observation of the recipient host;

3. infectious during the 5 days preceding the beginning of the infectious period of the infected host

Temporal constraint 3 is graphically represented in Figure 1. It can be applied by reconstructing the infectious period of each host and by assuming that an host is infected from 5 to 0 days before the infectious period.

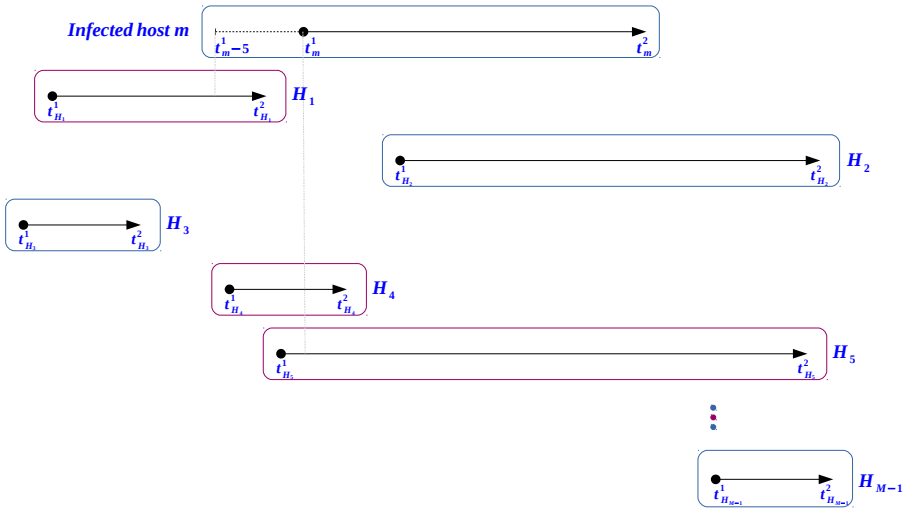


Figure 1: Graphical representation of the third temporal constraint used to select the putative sources of an infected host. Each box represents a host. Within each box, there is an axis representing the infectious period estimated with a preliminary analysis of equine influenza data. t^1 and t^2 are respectively the start and the end of the infectious period. The first box corresponds to the infected host m, the red boxes refer to the selected putative sources among the $M-1$ hosts. The blue boxes correspond to the non-selected hosts. A host H_i is selected as a putative source if $[t^1_{H_i};t^2_{H_i}] \cap [t^1_m - 5;t^1_m] \neq \varnothing$.

2.4 Penalization shapes

In the initial version of `SLAFEEL`, we considered two hypotheses for the penalization and derived different penalization shapes from these hypotheses. Under the first hypothesis, say H1, the contributions of sequences from the source to explain sequences in the recipient are homogeneous (in this study, H1 is implemented using the H1-χ^2 shape defined by Alamil et al., 2019). Under the second hypothesis, say H2, the distances between sequences in the recipient and their contributing sequences in the source are consistent with some known features, typically the expected value and the

standard deviation of these distances computed from training data (in this study, H2 is implemented using the H2-normal shape defined by Alamil et al., 2019). Here, we propose a joint penalization on the genetic distance $\bar{d} = (1/J)\sum_{j=1}^{J} \tilde{d}_j$ and the difference between the observation times of the infected host and the putative source $\Delta_{\text{obs}} = T_{\text{obs}} - T_{\text{obs}}^{(0)}$. This penalization corresponds to the following hypothesis, say H3:

H3: The distances between sequences in the recipient and their contributing sequences in the source are consistent with some known features, and the relationship between these distances and the lag in the observation of the source and the recipient is consistent with some known features.

As for H2, the *known features* mentioned in H3 are learned from training data. Supporting Text S1 makes explicit these known features as well as the shape of the penalization P_θ (called H3-gamma.LM) which depends on the sequence sets and the observation times of the recipient and the source, and which satisfies:

$$P_\theta(\mathbf{S}, \mathbf{S}^{(0)}, T_{\text{obs}}, T_{\text{obs}}^{(0)}) = \left\{\phi\left(\Delta_{\text{obs}} \mid \bar{d}\,; \delta, \nu, \sigma^2\right)\gamma\left(\bar{d}\,; \alpha_1, \alpha_2\right)\right\}^\theta,$$

where the first term $\phi\left(\Delta_{\text{obs}} \mid \bar{d}\,; \delta, \nu, \sigma^2\right)$ corresponds to a Gaussian linear regression between Δ_{obs} and $\bar{d}$ whose parameters are learned from training data, and the second term $\gamma\left(\bar{d}\,; \alpha_1, \alpha_2\right)$ is the probability density function of a gamma distribution, which is expected to be followed by $\bar{d}$ and whose parameters are also learned from training data.

3 Data

The data set used in this study was collected during the outbreak of equine influenza (H3N8) in the United Kingdom between March and May 2003 (Newton et al., 2006). We analysed 48 confirmed H3N8 horses from 22 yards denoted A to W. The virus population within each host was sampled on a daily basis, and multiple clones of hemagglutinin gene were sequenced using fluorescent sequencing chemistry and ABI 3730xl capillary sequencers. On average over the 48 hosts, 9 strains were sampled from each hosts (see Table 1). These samples were identified according to the ID of the yard and a horse number (e.g., sample A01 refers to the horse 1 in the yard A). Supplementary details about the data set are provided in Tables 1 and S1.

In addition, we have at our disposal information about seven recipient-source training pairs identified with a relatively high confidence by (Hughes et al., 2012). These pairs are given by Table 2. We notice that two of the pairs concern the same

Statistics	
Number of host units	48
Number of sequence fragments	1
Fragment length	903
Mean (SD) sequence depth	8.97 (5.69)
Number of different variants	429
Mean (SD) number of different variants per host unit	8.94 (5.57)
Mean (SD) genetic distance* between variants	2.58 (1.13)
Mean (SD) within-host distance between variants	1.96 (0.69)

* The genetic distance between two sequences is the number of different nucleotides.

Table 1: Statistics computed from the Equine influenza virus data.

Recipient host	Source host
L42	L25
H24	H16
N28	E10
N37	J21 or N28
L39	L40
M32	L25

Table 2: Contact information used for the reconstruction of transmission links of the Equine influenza outbreak.

recipient host, namely N37, whose source may be either J21 or N28 with a relatively high confidence.

4 Preliminary results

4.1 Inferring epidemiological links of an Equine influenza outbreak

To illustrate the type of results that we can obtain, we first show those obtained with the following specification: a tolerance $\eta = 1$ is applied in the identification of more likely sources; putative sources are selected under the third temporal constraint; the penalization corresponds to the hypothesis H3. In addition, the penalization parameter θ is calibrated over the set of values $\{0, 0.25, 0.5, ..., 10\}$, and we use contact tracing information provided by Table 2. The specification considered here was one of those reaching the highest consistency between contact information and inferred transmission; see Supporting Table S2. In this table, the measure of the consis-

tency was assessed without cross-validation. In a revised version of this work, we will compute this measure with cross-validation that allows a better assessment of the performance of each specification. We precisely carried out a leave-one-out-cross validation analysis for the specification considered in this section, i.e., for each recipient in the training data set, we remove the corresponding training information from the training data set, we apply `SLAFEEL` with this reduced training data set, and we analyse how the removed training recipient-source pair(s) is (are) reconstructed (we used the plural as a possibility since N37 has two potential sources based on training information). This analysis is summarized in Figure 2. Sources identified with contact tracing are ranked first (with other putative sources) for horses L42, H24 and N37. For horses N28 and L40, the traced sources are detected but not ranked first. Whereas, for horse M32, the traced source is not inferred as a likely source by `SLAFEEL`. We observe in the latter case an extreme behaviour, in the relationship ($\Delta_{\text{obs}} \sim \bar{d}$), of the recipient-source training pair (M32,L25); see Supporting Figure S1 This observation could either indicate that the H3-gamma.LM shape is not adequate or that the pair (M32,L25) is not actually a recipient-source pair.

Figure 3 displays the transmission links of intensity higher than 0.05 between the 48 infected horses sampled from 22 different yards. This figure shows that horses from yards D, E and H seem to be interfaces between horses from yards A, B, C and those from yards L, M, N which in turn transmit the virus for the horses of yards T, U, V and W. This reflects the chronology of yard first infection pointed out by Hughes et al. (2012).

4.2 Impact of the penalization shape and the temporal constraint on `SLAFEEL` output

Above, we compared the different `SLAFEEL` specifications (temporal constraint × penalization shape) in terms of consistency between the traced sources and the inferred sources for recipients in the training data set; see Supporting Table S2, which will be made with cross-validation in a further version of this work. Here, we compare these specifications by comparing `SLAFEEL` output for the whole data set (not only training hosts) with the transmission tree inferred with `BadTrIP` (De Maio et al., 2018). Note that, in the inference made with `BadTrIP`, two recipient-source pairs in the training data set were inferred as the most likely transmissions for the corresponding recipients. Therefore, `BadTrIP`-based transmission tree is considered as a benchmark here, it is not viewed as the truth.

First of all, the most likely sources are the same for less than 2% of recipient hosts whatever the penalization shape and the temporal constraint (Figure 4).

Moreover, the most likely sources inferred by SLAFEEL are among the 20 most likely sources identified by BadTrIP for less than 30% of the recipients. Thus, the consistency between BadTrIP and SLAFEEL is rather low. We do see some differences between SLAFEEL specifications: e.g., if we compare only the most likely sources provided by BadTrIP and SLAFEEL, the highest consistency is reached with the temporal constraint 3 and the H1-χ^2 penalization shape; if we see if the most likely source inferred by SLAFEEL is among the 20 most likely sources provided by BadTrIP, the highest consistency is reached with the temporal constraint 2 and the H2-normal penalization shape.

5 Short discussion

As mentioned above, the diverse specifications that we built should be compared by computing, with cross-validation, the number of training recipient-source pairs that are correctly inferred.

When we use the temporal constraint 3 and the H3 hypothesis for the penalization shape, SLAFEEL ranks first 6 training pairs without cross-validation and 3 training pairs with cross-validation. In contrast, BadTrIP ranks first 2 training pairs. However, for a fair comparison, we should take into account the ties in the ranking. We could also consider a slightly different criterion based on the rank: the average rank (taking into account the ties) of all the traced sources; or a probabilistic criterion: the average probability of the traced transmissions.

The results that we produced were obtained for a fixed tolerance value. Preliminary results (not shown in this chapter) indicate that the tolerance value plays a role in hindering or improving SLAFEEL accuracy. Thus, the relationship between the tolerance value and SLAFEEL performance should be explored in the future. Note that the tolerance value could be selected (with the penalization shape and the temporal constraint) by maximizing the number of training recipient-source pairs that are correctly inferred with cross-validation. This criterion, which is rather impartial, could be used to explore in further work other temporal assumptions allowing to select the putative sources and other penalization shapes correcting the pseudo-likelihood.

Finally, identifying a training recipient-source pair that has an extreme behaviour (like (M32,L25) in this study) should lead to consider again the information that was used for considering this pair as a training pair. It could be also interesting to make a sort of influence analysis, by analyzing how the inclusion or the deletion of this pair from training data change the results for the whole transmission tree.

References

Alamil, M., J. Hughes, K. Berthier, C. Desbiez, G. Thébaud, and S. Soubeyrand (2019). Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. *Philosophical Transactions of the Royal Society B 374*(1775), 20180258.

Campbell, F., A. Cori, N. Ferguson, and T. Jombart (2019). Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS computational biology 15*(3), e1006930.

De Maio, N., C. J. Worby, D. J. Wilson, and N. Stoesser (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS computational biology 14*(4), e1006117.

De Maio, N., C.-H. Wu, and D. J. Wilson (2016). Scotti: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS computational biology 12*(9), e1005130.

Didelot, X., J. Gardy, and C. Colijn (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution 31*(7), 1869–1879.

Hall, M., M. Woolhouse, and A. Rambaut (2016). Using genomics data to reconstruct transmission trees during disease outbreaks. *Revue scientifique et technique (International Office of Epizootics) 35*(1), 287.

Hughes, J., R. C. Allen, M. Baguelin, K. Hampson, G. J. Baillie, D. Elton, J. R. Newton, P. Kellam, J. L. Wood, E. C. Holmes, et al. (2012). Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS Pathog 8*(12), e1003081.

Jombart, T., A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol 10*(1), e1003457.

Newton, J., J. Daly, L. Spencer, and J. Mumford (2006). Description of the outbreak of equine influenza (h3n8) in the united kingdom in 2003, during which recently vaccinated horses in newmarket developed respiratory disease. *Veterinary Record 158*(6), 185–192.

Suchard, M. A., P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut (2018). Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus evolution 4*(1), vey016.

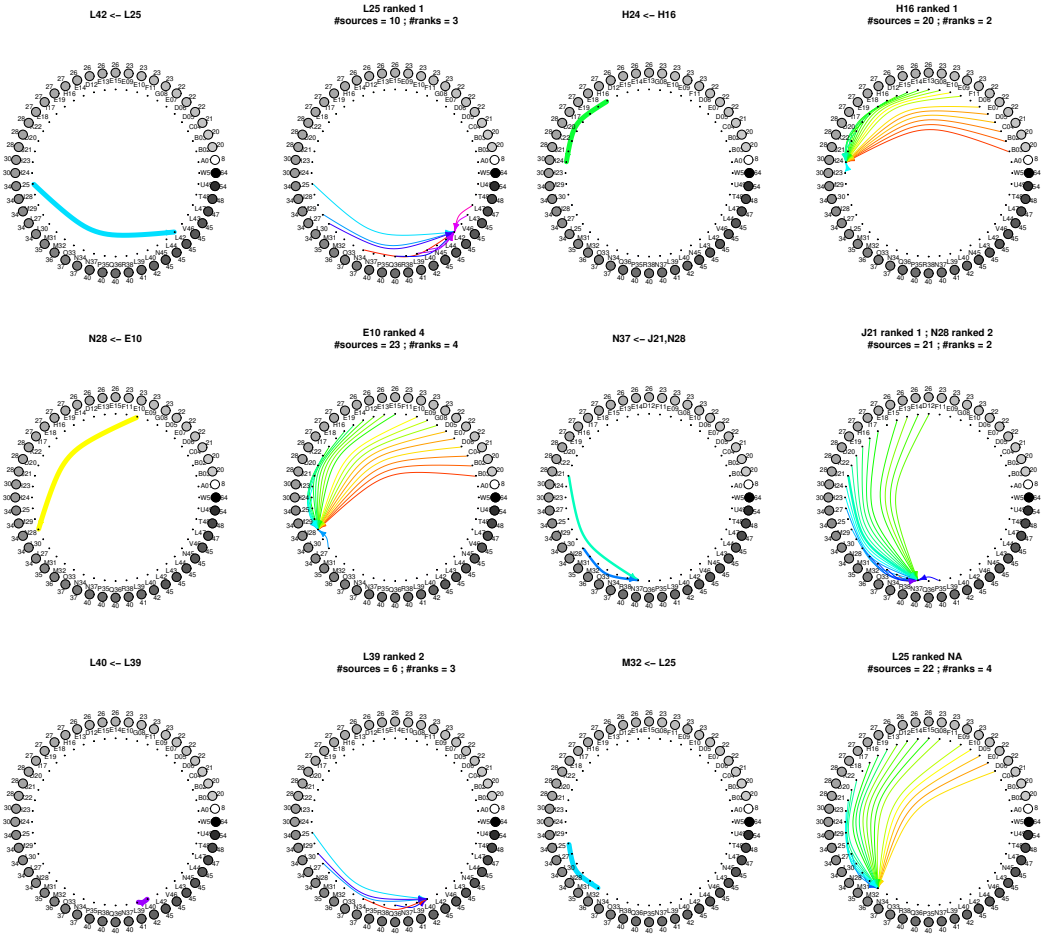


Figure 2: Training information and inferred sources for each recipient in the training data set inferred by cross-validation. In each pair of panels, corresponding to a given recipient host in the training data set, the left panel gives the traced source(s) of the recipient, and the right panel gives the likely sources inferred by `SLAFEEL`.

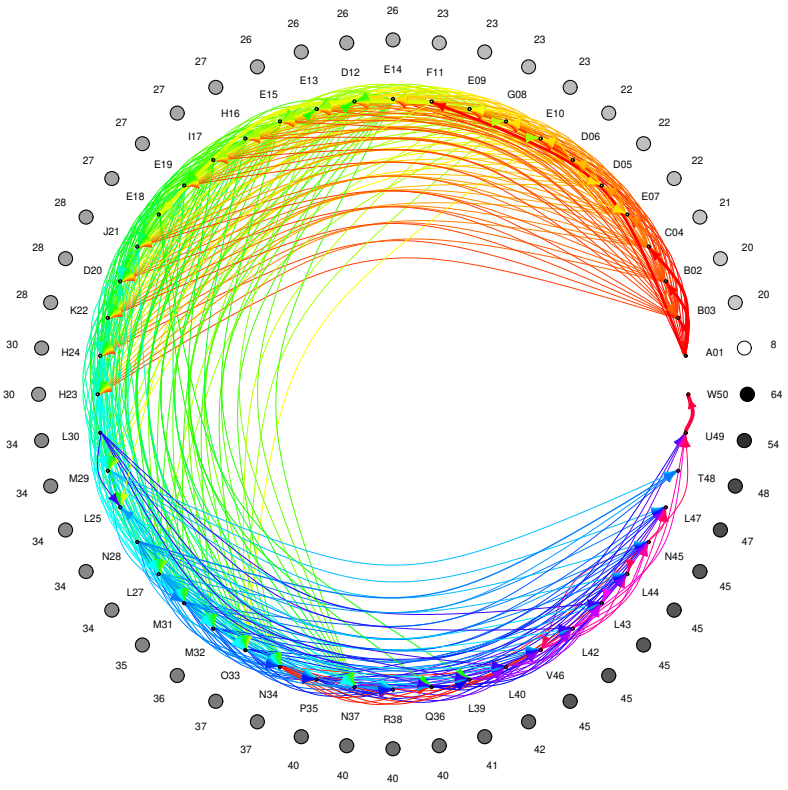


Figure 3: Transmission links of intensity higher than 0.05 inferred between 48 confirmed EIV horses. Circles represent hosts denoted with their IDs and their sampling times.

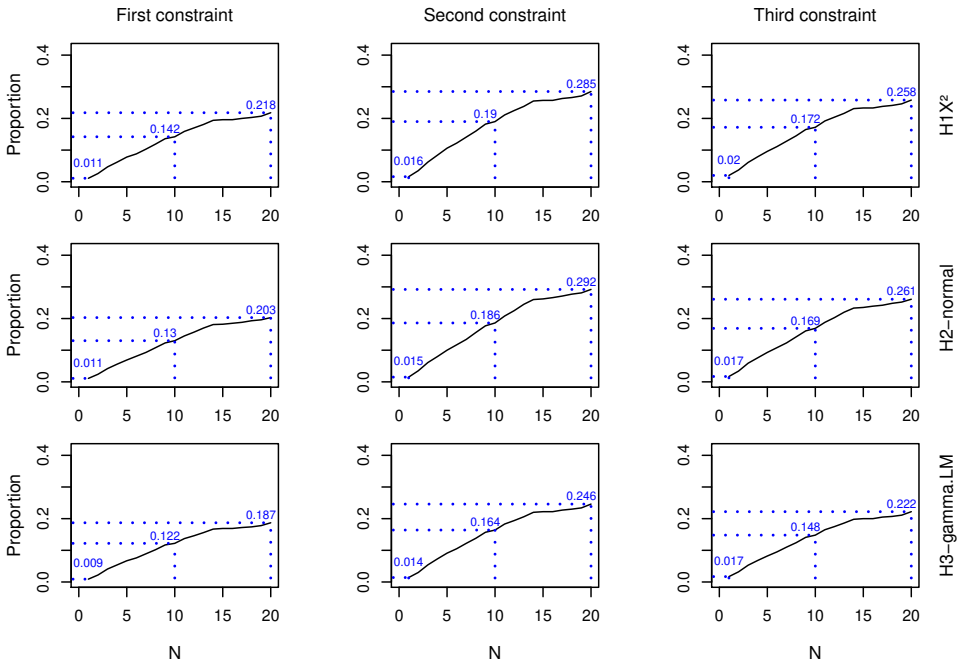


Figure 4: Proportion of recipient hosts whose `SLAFEEL`-based most likely sources are among the N `BadTrIP`-based most likely sources. Each line correspond to a penalization shape and each column to a temporal constraint (the tolerance $\eta = 0.1$ was used for every specifications).

3.4. Key points of Chapter 3

- ☞ Accurate inference of transmission links between hosts is becoming more feasible combining epidemiological data with genetic data.
- ☞ The obvious direction to follow for estimating transmission links of infectious diseases from genomic data reflecting the within-host genetic diversity of pathogen is adopting an approach based on a mechanistic vision of transmission and micro-evolutionary processes.
- ☞ However, alternative approaches, for instance grounded on statistical learning, may be explored to deal with non-standard situations where classical mechanistic assumptions may be erroneous, to handle sequencing errors, to tackle massive data sets and, more generally, to challenge approaches based on a mechanistic vision.
- ☞ In the context of disease-transmission reconstruction, statistical learning may rely on partial contact information used as learning data, guiding the inference of epidemiological links.
- ☞ Such an approach is adaptable to very different contexts and data from animal, human and plant epidemics.

Chapter 4

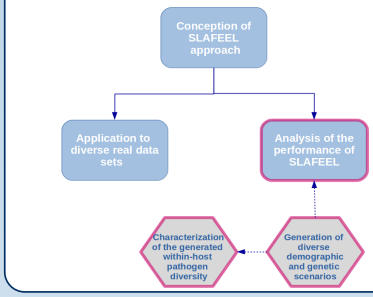
Validation of our approach efficiency

This chapter introduces two articles in progress:

Table of contents

4.1. Graphical summary

Chapter tasks

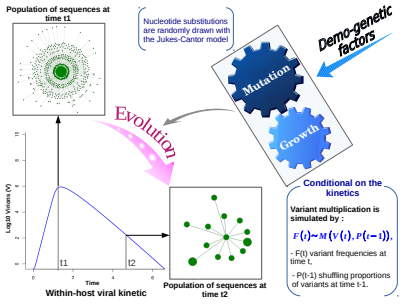


Answer to:

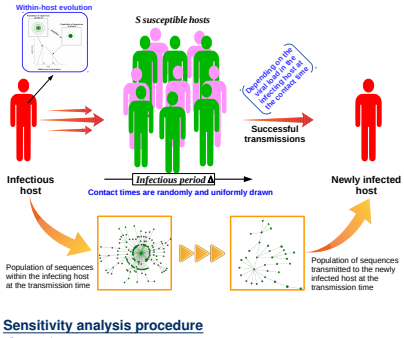
- How robust, effective and versatile is the developed inference approach?
- What sort of within-host pathogen diversity can be generated by the stochastic model proposed for simulating data?

Methodological ingredients

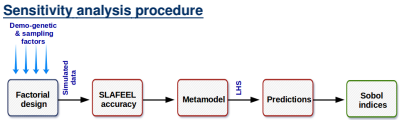
Simulation model for the within-host viral population



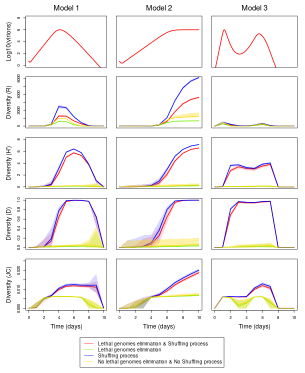
Simulation model for the inter-host viral transmission



Sensitivity analysis procedure

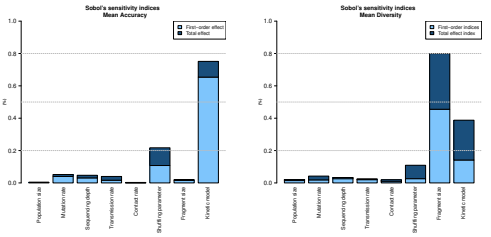


Characterization of within-host genetic diversity

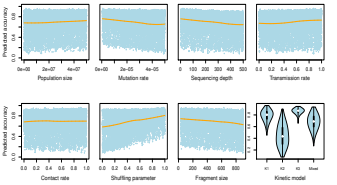


Temporal variations in within-host genetic diversity under various demo-genetic conditions.

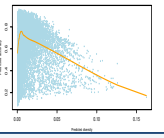
Sensitivity analysis results



Sobol indices of SLAFEEL accuracy and the observed total genetic diversity.



Relationship between input simulation model parameters and the predictions of SLAFEEL mean accuracy.



Predicted mean accuracy versus predicted total observed diversity.

Conclusion & perspectives

- **The characterization of the within-host genetic diversity shows that:** the viral kinetics and the variation in genotype proportions governed by the shuffling parameter are the drivers of the within-host diversity.
- **Sensitivity analysis of the relationship between SLAFEEL accuracy and some demo-genetic and sampling factors shows that:**
  - the factors that have the greatest effect on SLAFEEL accuracy are important factors affecting the genetic diversity;
  - an optimum for the total observed diversity allows the transmission network to be reconstructed with a relatively high level of accuracy.
- **Further works could aim to:**
  - Investigate the relationship between SLAFEEL and:
    - other input model factors,
    - the tuning options in SLAFEEL, such as the penalization shape and the smoother,
    - the training data.

4.2. Article 1

Characterizing viral within-host diversity in fast and non-equilibrium demo-genetic dynamics

Alamil M.[1,2,*], Thébaud G.[3], Berthier K.[4], and Soubeyrand S.[1]

[1]INRAE, BioSP, 84914 Avignon, France
[2]LMA, Université d'Avignon, 84140, AVIGNON
[3]BGPI, Univ Montpellier, INRAE, CIRAD, Institut Agro, Montpellier, France
[4]INRAE, Pathologie Végétale, 84140, Montfavet, France
[*]Corresponding author: maryam.alamil@inrae.fr

Abstract

High-throughput sequencing has opened the route for a deep assessment of within-host genetic diversity that can be used, e.g., to characterize microbial communities and to infer transmission links in infectious disease outbreaks. For assessing the performance of such characterization and inference approaches, which are often grounded on computer-intensive techniques and which cannot be theoretically analyzed, being able to simulate within-host genetic diversity across time under various demo-genetic assumptions is paramount (indeed, one can generate a high number of simulated data sets, apply the approach of interest to each of them, and quantitatively evaluate the accuracy of the results since we know the *truth* underlying the simulated data sets). In this article, we precisely develop a simulation model of viral within-host genetic diversity and characterize the generated diversity under various assumptions. The model that we propose provides the temporal evolution of genotypes and their frequencies under various demo-genetic conditions, and allows the generation of fast non-equilibrium demo-genetic dynamics. The characterization of the within-host genetic diversity is performed numerically with several classical diversity indices. This study enables us to point out key drivers of the within-host viral diversity, namely the viral kinetics and the fast variation in genotype proportions differently influencing genetic selection and drift.

Keywords: diversity indices, genome evolution, kinetic model, simulation model, virus evolution, within-host pathogen diversity

1 Introduction

RNA viruses, such as Influenza A, Ebola and Hepatitis C viruses, are often referred as fast evolving pathogens because of their high mutation rates and rapid generation time (Nelson and Hughes, 2015; Biek et al., 2015; Picard et al., 2017). These characteristics hold at the multi-host level as well as at the within-host level. The development of sequencing technologies has specifically contributed to unravel how virus genetic diversity can be significant within a single host and that it may vary spatially within the host as well as temporarily during the course of the infection due to mutation, selection and genetic drift processes acting at the within-host scale (Pybus and Rambaut, 2009; Alizon et al., 2011; Gutiérrez et al., 2012; Simmons et al., 2012; Abel et al., 2015; Cuevas et al., 2015; Nelson and Hughes, 2015; Poirier and Vignuzzi, 2017). Typically, a deep assessment of within-host genetic diversity can be achieved using whole genome high-throughput sequencing (HTS) approaches on serial samples collected from an important number of infected hosts. However, and although most RNA viruses have relatively small genome sizes, accurate whole genome sequencing of numerous samples still remains costly, time consuming and require expertise in bioinformatics to select appropriate tools and approaches for data analysis (Kulkarni and Frommolt, 2017). Alternatively, within-host genetic diversity can be approached by high-throughput amplicon sequencing (HTAS) techniques, which can be used to identify distinct genotypes for a target marker of a few hundred bases length within the host while genotyping a high number of samples through ad hoc multiplexing techniques (Galan et al., 2010, 2012; Piry et al., 2017). Such techniques are less costly and produce data that can be easily handled and analyzed with limited computational resources and bioinformatics expertise (e.g., using the `R` package `dada2` for example; Callahan et al., 2016).

Today, within-host genetic diversity of virus is of particular interest for inferring (potentially indirect) epidemiological links between hosts and even reconstructing chains of transmission in outbreaks. Before the use of within-host genetic diversity for such inferences, one essentially exploited the high mutation rate and rapid generation time of viruses and analyzed the spatio-temporal structure of the viral genetic diversity at the multi-host level (Brunker et al., 2012; Picard et al., 2017). Typically, empirical and model-based approaches were designed to use information on virus genetic diversity at the multi-host level and hence reconstruct transmission links during outbreaks (Cottam et al., 2008; Morelli et al., 2012; Hall et al., 2015; Jombart et al., 2014; Mollentze et al., 2014; Ypma et al., 2012, 2013; Lau et al., 2015; Valdazo-González et al., 2015). In most of the earliest approaches that have been developed, the host unit was (implicitly) considered as a spatially homogeneous environment, within which the viral population at a fixed time was represented by a unique sequence, such as the consensus or the majority sequence.

However, recent approaches have exploited within-host genetic diversity and the degree of genetic similarity (in a broad sense) between viral genotypes collected from different hosts for transmission chain reconstruction (Alamil et al., 2019; De Maio et al., 2018; Wymant et al., 2018; Leitner and Romero-Severson, 2018; Jombart et al., 2014; Worby et al., 2014; Didelot et al., 2014; Walker et al., 2013; Morelli et al., 2012; Murcia et al., 2012; Hughes et al., 2012). To evaluate the performance of these approaches in numerous diverse and challenging settings, we need a simulation model of viral within-host genetic diversity and some tools to characterize this diversity. We precisely propose such a framework here, based on the work of Worby and Read (2015) on the simulation of evolutionary and epidemiological dynamics, as well as classical viral kinetic model and widely used diversity indices.

In our approach, within-host virus population is simulated by generating genotypes (i.e., sequence fragments) and their proportions under different demographic kinetics. The resulting computer-based demo-genetic dynamics can be generated under numerous conditions and can be monitored like in real situations using HTAS longitudinal samples (i.e., samples collected from a unique host at different times during the infection). In the model, demographic effects are essentially represented by a founder effect (i.e., the set of genotypes initiating the infection), which can be relatively strong (Abel et al., 2015; Poirier and Vignuzzi, 2017), and a demographic kinetic described by a set of differential equations and quantifying the variation of the viral load during the course of the infection. We consider three examples of kinetic models: 1) a latency model representing an acute infection, 2) a latency model representing a chronic infection and 3) a latency model with an immune response. Genetic effects correspond to the mutation and replication processes. Nucleotide substitutions are assumed to occur randomly at a constant rate. Mutation effects are handled by classifying substitutions into lethal (leading to negative selection) and non-lethal.

Genotype replication is simulated by successive over-dispersed multinomial draws with a size equal to the current quantity of virions that is governed by the chosen kinetic model. The replication success represents the fitness of the genotypes, which can vary during the course of the infection via the over-dispersion of the multinomial draws. The over-dispersion is governed by a so-called shuffling process noising the current vector of genotype proportions. When this process is applied, a rare genotype at generation t can significantly increase in proportion at generation $t+1$. This process implicitly mimics positive selection, genetic drift and spatio-temporal variation in genotype multiplication (occurring, e.g., when a genotype invades a new part of the host that is more favorable to it). Thus, overall, the stochastic model that we propose implicitly or explicitly encompasses several biological mechanisms such as natural selection and genetic drift and produces fast and non-equilibrium

demo-genetic dynamics.

The model briefly described above was designed for the evaluation, in diverse and challenging demo-genetics situations, of the performance of methods that reconstruct transmission trees by exploiting within-host genetic diversity data. However, we focus in this article on the characterisation of the genetic diversity resulting from this simulation model. Thus, in what follows, we propose a comprehensive mathematical description of the model and we investigate the influential parameters in terms of temporal variation in genetic diversity. This investigation is performed using several diversity indices, and contributes to a better understanding of the main drivers of within-host genetic evolution and pathogen population divergence. These elements are discussed in the last section of this article.

2 Theory and calculation

2.1 Kinetic models

In our study, we consider that within-host pathogen population size varies over time. To quantify this temporal variation, we use kinetic models that were developed to study non-equilibrium within-host dynamics of pathogens (Baccam et al., 2006; Smith and Perelson, 2011; Beauchemin and Handel, 2011; Nowak and May, 2000; Beauchemin et al., 2008; Saenz et al., 2010; Handel et al., 2010; Pawelek et al., 2012; Canini and Perelson, 2014). These models are grounded on sets of ordinary differential equations (ODE) basically governing the numbers of susceptible target cells, infected cells and virions. We chose three of these models (presented below) corresponding to different situations: a model incorporating a latency in virus production (Baccam et al., 2006; Beauchemin and Handel, 2011), a latency model corresponding to a chronic infection, and a latency model incorporating an immune response (Smith and Perelson, 2011; Saenz et al., 2010; Pawelek et al., 2012).

2.1.1 Acute and chronic infection models

The acute infection model is derived from a simple viral kinetic model describing the dynamics between susceptible target cells (S), infected cells (I) and virions (V) (Baccam et al., 2006; Beauchemin and Handel, 2011). It illustrates the eclipse phase dynamics. The eclipse phase is the time span between the entry of the virus into the target cells and the release of the virions produced by these newly infected cells. The delay in the viral production is modeled by defining two separate populations of infected cells: the infected population that is not yet producing virions, I_1, and the infectious population that is actively producing virions, I_2. The following set of

differential equations (Baccam et al., 2006; Beauchemin and Handel, 2011) defines the latent model:

$$\begin{cases} \dfrac{dS}{dt} = -\beta SV \\ \dfrac{dI_1}{dt} = \beta SV - kI_1 \\ \dfrac{dI_2}{dt} = kI_1 - \delta I_2 \\ \dfrac{dV}{dt} = pI_2 - cV, \end{cases} \tag{1}$$

where the susceptible cells, S, are converted at rate β into infected cells, I_1 upon interaction with virions, V. Infected cells become infectious at rate k; in other words, $1/k$ is the average transition time from I_1 to I_2. The virions, V, are assumed to be produced at rate p and cleared at rate c.

To model a chronic infection, we use the acute model of Eq. (1) and we assume that the infectious cells I_2 directly responsible for the production of virions are not removed or lost. In that respect, the death rate of infectious cells I_2 is set to zero and the chronic model is defined by:

$$\begin{cases} \dfrac{dS}{dt} = -\beta SV \\ \dfrac{dI_1}{dt} = \beta SV - kI_1 \\ \dfrac{dI_2}{dt} = kI_1 \\ \dfrac{dV}{dt} = pI_2 - cV. \end{cases} \tag{2}$$

A schematic diagram of these acute and chronic models is shown in Figure 1.

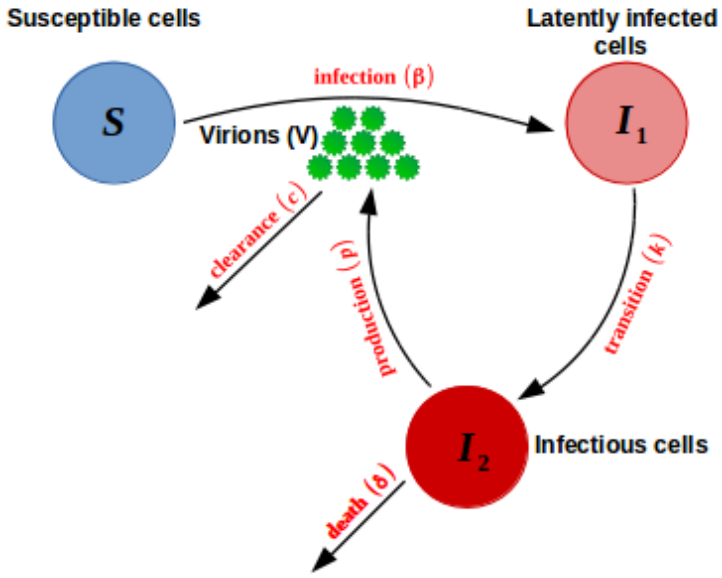


Figure 1: Schematic representation of the acute and chronic models defined respectively by Eq. (1) and (2). The chronic model is obtained by setting the death rate of infectious cells δ to zero.

2.1.2 Immunity-cured infection model

A third model accounts for the immune response.Innate immunity through interferon (IFN) induction is modelled by adding two compartments to the acute-infection model defined by Eq.(1): the IFNs (F) and the refractory uninfected cells (R). The rising adaptive immune response is modelled as an increase in the death rate of the infectious cells, δ, after an initial delay. Thus the model with an immune response

is defined by:

$$\begin{cases} \dfrac{dS}{dt} = -\beta SV - \phi SF + \rho R \\ \dfrac{dI_1}{dt} = \beta SV - kI_1 - mI_1F \\ \dfrac{dI_2}{dt} = kI_1 - \delta I_2 - mI_2F \\ \dfrac{dR}{dt} = \phi SF - \rho R \\ \dfrac{dV}{dt} = pI_2 - cV \\ \dfrac{dF}{dt} = qI_2 - dF, \end{cases} \tag{3}$$

where: IFNs are secreted only by infectious cells I_2 at rate q and decay at rate d; upon exposure to these signalling proteins, all infected cells incur an (additional) death rate m, and susceptible cells become refractory to infection at rate ϕ (refractory cells revert to the susceptible state at rate ρ); δ is defined as follows:

$$\delta = \begin{cases} \delta_I & \text{if } t < s \\ \delta_I e^{\sigma(t-s)} & \text{otherwise,} \end{cases}$$

where $1/\delta_I$ is the mean lifespan of the infectious cells before the rise of the immune response, and σ determines the speed at which the death rate increases after the time s when the adaptive immune response starts (Pawelek et al., 2012).

A schematic diagram of the immunity-cured infection model (Eq. (3)) is given in Figure 2.

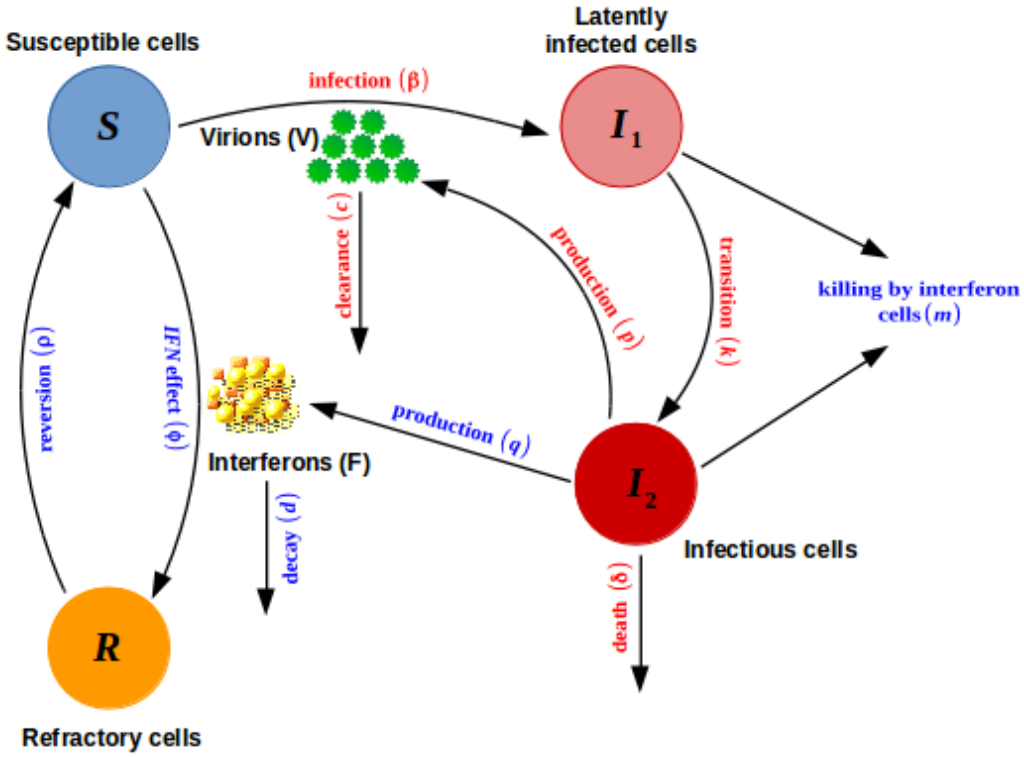


Figure 2: Schematic representation of the immunity-cured infection model defined in Eq. (3). This schematic representation is an edited version (with permission) of Pawelek representation (Pawelek et al., 2012).

2.1.3 Simulation settings

Values of parameters and initial values of variables used thereafter for simulating changes in the viral load during 10 days are provided in Tables 1 and 2 for the three kinetic models. All these parameters except the viral production rate are taken from previous studies simulating the within-host viral kinetic (Baccam et al., 2006; Pawelek et al., 2012). In these studies, parameters are estimated with a least square approach between the kinetic model and experimental data collected from patients infected by H1N1 (Baccam et al., 2006) or from unvaccinated ponies infected by EIV (Pawelek et al., 2012).

The viral production rate, p, is chosen such that the maximum viral load reached during the infection period is the same for the three different models. Let $V_{\max}$ denote the maximum viral load to be reached (we use $V_{\max} = 10^6$ virions). For each model, parameter p is computed by minimizing (with respect to p) the squared

deviation, $\Delta_p = (V_{\max} - \bar{V}_p)^2$, between $V_{\max}$ and the maximum value $\bar{V}_p$ (over a 10-day time period) of the number of virions V obtained by solving the system of ODEs.

2.2 Demo-genetic model with fast variation

To generate within-host genetic diversity of a pathogen population with a non-equilibrium fast evolutionary dynamics, we build a discrete-time stochastic model simulating genotypes and their frequencies at each generation during an infection period. Numerous data sets can be generated with this model under various demo-genetic situations that can lead to fast-evolving dynamics and consequently to significant changes in the viral composition.

To generate the demo-genetic situations, we integrate several varying demo-genetic factors, namely the kinetic model, the mutations and two fitness components described in the following sections. The sum of genotype frequencies at each generation (i.e., the pathogen population size) is assumed to be the quantity of virions, V, given by one of the three viral kinetic models presented in Section 2.1 (we only need values of V at the discrete times corresponding to the generations; thereafter, the generation and the day coincide). This conditional construction of the population genetics given the demography allows us to consider very diverse demo-genetic scenarios.

Host infection is initiated by the introduction of a single genotype defined by a nucleotide sequence of length L, each nucleotide being uniformly drawn among {A,C,G,T}. At any time t (i.e. generation) during the infection period, the within-host pathogen population is represented by a set of $n(t)$ different genotypes $G(t) = \{g_1(t), ..., g_{n(t)}(t)\}$ and their absolute frequencies $F(t) = \{f_1(t), ..., f_{n(t)}(t)\}$. Below, to complement the definition of the stochastic demo-genetic model, we describe how $\{G(t), F(t)\}$ are generated by a sequential procedure, conditionally on $\{G(t-1), F(t-1)\}$ and $V(t)$.

2.2.1 Growth

First, genotypes undergo a growth stage constrained by the fact that the total quantity of genomes goes from $V(t-1) = \sum_{i=1}^{n(t-1)} f_i(t-1)$ to $V(t)$. This stage is performed with a conditional multinomial draw with size $V(t)$ and probabilities $P^*(t-1)$ equal to standardized noisy versions of the proportions $P(t-1) = \frac{1}{V(t-1)} F(t-1)$ of the genotypes in the set $G(t-1)$ (Section 2.2.3 specifies P^*):

$$F'(t) \mid P^*(t-1), V(t) \sim \text{Multinomial}\left(V(t), P^*(t-1)\right), \tag{4}$$

where $F'(t) = (f'_1(t), ..., f'_{n(t-1)}(t))$ is the frequency vector of the $n(t-1)$ genotypes constituting the $G(t-1)$ family after the growth stage.

After the growth stage and before the mutation stage, all genotypes with zero-frequencies are removed. Hence, we introduce:

$$\begin{aligned} G^*(t) &= \{g_i(t-1) : i = 1, \ldots, n(t-1), f'_i(t) > 0\} \subset G(t-1) \\ &= \{g^*_1(t), ..., g^*_{m(t)}(t)\}, \end{aligned}$$

the set of nonzero frequency genotypes ($m(t) \leq n(t-1)$ is the number of these genotypes), and $F^*(t) = (f^*_1(t), ..., f^*_{m(t)}(t))$ the vector of the corresponding frequencies ($F^*(t)$ is obtained by removing the null elements of the vector $F'(t)$).

2.2.2 Mutations

Second, genomes undergo a mutation stage (followed by the elimination of the lethal genomes; see Section 2.2.4).

At this stage, the number of the mutations $N_v(t)$ occuring in the genome $v \in \{1, \ldots, V(t)\}$ whose genotype is $\gamma_v = (\gamma_v(1), \ldots, \gamma_v(L)) \in G^*(t)$ follows a binomial distribution with size L (which is the genome length) and probability μ (per nucleotide per generation):

$$N_v(t) \underset{\text{indep.}}{\sim} \text{Binomial}(L, \mu), \quad \forall v \in \{1, \ldots, V(t)\}.$$

Let $\mathcal{V}(t) = \{v = 1, \ldots, V(t) : N_v(t) > 0\}$ denote the set of genomes undergoing at least one mutation. For each $v \in \mathcal{V}(t)$, $N_v(t)$ indexes, noted $j_1, \ldots, j_{N_v(t)}$, are selected uniformly with replacement from $\{1, \ldots, L\}$ (using a drawing with replacement allows us to take into account multiple mutations on the same nucleotide, the consequence of this choice leading to a lower efficient mutation rate than μ; note however that given the parameter values that we use in the application section, this event is extremely rare). Then, for j from j_1 to $j_{N_v(t)}$, the nucleotide $\gamma_v(j)$ is updated by drawing randomly and uniformly a new nucleotide from the set {A,C,G,T} and excluding the current value of $\gamma_v(j)$.

Let $\tilde{\gamma}_v$ denote the genotype obtained using this iterative procedure. Eliminating the lethal genomes (see Section 2.2.4), $\tilde{\mathcal{V}}(t)$ designates the set of remaining genomes in $\mathcal{V}(t)$. Assigning (in an arbitrary order) the indices $m(t)+1, \ldots, m(t)+q(t)$ to these $q(t)$ genotypes (where $q(t)$ is the length of $\tilde{\mathcal{V}}(t)$), noting $\{\tilde{g}_{m(t)+1}(t), \ldots, \tilde{g}_{m(t)+q(t)}(t)\} = \{\tilde{\gamma}_v : v \in \tilde{\mathcal{V}}(t)\}$ and $\tilde{g}_i(t) = g^*_i(t)$ for each $i \in \{1, \ldots, m(t)\}$, the genotype set is henceforth:

$$\begin{aligned} \tilde{G}(t) &= G^*(t) \cup \{\tilde{\gamma}_v : v \in \mathcal{V}(t)\} \\ &= \{\tilde{g}_1(t), ..., \tilde{g}_{m(t)+q(t)}(t)\}. \end{aligned}$$

In that respect, the set of frequencies corresponding to the genotypes in the new set $\tilde{G}(t)$ is defined by:

$$\begin{aligned}\tilde{F}(t) &= \tilde{F^*}(t) \cup \{\tilde{f}_{\gamma_v} : v \in \mathcal{V}(t)\} \\ &= \{\tilde{f}_1(t), ..., \tilde{f}_{m(t)+q(t)}(t)\}.\end{aligned}$$

where $\tilde{F^*}$ is the set of frequencies F^* uptaded by deducing the frequency of genomes that were mutated and $\{\tilde{f}_{m(t)+1}, \ldots, \tilde{f}_{m(t)+q(t)}(t)\}$ is the vector of the $q(t)$ genotype frequencies; $\forall k = m(t)+1, ..., m(t)+q(t),\ f_k = 1$.

Then, genotypes whose frequencies are zero in $\tilde{G}(t)$ are deleted, identical genotypes are aggregated and their frequencies are summed. Thus, we obtain the set $G(t)$ of genotypes present in the host at time t, after the growth and mutation stages, and $F(t)$ the frequency vector of these genotypes.

2.2.3 Shuffling process

Here, we describe how we build probabilities $P^*(t-1)$ equal to standardized noisy versions of the proportions $P(t-1)$ and how it is used in the growth stage of the demo-genetic model. Beyond the effect of mutation, genotype frequencies may vary due to other mechanisms such as natural selection and random genetic drift (Lande, 1976). To implicitly account for the effect of such mechanisms into our within-host pathogen evolutionary model, we incorporate a shuffling process into the model. This process consists of drawing genotype proportions with an over-dispersion to simulate the extra multiplication of low-proportion genotypes and/or the reduced multiplication of high-proportion genotypes.

Let P denote a vector of proportions that sum to one (typically, $P(t-1)$ in Section 2.2.1). The vector of proportions P^* provided by the shuffling process applied to P is obtained by noising P with a centered Gaussian distribution:

$$\tilde{P} \mid P \sim \mathcal{N}\left(P, \sigma^2\right), \tag{5}$$

where $\sigma^2 = \gamma_1 \times P^{\gamma_2} \times (1-P)^{\gamma_3}$ ($\gamma_1, \gamma_2, \gamma_3 \geq 0$); cutting $\tilde{P}$ off: $\hat{P} = \min(1, \max(0, \tilde{P}))$; and standardizing $\hat{P}$:

$$P^* = \frac{1}{\sum_{i=1}^{n} \hat{p}_i} \hat{P}, \tag{6}$$

where $\hat{P} = (\hat{p}_1, \ldots, \hat{p}_n)$, $n \in \mathbb{N}^*$. The effects of the shuffling parameters $(\gamma_1, \gamma_2, \gamma_3)$ are detailed in Supporting Text S1. Briefly, the larger γ_1, the larger the noise; the smaller γ_2, the more some low-proportion genotypes may reach high frequencies; the smaller γ_3, the more some high-proportion genotypes may reach low frequencies.

2.2.4 Elimination of lethal genomes

A large proportion of mutations incurred by viral genomes are lethal (Fudala and Korona, 2009; Sanjuán et al., 2004). We accounted for a proportion $\alpha = 0.4$ of lethal mutations by discarding the genomes with mutations in the first 40% of the nucleotide positions along the sequence; the other mutations are considered neutral. To allow the assessment of the presence or absence of lethal-genome elimination given the viral kinetics, the proportion and the frequency of each genotype are then re-scaled such that the sum of proportions is one and the sum of frequencies equals $V(t)$.

2.3 Genetic diversity indices

To measure the level of genetic diversity of the pathogen population within an infected host at each generation t, we used several diversity indices. The first three indices are haplotype diversity indices that depend on genotype abundance (Morris et al., 2014). The fourth index quantifies pairwise genetic distances that depend on sequence variation.

2.3.1 Richness (R)

The richness estimator $R(t)$ is the simple count of different genotypes existing at time t. It is equal to $n(t)$. This index is therefore highly sensitive to rare genotypes.

2.3.2 Shannon index (H')

The Shannon diversity index is calculated as follows:

$$H'(t) = -\sum_{i=1}^{R(t)} p_i(t) \log(p_i(t)), \tag{7}$$

where $R(t)$ is the number of existing genotypes (richness) at time t and $p_i(t)$ is the proportion of the i^{th} genotype at time t. This index is both sensitive to rare and abundant genotypes.

2.3.3 Gini-Simpson index (D)

The Gini-Simpson index also depends on the genotype proportions and is defined as follows:

$$D(t) = 1 - \sum_{i=1}^{R(t)} p_i^2(t), \tag{8}$$

This index is sensitive to abundant genotypes.

2.3.4 Jukes-Cantor distance

Pairwise indices require the comparison between each pair of sequences. Here, we used the Jukes-Cantor distance (Jukes et al., 1969) to evaluate the within-host genetic diversity. Supposing that the rate of nucleotide substitution is the same between any pair of nucleotides, the Jukes-Cantor distance is defined in the following way:

$$\bar{d}(t) = \mathbb{E}_{ij}[d(g_i(t), g_j(t))], \tag{9}$$

where i and j represent two genotypes drawn randomly, independently and uniformly from the genotype space and $d(g_i(t), g_j(t))$ is given by:

$$d(g_i(t), g_j(t)) = -\frac{3}{4}\log(1 - \frac{4}{3}p(g_i(t), g_j(t))),$$

with $p(g_i(t), g_j(t))$ the mean pairwise distance (p-distance) between the two sequences $g_i(t)$ and $g_j(t)$. This p-distance is the proportion of nucleotide sites at which $g_i(t)$ and $g_j(t)$ differ, and it is estimated by $\hat{p}(t) = n_d/L$ (n_d being the number of nucleotide differences).

2.4 Methods

In order to study the impact of the demo-genetic factors on the within-host genetic diversity, we measured the genetic diversity of pathogen populations by the above-mentioned indices during 10 generations. Each pathogen population is characterized by a set of viral genotypes generated via our evolutionary model where the length of each genetic sequence was set to $L = 330$ nucleotides and the mutation rate was set to $\mu = 10^{-5}$ mutation per nucleotide per generation. These populations differ in the demo-genetic characteristics that are included through the kinetic model, the shuffling process and the elimination of lethal genomes. For each demo-genetic scenario, we performed 100 independent simulations of the temporal dynamics of the within-host population.

3 Results

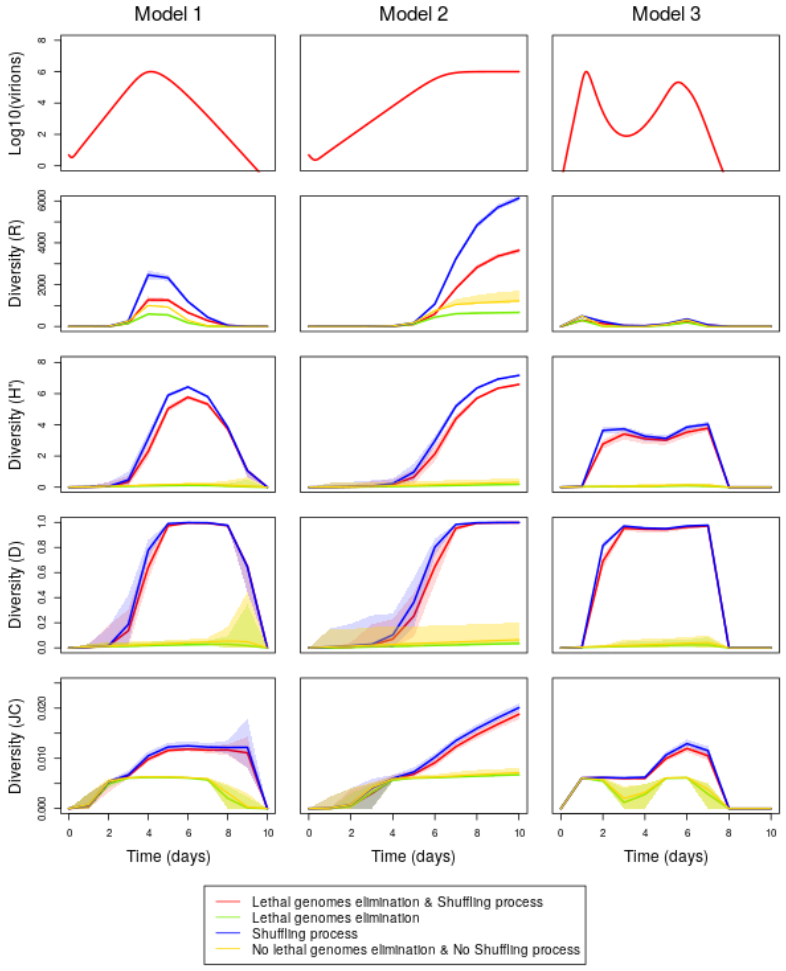


Figure 3: Temporal variations in within-host genetic diversity under various demo-genetic conditions. Row 1: within-host virion quantity under the three models of viral load dynamics (in columns); each day corresponds to one generation. Rows 2 to 5: within-host genetic diversity measured by richness (R), Shannon (H'), Gini-Simpson (D) and Jukes-Cantor (JC) indices, respectively. In each diversity panel, the colors of the lines correspond to different demo-genetic conditions with or without lethal-genome elimination and the shuffling process. Shaded areas delimit the 95% confidence envelopes of the diversity.

3.1 Cross-effects of the viral kinetic, the shuffling process and the elimination of lethal genomes

Figure 3 shows, for three different viral kinetics, the temporal evolution of the genetic diversity of the viral population within a host during an infection, computed from 100 replicates for each kinetic. The diversity is assessed with the four indices described in Section 2.3: richness (R), Shannon (H'), Gini-Simpson (D) and Jukes-Cantor (JC). The kinetic models, which quantify the temporal variation of the viral load during the infection, are those presented in Section 2.1: the acute model, the chronic infection model and the immunity-cured infection model. The simulations are performed with default parameter values, namely the kinetic parameters given in Tables 1 and 2, $\alpha = 0.4$ when lethal genomes are eliminated and $(\gamma_1, \gamma_2, \gamma_3) = (0.8, 0.4, 70)$ when the shuffling process is applied.

Richness (R), Gini-Simpson (D) and Jukes-Cantor (JC) diversity indices are more or less smoothed and delayed versions of the temporal dynamics of virions. We however note that the number of different genotypes is strongly reduced by a fast onset of the immune response (index R, Model 3). In contrast with the three above-mentioned indices, there is a clear difference between the temporal patterns of Shannon index (H') and virion abundance. In other words, the two Shannon peaks in the acute infection model do not represent the same processes. The first peak is linked to the appearance of diversity due to mutations and the increase in the size of the pathogen population. The second peak is due to a strong genetic drift with the decrease in the size of the pathogen population: the number and the abundance of different genotypes decrease and the rare alleles remain (shuffling effect). In the end, there is a high probability that two alleles taken at random are different. In the chronic infection model where there is no bottleneck, we only see the first peak related to the appearance of diversity. Shannon index is sensitive to the presence of rare alleles, which occurs early in the dynamics when new genotypes appear. This index collapses rapidly when maximum population size is reached, probably due to the presence of ultra-dominant genotypes (the Gini-Simpson index is more sensitive to the presence of dominant alleles). In the immunity-cured infection model, the immune response provides rapid fluctuations in the population size and constrains the number of genotypes (little diversity), which seems to generate a rapid succession of small peaks that form a block.

Figure 3 shows that promoting non-equilibrium and fast variations with the shuffling process induces a marked increase in the within-host genetic diversity, whatever the index, even with lethal genomes (red and blue lines). In addition, the comparison with Fig. S1 shows that the shuffling process also results in major qualitative changes in the within-host diversity measured by the H' and S indexes, and to a lesser extent by the JC index. There are two explanations to these observations.

First, the shuffling process favors the number of genotypes (i.e. the richness R) despite the mass at zero of the noisy proportions (see Section 2.2.3). Second, the shuffling process favors the presence of a larger number of abundant genotypes, as particularly illustrated with Shannon (H') and Gini-Simpson (D) indices that are sensitive to abundant genotypes.

Figure 3 also shows, as intuitively expected, that negative selection against lethal mutations (red and green lines) reduces the richness (R) by 60% both in the presence and in the absence of the shuffling process (i.e., when viral multiplication probabilities are noised). In contrast, lethal genome elimination seems to have little impact on Shannon (H'), Gini-Simpson (D) and Jukes-Cantor (JC) diversity indices. Supporting Figures S1 and S2, which show the temporal changes in the four diversity indices when the proportion of lethal mutations α varies between 0.2 and 0.4 (Sanjuán, 2010), essentially confirms this observation.

3.2 Fast changes in genotype proportions

In the shuffling process, the enhancement of low-proportion genotypes is governed in particular by parameter γ_2: the lower γ_2, the larger the dispersion of the noise affecting genotype proportions in the multiplication stage and, consequently, the faster low-proportion genotypes reach large proportions. Figure 4 and Supporting Figure S3 illustrating respectively simulations without and with lethal genome elimination show that variation in γ_2 generates significantly different temporal profiles for all the diversity indices. The overdispersion obtained with small γ_2 increases the number of genotypes (R), the probability of substitutions (JC) and the evenness in genotypes abundance (H'). In addition, small γ_2 values rapidly lead to a maximum Gini-Simpson diversity (D). The two other shuffling parameters, γ_1 and γ_3, have much less influence (apart for $\gamma_1 = 0$) on the diversity indexes (Supporting Figures S4 and S5).

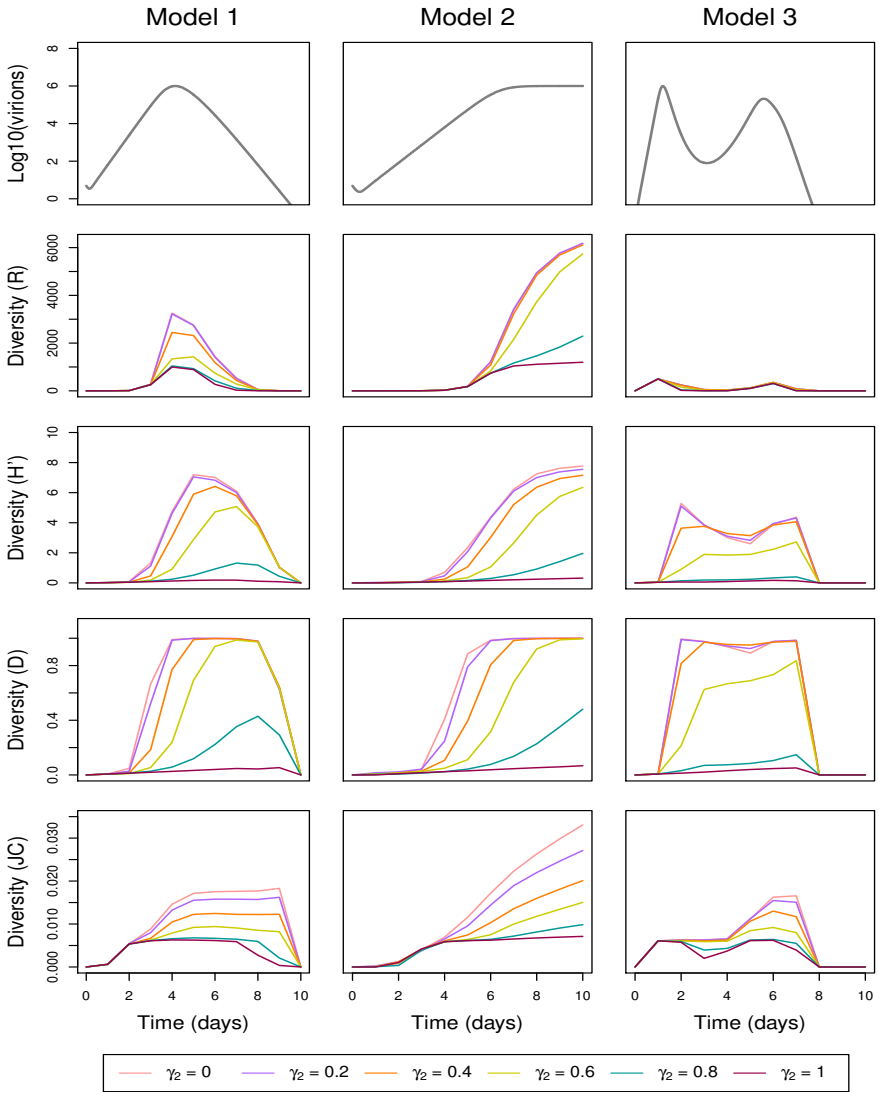


Figure 4: Simulated effects of the shuffling parameter γ_2 on the levels of within-host genetic diversity. Simulations performed without the elimination of lethal genomes (i.e. $\alpha = 0$). Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2–5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the shuffling parameter γ_2.

3.3 Changes in the number of mutations

The proportion of mutated genomes increases with the mutation rate μ and the genome size L. Thus, higher μ and L lead to a faster increase in the proportion of genotypes. Applying the shuffling process and eliminating the lethal genomes or not, Figures 5–6 and Supporting Figures S6–S7 show that the diversity indices are affected in a qualitatively similar manner by an increase in μ or L, which also resembles to the effect of decreasing γ_2 in the shuffling process. Unsurprisingly, the proportion of different nucleotide sites (JC) tends to increase with the mutation rate μ. Moreover, JC tends to decrease when the genome size L increases. For low values of L (typically less than 100), JC is generally over-estimated (Tajima, 1993) and, hence, the curve for $L = 30$ should be cautiously analysed. The decrease of JC with increasing L (for values of L larger than 100) indicates that the typical approximation of the p-distance arising in the formula of JC, namely p-distance $= 2\mu L/L = 2\mu$ (Weir and Basten, 1990), does not hold with the settings and the parameter values that we use. If this approximation was satisfied, JC should not depend on L. Instead, given that we estimated the p-distance by the ratio between the number n_d of nucleotide differences and L, n_d seems to increase more slowly than $2\mu L$ when L increases. The increased proportion of genotypes is reflected by an increased richness (R) and a faster increase of the Gini-Simpson diversity D up to its maximum. The immune response mitigates the replication of new genotypes, which does not ensure the genotype abundance evenness (H') even with high values of mutation rate and genome size.

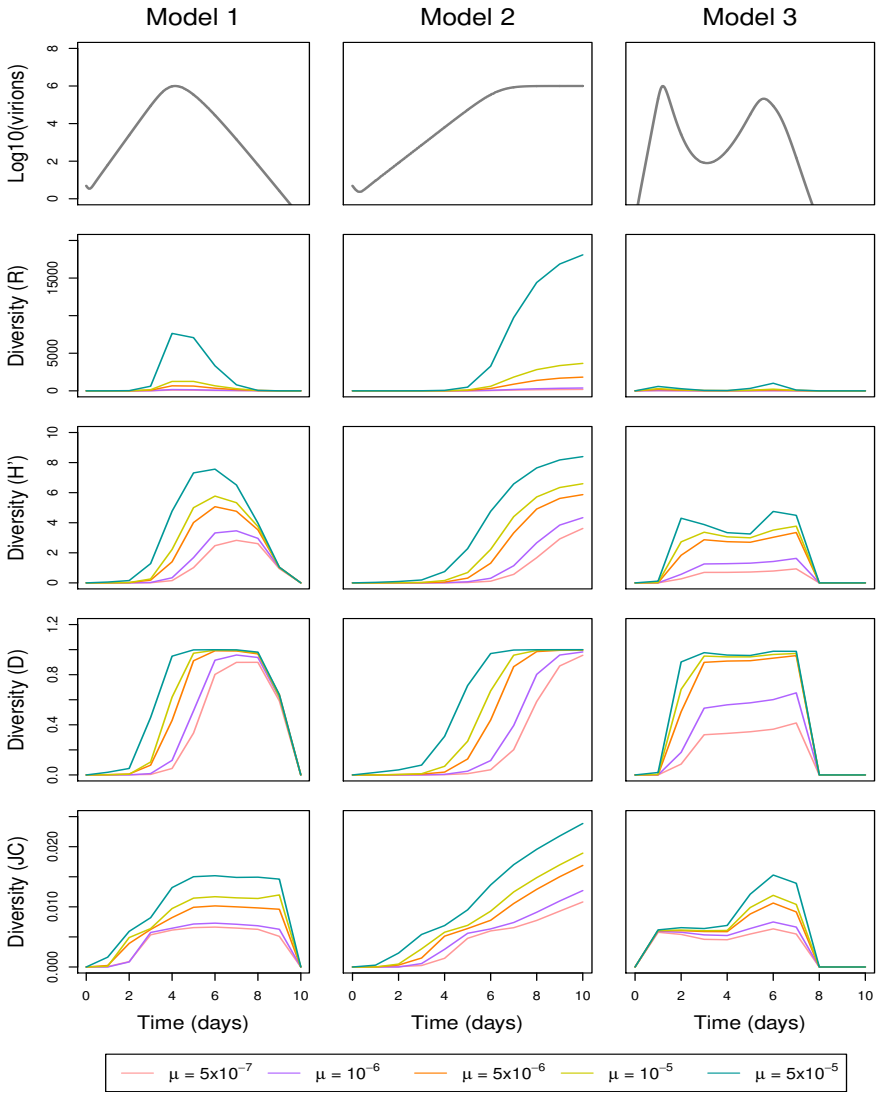


Figure 5: Simulated effects of the mutation rate μ on the levels of within-host genetic diversity. Simulations performed with the shuffling process and the elimination of lethal genomes (with $\gamma_2 = 0.4$ and $\alpha = 0.4$). Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2–5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the mutation rate μ.

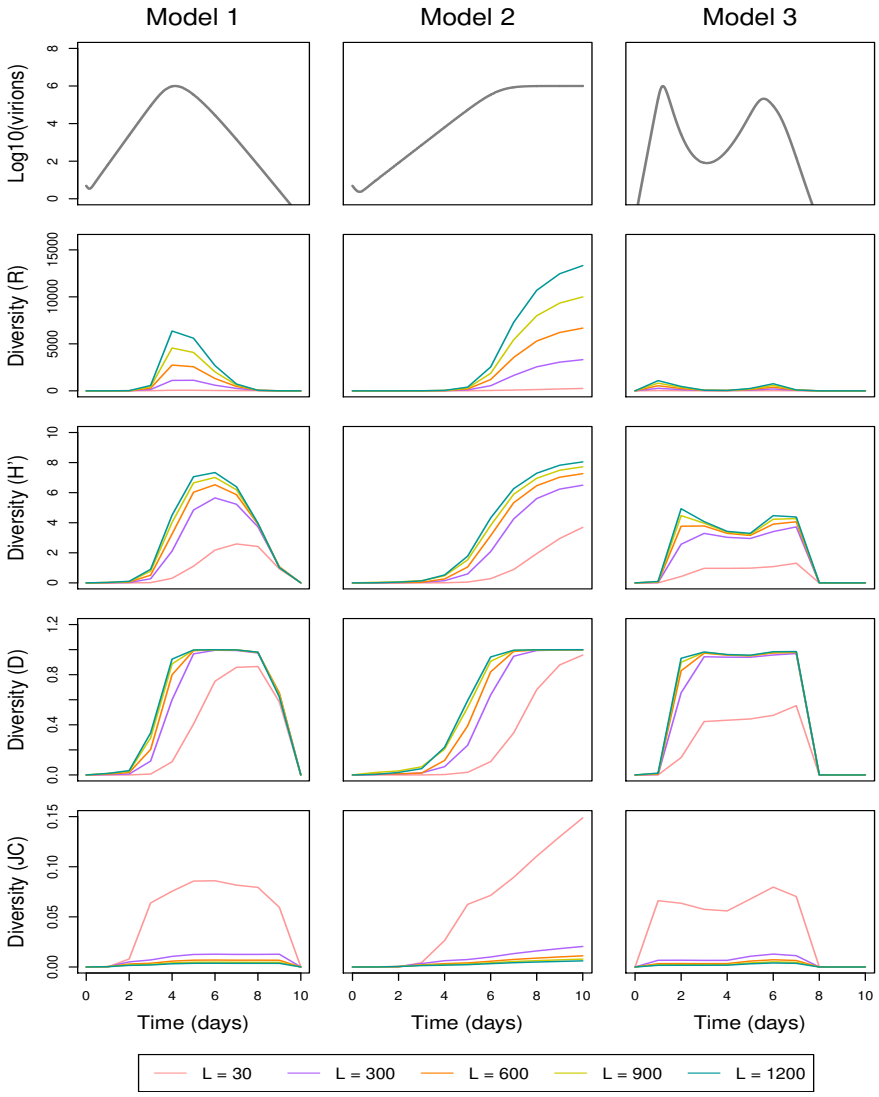


Figure 6: Simulated effects of the genetic sequence size L on the levels of within-host genetic diversity. Simulations performed with the shuffling process and the elimination of lethal genomes (with $\gamma_2 = 0.4$ and $\alpha = 0.4$). Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2–5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the genetic sequence size L.

4 Discussion

In this work, we introduced a stochastic model to simulate within-host pathogen evolution during an infection in order to outline the demographic and genetic factors shaping viral within-host genetic diversity. Our explicit model developed in a forward framework allows us to monitor temporal changes (i.e., across generations) in within-host genetic diversity computed under various demo-genetic scenarios. This model is able to generate very diverse within-host scenarios in terms of viral load and genetic diversity as we illustrate in the result section. Demographic effects are considered mainly through the kinetic model quantifying the temporal variation of the viral load. Genetic effects are considered through mutation and replication modes approximately mimicking natural selection and genetic drift. These modes are based, in particular, on the elimination of lethal genomes (leading to negative selection) and the shuffling of genotype proportions generating over-dispersion with respect to multinomial draws (leading to genetic drift and positive selection). Thus, by coupling the model that we propose with a host-to-host transmission model, we will obtain a flexible basis to challenge, in very diverse settings, the methods that reconstruct transmission trees using within-host genetic diversity data (e.g. Alamil et al., 2019; De Maio et al., 2018).

In contrast to the Wright-Fisher process considering that the total pathogen population size is constant (Fisher, 1923; Wright, 1931; Imhof and Nowak, 2006) and to the Worby and Read process (Worby and Read, 2015) assuming that the size of the pathogen population converges to an attraction function via the sum of binomial jumps, virion-quantity changes during an infection are explicitly modeled (and hence controlled) in our approach, and we can use many existing viral kinetic models found in the specialized literature. Previous studies often based on estimated effective population size (N_e) or viral load show a positive relationship between population size and genetic diversity (Golubchik et al., 2013; Bailey et al., 2014; Nelson and Hughes, 2015) supporting the neutral theory (Kimura, 1983). By accounting for temporal variation in virus load (under different kinetic assumptions) and contrasting diversity measures, we however observe a non-monotonous relationship between pathogen population size and genetic diversity. This may result from the complex interplay between diversity accumulation through time and changes in the size of the pathogen population. Analysing and confronting the variations of different diversity indices in further analyses may provide some clues on the main processes shaping genetic diversity across time.

Interactions between genetic and demographic forces have been pointed out in numerous studies; e.g., the pathogen population size can impact the mutational robustness (Elena et al., 2007) as well as the random genetic drift robustness (LaBar and Adami, 2017; Didelot et al., 2016; Kuo et al., 2009) and the intensity of selection

(Frickel et al., 2018; Didelot et al., 2016; Gutiérrez et al., 2012), which directly affects the composition of the viral population. Our study is an additional illustration of such interactions. Consider as an example the demographic force consisting of the immune response included in the kinetic model 3. The level of within-host genetic diversity and the mutation rate are known to be positively correlated (Castellano et al., 2020; Xu et al., 2019) and we clearly see this with the assessment of Richness and Jukes-Cantor indices in Figure 5. However, the immune response reduces, in general, the impact of mutation on diversity and reduces, in particular, the evenness of mutant genotypes (Shannon index). By considering that the immune response *de facto* induces an additional selective pressure, the negative effect of the immune response on diversity can be viewed as a manifestation of the overall quick response of rapidly mutating viruses (such as RNA viruses) to selection (Domingo and Holland, 1997; Holmes, 2009; Sanjuán, 2010). Interestingly, we observe a non-monotonic effect of the immune response on the Shannon index since a higher diversity (i.e., here, a higher level of homogeneity in genotype abundances) is achieved for an intermediate value of the mutation rate, namely $\mu = 10^{-6}$ (Figure 5, 3rd row, 3rd column). This observation results from a combined effect of the shuffling process and the immune response since the non-monotonic effect does not hold when the shuffling process is removed; see Supporting Figure S8.

As mentioned above, the model can easily incorporate more advanced kinetic models of the number of virions and, hence, be used for example to study the within-host pathogen diversity in the presence of treatment with therapeutic antiviral agents (Smith and Perelson, 2011; Beauchemin et al., 2008), variation of virion infectivity over time (Vaidya et al., 2010; Beauchemin and Handel, 2011), decay of viral infectivity (Smith and Ribeiro, 2010), co-receptor switch (Alizon and Boldin, 2010) and virion loss due to cell entry (Beauchemin et al., 2008). Another perspective is the study of diversity by using more realistic mutation processes (Kimura, 1980; Tavaré, 1986), or by including relative fitness depending on the genetic sequence or on the frequency, which induces frequency-dependent selection (Alizon and Boldin, 2010; Sanjuán et al., 2004). However, to improve model realism, one must not only consider the way the model components are defined, but one must also use realistic parameter values. The statistical estimation of the model parameters from within-host genetic data is likely to be a challenge that firstly requires to assess what accuracy level of data and what inference approach could be adequate.

Table 1: Variables, parameters and values used in acute and chronic models (Eq. 1 & 2) (Baccam et al., 2006).

Symbol	Definition	Unit	Value
S	Uninfected cells that are susceptible to infection	cells	Initial value: 4×10^8
I_1	Infected cells not producing virus	cells	Initial value: 0
I_2	Infected cells actively producing virus	cells	Initial value: 0
V	Viral load	$\text{TCID}_{50}/\text{ml}$	Initial value: 4.9
β	Rate of susceptible target cell infection	$(\text{TCID}_{50}/\text{ml})^{-1}.\text{d}^{-1}$	5.3×10^{-6}
k	$1/k$ is the average transition time from I_1 to I_2	d^{-1}	4
δ	Death rate of infected cells I_2 that actively produce virus	d^{-1}	3.8
p	Viral production rate	$(\text{TCID}_{50}/\text{ml}).\text{d}^{-1}$	0.05
c	Clearance rate of virions	d^{-1}	3.8

*d: day

Table 2: Variables, parameters and values used in the delay model incorporating an immune response (Eq. 3).

Symbol	Definition	Unit	Value
S	Uninfected cells that are susceptible to infection	cells	Initial value: 3.5×10^{11}
I_1	Infected cells not producing virus	cells	Initial value: 0
I_2	Infected cells actively producing virus	cells	Initial value: 0
R	Uninfected refractory cells	cells	Initial value: 0
F	Interferon	IFN fold change	Initial value:5.3
V	Viral load	$\text{TCID}_{50}/\text{ml}$	Initial value: 3.5×10^{-1}
β	Rate of susceptible target cell infection	$(\text{TCID}_{50}/\text{ml})^{-1}.\text{d}^{-1}$	8.3×10^{-6}
ϕ	Rate of the IFN-induced antiviral efficacy	$(\text{IFN fold change})^{-1}.\text{d}^{-1}$	9×10^{-4}
ρ	Reversion rate from refractory to susceptible state	d^{-1}	1.5
k	$1/k$ is the average transition time from I_1 to I_2	d^{-1}	0.55
δ_I	Death rate of infected cells before the emergence of the adaptive immune response	d^{-1}	4
m	killing rate of infected cells by NK cells activated by IFN	$(\text{IFN fold change})^{-1}.\text{d}^{-1}$	2.9×10^{-3}
p	Viral production rate	$(\text{TCID}_{50}/\text{ml}).\text{d}^{-1}$	4.8×10^{-3}
c	Clearance rate of virions	d^{-1}	11.5
q	Rate of IFN production	$(\text{IFN fold change})\ \text{cell}^{-1}$	1.1×10^{-5}
d	Rate of IFN decay	d^{-1}	0.72
σ	Speed of the death rate increase		4

*d: day

References

Abel, S., P. A. zur Wiesch, B. M. Davis, and M. K. Waldor (2015). Analysis of bottlenecks in experimental models of infection. *PLoS pathogens 11*(6).

Alamil, M., J. Hughes, K. Berthier, C. Desbiez, G. Thébaud, and S. Soubeyrand (2019). Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. *Philosophical Transactions of the Royal Society B 374*(1775), 20180258.

Alizon, S. and B. Boldin (2010). Within-host viral evolution in a heterogeneous environment: insights into the hiv co-receptor switch. *Journal of evolutionary biology 23*(12), 2625–2635.

Alizon, S., F. Luciani, and R. R. Regoes (2011). Epidemiological and clinical consequences of within-host evolution. *Trends in microbiology 19*(1), 24–32.

Baccam, P., C. Beauchemin, C. A. Macken, F. G. Hayden, and A. S. Perelson (2006). Kinetics of influenza a virus infection in humans. *Journal of virology 80*(15), 7590–7599.

Bailey, A. L., M. Lauck, A. Weiler, S. D. Sibley, J. M. Dinis, Z. Bergman, C. W. Nelson, M. Correll, M. Gleicher, D. Hyeroba, et al. (2014). High genetic diversity and adaptive potential of two simian hemorrhagic fever viruses in a wild primate population. *PloS one 9*(3), e90714.

Beauchemin, C. A. and A. Handel (2011). A review of mathematical models of influenza a infections within a host or cell culture: lessons learned and challenges ahead. *BMC public health 11*(1), S7.

Beauchemin, C. A., J. J. McSharry, G. L. Drusano, J. T. Nguyen, G. T. Went, R. M. Ribeiro, and A. S. Perelson (2008). Modeling amantadine treatment of influenza a virus in vitro. *Journal of theoretical biology 254*(2), 439–451.

Biek, R., O. G. Pybus, J. O. Lloyd-Smith, and X. Didelot (2015). Measurably evolving pathogens in the genomic era. *Trends in ecology & evolution 30*(6), 306–313.

Brunker, K., K. Hampson, D. Horton, and R. Biek (2012). Integrating the landscape epidemiology and genetics of rna viruses: rabies in domestic dogs as a model. *Parasitology 139*(14), 1899–1913.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes (2016). Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods 13*(7), 581.

Canini, L. and A. S. Perelson (2014). Viral kinetic modeling: state of the art. *Journal of pharmacokinetics and pharmacodynamics 41*(5), 431–443.

Castellano, D., A. Eyre-Walker, and K. Munch (2020). Impact of mutation rate and selection at linked sites on dna variation across the genomes of humans and other homininae. *Genome biology and evolution 12*(1), 3550–3561.

Cottam, E. M., G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, and D. T. Haydon (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society of London B: Biological Sciences 275*(1637), 887–895.

Cuevas, J. M., R. Geller, R. Garijo, J. López-Aldeguer, and R. Sanjuán (2015). Extremely high mutation rate of hiv-1 in vivo. *PLoS biology 13*(9).

De Maio, N., C. J. Worby, D. J. Wilson, and N. Stoesser (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS computational biology 14*(4), e1006117.

Didelot, X., J. Gardy, and C. Colijn (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution 31*(7), 1869–1879.

Didelot, X., A. S. Walker, T. E. Peto, D. W. Crook, and D. J. Wilson (2016). Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology 14*(3), 150.

Domingo, E. and J. Holland (1997). Rna virus mutations and fitness for survival. *Annual review of microbiology 51*(1), 151–178.

Elena, S. F., C. O. Wilke, C. Ofria, and R. E. Lenski (2007). Effects of population size and mutation rate on the evolution of mutational robustness. *Evolution 61*(3), 666–674.

Fisher, R. A. (1923). Xxi.—on the dominance ratio. *Proceedings of the royal society of Edinburgh 42*, 321–341.

Frickel, J., P. G. Feulner, E. Karakoc, and L. Becks (2018). Population size changes and selection drive patterns of parallel evolution in a host–virus system. *Nature communications 9*(1), 1–10.

Fudala, A. and R. Korona (2009). Low frequency of mutations with strongly deleterious but nonlethal fitness effects. *Evolution: International Journal of Organic Evolution 63*(8), 2164–2171.

Galan, M., E. Guivier, G. Caraux, N. Charbonnel, and J.-F. Cosson (2010). A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC genomics 11*(1), 296.

Galan, M., M. Pages, and J.-F. Cosson (2012). Next-generation sequencing for rodent barcoding: species identification from fresh, degraded and environmental samples. *PLoS One 7*(11).

Golubchik, T., E. M. Batty, R. R. Miller, H. Farr, B. C. Young, H. Larner-Svensson, R. Fung, H. Godwin, K. Knox, A. Votintseva, et al. (2013). Within-host evolution of staphylococcus aureus during asymptomatic carriage. *PLoS One 8*(5), e61319.

Gutiérrez, S., Y. Michalakis, and S. Blanc (2012). Virus population bottlenecks during within-host progression and host-to-host transmission. *Current opinion in virology 2*(5), 546–555.

Hall, M., M. Woolhouse, and A. Rambaut (2015). Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS computational biology 11*(12), e1004613.

Handel, A., I. M. Longini, and R. Antia (2010). Towards a quantitative understanding of the within-host dynamics of influenza a infections. *Journal of the Royal Society Interface 7*(42), 35–47.

Holmes, E. C. (2009). *The evolution and emergence of RNA viruses*. Oxford University Press.

Hughes, J., R. C. Allen, M. Baguelin, K. Hampson, G. J. Baillie, D. Elton, J. R. Newton, P. Kellam, J. L. Wood, E. C. Holmes, et al. (2012). Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS pathogens 8*(12), e1003081.

Imhof, L. A. and M. A. Nowak (2006). Evolutionary game dynamics in a wright-fisher process. *Journal of mathematical biology 52*(5), 667–681.

Jombart, T., A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS computational biology 10*(1), e1003457.

Jukes, T. H., C. R. Cantor, et al. (1969). Evolution of protein molecules. *Mammalian protein metabolism 3*(21), 132.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution 16*(2), 111–120.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kulkarni, P. and P. Frommolt (2017). Challenges in the setup of large-scale next-generation sequencing analysis workflows. *Computational and structural biotechnology journal 15*, 471–477.

Kuo, C.-H., N. A. Moran, and H. Ochman (2009). The consequences of genetic drift for bacterial genome complexity. *Genome research 19*(8), 1450–1454.

LaBar, T. and C. Adami (2017). Evolution of drift robustness in small populations. *Nature communications 8*(1), 1–12.

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution 30*(2), 314–334.

Lau, M. S., G. Marion, G. Streftaris, and G. Gibson (2015). A systematic bayesian integration of epidemiological and genetic data. *PLoS computational biology 11*(11).

Leitner, T. and E. Romero-Severson (2018). Phylogenetic patterns recover known hiv epidemiological relationships and reveal common transmission of multiple variants. *Nature microbiology 3*(9), 983–988.

Mollentze, N., L. H. Nel, S. Townsend, K. Le Roux, K. Hampson, D. T. Haydon, and S. Soubeyrand (2014). A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B 281*(1782), 20133251.

Morelli, M. J., G. Thébaud, J. Chadœuf, D. P. King, D. T. Haydon, and S. Soubeyrand (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol 8*, e1002768.

Morris, E. K., T. Caruso, F. Buscot, M. Fischer, C. Hancock, T. S. Maier, T. Meiners, C. Müller, E. Obermaier, D. Prati, et al. (2014). Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories. *Ecology and evolution 4*(18), 3514–3524.

Murcia, P. R., J. Hughes, P. Battista, L. Lloyd, G. J. Baillie, R. H. Ramirez-Gonzalez, D. Ormond, K. Oliver, D. Elton, J. A. Mumford, et al. (2012). Evolution of an eurasian avian-like influenza virus in naive and vaccinated pigs. *PLoS Pathogens 8*(5), e1002730.

Nelson, C. W. and A. L. Hughes (2015). Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. *Infection, Genetics and Evolution 30*, 1–7.

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

Pawelek, K. A., G. T. Huynh, M. Quinlivan, A. Cullinane, L. Rong, and A. S. Perelson (2012). Modeling within-host dynamics of influenza virus infection including immune responses. *PLoS Comput Biol 8*(6), e1002588.

Picard, C., S. Dallot, K. Brunker, K. Berthier, P. Roumagnac, S. Soubeyrand, E. Jacquot, and G. Thébaud (2017). Exploiting genetic information to trace plant virus dispersal in landscapes. *Annual review of phytopathology 55*, 139–160.

Piry, S., C. Wipf-Scheibel, J.-F. Martin, M. Galan, and K. Berthier (2017). High throughput amplicon sequencing to assess within-and between-host genetic diversity in plant viruses. *bioRxiv*, 168773.

Poirier, E. Z. and M. Vignuzzi (2017). Virus population dynamics during infection. *Current opinion in virology 23*, 82–87.

Pybus, O. G. and A. Rambaut (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics 10*(8), 540–550.

Saenz, R. A., M. Quinlivan, D. Elton, S. MacRae, A. S. Blunden, J. A. Mumford, J. M. Daly, P. Digard, A. Cullinane, B. T. Grenfell, et al. (2010). Dynamics of influenza virus infection and pathology. *Journal of virology 84*(8), 3974–3983.

Sanjuán, R. (2010). Mutational fitness effects in rna and single-stranded dna viruses: common patterns revealed by site-directed mutagenesis studies. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1548), 1975–1982.

Sanjuán, R., A. Moya, and S. F. Elena (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an rna virus. *Proceedings of the National Academy of Sciences 101*(22), 8396–8401.

Simmons, H., J. Dunham, J. Stack, B. Dickins, I. Pagan, E. Holmes, and A. Stephenson (2012). Deep sequencing reveals persistence of intra-and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. *Journal of general virology 93*(8), 1831–1840.

Smith, A. M. and A. S. Perelson (2011). Influenza a virus infection kinetics: quantitative data and models. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine 3*(4), 429–445.

Smith, A. M. and R. M. Ribeiro (2010). Modeling the viral dynamics of influenza a virus infection. *Critical Reviews^{TM} in Immunology 30*(3).

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics 135*(2), 599–607.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences 17*(2), 57–86.

Vaidya, N. K., R. M. Ribeiro, C. J. Miller, and A. S. Perelson (2010). Viral dynamics during primary simian immunodeficiency virus infection: effect of time-dependent virus infectivity. *Journal of virology 84*(9), 4302–4310.

Valdazo-González, B., J. T. Kim, S. Soubeyrand, J. Wadsworth, N. J. Knowles, D. T. Haydon, and D. P. King (2015). The impact of within-herd genetic variation upon inferred transmission trees for foot-and-mouth disease virus. *Infection, Genetics and Evolution 32*, 440–448.

Walker, T. M., C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, et al. (2013). Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases 13*(2), 137–146.

Weir, B. and C. Basten (1990). A biometrics invited paper with discussion. sampling strategies for distances between dna sequences. *Biometrics*, 551–582.

Worby, C. J., M. Lipsitch, and W. P. Hanage (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS computational biology 10*(3).

Worby, C. J. and T. D. Read (2015). 'seedy'(simulation of evolutionary and epidemiological dynamics): An r package to follow accumulation of within-host mutation in pathogens. *PloS one 10*(6), e0129745.

Wright, S. (1931). Evolution in mendelian populations. *Genetics 16*(2), 97.

Wymant, C., M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, and M. Cornelissen (2018). Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular biology and evolution 35*(3), 719–733.

Xu, S., J. Stapley, S. Gablenz, J. Boyer, K. J. Appenroth, K. S. Sree, J. Gershenzon, A. Widmer, and M. Huber (2019). Low genetic variation is associated with low mutation rate in the giant duckweed. *Nature communications 10*(1), 1–6.

Ypma, R. J., A. Bataille, A. Stegeman, G. Koch, J. Wallinga, and W. M. Van Ballegooijen (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society of London B: Biological Sciences 279*(1728), 444–450.

Ypma, R. J., W. M. van Ballegooijen, and J. Wallinga (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics 195*(3), 1055–1062.

4.3. Article 2

Factors influencing the inference of transmission events in disease outbreaks

Alamil M.[1,2,*], Bruchou C.[1], Ribaud M.[1,], Thébaud G.[3], and Soubeyrand S.[1]

[1]INRAE, BioSP, 84914 Avignon, France
[2]LMA, Université d'Avignon, 84140, AVIGNON
[3]BGPI, INRAE, Univ Montpellier, Cirad, Institut Agro, Montpellier, France
[*]Corresponding author: maryam.alamil@inrae.fr

Abstract

Collecting pathogen sequence data from multiple hosts in an infectious disease outbreak offers the possibility to infer who infected whom, potentially giving valuable insights into the disease dynamics that can be useful to improve control strategies. However, this general idea requires an effective approach to infer transmission pathways from such genetic data. The accuracy of the inference approach can be affected by many factors and processes (e.g. sequencing properties and within-host pathogen evolution). Here, we precisely investigate how such factors and processes influence transmission reconstruction in various challenging demo-genetic situations. This investigation is carried out on `SLAFEEL` method, a Statistical Learning Approach For Estimating Epidemiological Links from deep sequencing data. We performed a sensitivity analysis of the relationship between `SLAFEEL` accuracy and the sampling and demo-genetic factors that may impact virus evolution and transmission as well as the collected data. Using Sobol's indices, we quantified the impact of all factors on `SLAFEEL` accuracy. Results show that the most influential factors are those affecting the total diversity covering within- and between-host genetic diversity. Furthermore, we identified an optimum for the total observed diversity, which allows the transmission network to be reconstructed with a relatively high level of accuracy.

Keywords: genetic diversity, inference accuracy, metamodel, sensitivity analysis, simulation model, Sobol's indices, viral transmission, within-host pathogen evolution

1 Introduction

Understanding how pathogens spread within host populations is crucial for effective epidemiological predictions and control strategies that limit the spread of infectious diseases. Reconstructing transmission links enables to get valuable insights, describing the history of infection between host units (Ferguson et al., 2001; Wallinga and Teunis, 2004; Spada et al., 2004; Lloyd-Smith et al., 2005). In the past dozen years, significant progress has been made in the development of methods for inferring epidemiological links from pathogen sequence data (Cottam et al., 2008; Morelli et al., 2012; Mollentze et al., 2014; Jombart et al., 2014; Didelot et al., 2014, 2017; Lau et al., 2015; Worby et al., 2016; De Maio et al., 2016, 2018; Skums et al., 2018; Leitner and Romero-Severson, 2018; Wymant et al., 2018; Campbell et al., 2019) As a result, estimating transmission links became increasingly accurate, robust and rapid. However, the level of accuracy of these approaches may be impacted by many complications such as sequencing properties and within-host viral evolution. Here, we present a study aiming at determining which forces may obstruct the inference of transmission events.

A panel of recent studies reveal that demographic and genetic factors playing a role in within-host pathogen evolution and host-to-host transmission, as well as sampling factors, may hinder or improve the ability to reconstruct transmission links (Hang et al., 2007; Worby et al., 2014; De Maio et al., 2016; Soubeyrand, 2016; Worby et al., 2016; De Maio et al., 2018; Campbell et al., 2018). These studies investigate the predictive ability using simulated data generated by different `R` software codes or packages like `seedy` (Worby and Read, 2015), `outbreaker` (Jombart et al., 2014) and `phybreak` (Klinkenberg et al., 2017), providing transmission histories under a variety of scenarios. The comparison between simulated and estimated transmission chains enabled to assess in which scenario the inference approach performs appropriately. This comparison was typically performed by assessing the proportion of correct transmission source identification or by examining the receiver-operating characteristic (ROC) curves (Krzanowski and Hand, 2009).

According to these studies, within-host variation (affected by the infectious period and the transmission bottleneck), genetic diversity at transmission (limited by transmission bottlenecks) and sampling techniques are the main factors influencing the accuracy of transmission inference approaches. Moreover, some of these studies were able to quantify the effect of the size of virtual populations and their mutation and recombination rates. However, they did not investigate the role of factors that significantly impact within-host genetic variability such as natural selection, and variations in within-host population size through time (Alamil et al., 2020; Frickel et al., 2018; Rousseau et al., 2017; LaBar and Adami, 2017; Didelot et al., 2016; Gutiérrez et al., 2012; Kuo et al., 2009). Assuming that these factors could influence

the accuracy of transmission routes reconstruction, we addressed their effect.

To gain insight into these issues, we study how well the SLAFEEL approach (Alamil et al., 2019) performs with simulated data generated under various scenarios depending on forces impacting within-host pathogen evolution, viral load and the observation. SLAFEEL is a Statistical Learning Approach For Estimating Epidemiological Links of infectious diseases (caused by fast-evolving pathogens) from deep sequencing data. Its principle is to learn the structure of the epidemiological links with a pseudo-evolutionary model applied to training data, and then to use this initial training stage for the inference of the links for the whole data set. Modelling pathogen evolution within the host, virus transmission between hosts, and sequencing, we simulate a broad range of outbreak scenarios under multiple demographic, genetic and sampling conditions impacting the observed pathogen diversity. Then, we apply sensitivity analysis methodology for exploring the relationship between SLAFEEL accuracy and the uncertainty about factors that may impact the virus evolution, transmission and sampling. This sensitivity analysis based on Sobol's indices (Sobol, 1967, 1976; Saltelli et al., 2000, 2008) is designed to uncover the most influential factors on the efficiency of SLAFEEL in inferring transmission links.

Beyond unravelling the links between SLAFEEL accuracy and the factors mentioned above, we investigate the relationship between SLAFEEL accuracy and an aggregated indicator measuring the *total observed genetic diversity* of the pathogen population within and across hosts. This indicator can be viewed as a measure of the information brought, through the prism of observation, by genetic data for the inference of epidemiological links during an outbreak.

Finally, this work provides an opportunity to identify the optimal operating range of the first version of SLAFEEL, and points out the need to further explore the advantage and the difficulty of exploiting the within-host genetic variation of the pathogen to infer transmission links.

2 Model and methods

To investigate the relationship between the accuracy of the reconstruction of transmission chains (using SLAFEEL) on the one hand, and demographic, genetic and sampling factors (that may impact observed data) on the other hand, we simulated data from a flexible stochastic model of evolution, transmission and sampling of pathogen sequence populations (which allows us to consider multiple scenarios characterized, e.g., by low/high mutation rates, low/high fitness differences between viral variants, low/high sampling effort, slow/rapid outbreak, weak/strong transmission bottleneck...), and performed a sensitivity analysis after applying the SLAFEEL method to each simulated data set.

2.1 Within-host pathogen evolution

Alamil et al. (2020) proposed a flexible stochastic model of the temporal changes in the within-host genetic composition and size of a viral population. This model provides, across virus generations, sequences and frequencies of variants under different demo-genetic situations. In this model, demographic effects are handled via the initialisation of infection (with a single variant or multiple variants) and the use of a demographic kinetic model (e.g., described by a set of ordinary differential equations) specifying the viral load during the course of infection. Genetic effects are conditional on the demography and correspond to the mutation and replication processes of virus variants subjected to natural selection and random genetic drift. Nucleotide substitutions are assumed to occur randomly at a constant rate $\mu > 0$, and are classified as lethal (extreme negative selection) and non-lethal. The replication of non-lethal variants can be impacted by a sort of frequency-dependent selection amplifying temporal fluctuations in the frequencies of variants. These fluctuations are governed by a shuffling process, which yields more or less noisy variant proportions with respect to the current ones and implicitly accounts for the effect of viral evolution by natural selection and random genetic drift. The shuffling process is parameterized by the vector $(\gamma_1, \gamma_2, \gamma_3) \in \mathbb{R}_+^3$. From Alamil et al. (2020), γ_2 (which directly allows rare variants to reach relatively high proportions) has a particularly large impact on the within-host genetic diversity.

2.2 Host-to-host transmission dynamics

We model the transmission dynamics as a stochastic, individual-based, SI (susceptible – infectious) model conditional on the within-host dynamics of the pathogen in infected hosts. The host units are assumed to form a completely mixing population of size M, in which two hosts have equal contact probability whatever the hosts and their infection state. The outbreak is initiated at time $t = 0$ with the introduction of a single pathogen genome in a single host, the remaining individuals being susceptible. At any time, a given host is either in the susceptible state (S) or in the infectious state (I). An infectious host remains in this state during exactly $\Delta = 10$ days; it cannot be re-infected during the infectious period, and returns in the susceptible state afterwards. The pathogen spreads within the host population through direct contacts between infectious and susceptible hosts (the viral load in the infectious host determines the probability for a contact to lead to a transmission event).

The set of contact times $\mathbf{T}_h$ between the infectious host h and susceptible hosts forms a non-homogeneous Poisson process (over the temporal window corresponding to the infectious period of h) with intensity function $\lambda S(t)$ depending on a constant

contact rate $\lambda > 0$ and the time-varying number of susceptible hosts $S(t)$. If $\mathbf{T}_h$ is not empty, for each contact time T_{hi} included in $\mathbf{T}_h$, $i = 1, \ldots, N_h$ with N_h the number of contacts between h and susceptible hosts during the infectious period of h, we denote H_{hi} the susceptible host in contact with h and $\mathbb{P}_h(T_{hi})$ the probability of transmission from h to H_{hi}. The probability $\mathbb{P}_h(T_{hi})$ does not depend on H_{hi} because the susceptible hosts are assumed to be equally susceptible, but depends on T_{hi} because the viral load in h at the contact time is assumed to impact the transmission success. It is assumed to satisfy:

$$\mathbb{P}_h(T_{hi}) = \frac{V_h(T_{hi})}{V_h^{\max}},$$

where $V_h(t)$ is the viral load in h at time t and $V_h^{\max}$ is the maximum value of V_h during the infectious period.

The viral kinetic and composition in the host H_{hi} is initiated at time T_{hi} by sub-sampling at the same time the viral population within the source host h. At transmission time T_{hi}, in h the size of this viral population is $V_h(T_{hi})$ and the vector of variant proportions is denoted by $p_h(T_{hi})$. The initial vector of variant frequencies $F_{H_{hi}}(T_{hi})$ in H_{hi} is modeled as a multinomial draw with size equal to the ceiling value of $\tau V_h(T_{hi})$ and with the vector of probabilities $p_h(T_{hi})$:

$$F_{H_{hi}}(T_{hi}) \sim \text{Multinomial}\left(\lceil \tau V_h(T_{hi}) \rceil, p_h(T_{hi})\right),$$

where $\tau > 0$ is the relative transmission bottleneck, i.e. the proportion of the viral population in h that is transmitted to host H_{hi}.

2.3 Sampling

The sampling time, denoted by T_h^{samp}, of the viral population in an infected host h is drawn from an exponential distribution whose 'zero' is the time at which the viral load in h exceeds 60% of its maximum value $V_h^{\max}$ and which is cut off for avoiding T_h^{samp} to be beyond the end of the infection:

$$\begin{aligned} T_h^{\text{samp}} - t_h &= \max\{X_h, T_h^{\text{inf}} + \Delta - t_h\} \\ X_h &\sim \text{Exponential}(\eta) \end{aligned}$$

where $t_h = \min_{t \in [T_h^{\text{inf}}, T_h^{\text{inf}} + \Delta]}\{V_h(t) = 0.6 \times V_h^{\max}\}$, T_h^{inf} is the time of infection of host h, and $\eta > 0$ is the rate of the exponential distribution (we remind that Δ is the infection duration and is equal to 10 days). Then, a set of N aligned genetic fragments of L nucleotides are sampled from the infected host h using a multinomial distribution with vector of probabilities equal to $p_h(T_h^{\text{samp}})$. L is the size of the

sequenced fragment, and N is the sequencing depth. We assume that T_h^{samp} is both the sampling time of the viral population in h as defined above and the time at which h is observed as infected.

2.4 Simulation of outbreak scenarios

To evaluate the accuracy of `SLAFEEL` in estimating epidemiological links, we investigate a broad range of outbreak scenarios.

In terms of demography, we consider four situations where the viral load in any infectious individual is drawn from one of the following viral kinetic model:

- $\mathcal{K}_1$: acute infection kinetic model (see Eq. 1-3 in Baccam et al. 2006);
- $\mathcal{K}_2$: acute infection kinetic model with a latent period (see Eq. 1 in Alamil et al. 2020 and Eq. 5-8 in Baccam et al. 2006);
- $\mathcal{K}_3$: immunity-cured infection kinetic model (see Eq. 3 in Alamil et al. 2020 and Eq. 3 in the Supporting Text S1 of Pawelek et al. 2012);
- $\mathcal{K}_{\text{mixed}}$: one of the three previous models randomly and equiprobably selected.

In addition, we vary two sampling parameters (the size of sequenced fragments L and sequencing depth N), two parameters relating to virus evolution (the mutation rate μ and the shuffling parameter γ_2), two epidemiological parameters (the contact rate λ and relative transmission bottleneck (transmission rate) τ), and the host population size M. Although the number of infected hosts is not controlled, we discard simulations with less than 10 infected hosts, and we stop outbreaks when they reach 30 infected hosts. The parameters and their values (or ranges) used to simulate outbreaks are summarized in Table 1.

Table 1: Parameters and values (or ranges) used to simulate outbreaks.

Symbol	Definition	Unit	Value/range
Within-host pathogen evolution			
μ	Mutation rate	mutation per nucleotide per generation	$[5 \times 10^{-7}; 5 \times 10^{-5}]$
γ_1	First shuffling parameter		0.8
γ_2	Second shuffling parameter		$[0; 1]$
γ_3	Third shuffling parameter		70
α	Proportion of lethal genomes		0.4
Acute infection model without delay			
S	Uninfected cells that are susceptible to infection	cells	Initial value: 4×10^8
I	Infected cells	cells	Initial value: 0
V	Viral load	$\text{TCID}_{50}/\text{ml}$	Initial value*: 3.5×10^{-1}
β	Rate of susceptible target cell infection	$(\text{TCID}_{50}/\text{ml})^{-1}.\text{d}^{-1}$	3.4×10^{-5}
δ	Death rate of infected cells	d^{-1}	3.4
p	Viral production rate	$(\text{TCID}_{50}/\text{ml}).\text{d}^{-1}$	0.0239
c	Clearance rate of virions	d^{-1}	3.3
Acute infection model with delay			
S	Uninfected cells that are susceptible to infection	cells	Initial value: 4×10^8
I_1	Infected cells not producing virus	cells	Initial value: 0
I_2	Infected cells actively producing virus	cells	Initial value: 0
V	Viral load	$\text{TCID}_{50}/\text{ml}$	Initial value*: 4.9
β	Rate of susceptible target cell infection	$(\text{TCID}_{50}/\text{ml})^{-1}.\text{d}^{-1}$	5.3×10^{-6}
k	$1/k$ is the average transition time from I_1 to I_2	d^{-1}	4
δ	Death rate of infected cells I_2 that actively produce virus	d^{-1}	3.8
p	Viral production rate	$(\text{TCID}_{50}/\text{ml}).\text{d}^{-1}$	0.05
c	Clearance rate of virions	d^{-1}	3.8
Immunity-cured infection model			
S	Uninfected cells that are susceptible to infection	cells	Initial value: 3.5×10^{11}
I_1	Infected cells not producing virus	cells	Initial value: 0
I_2	Infected cells actively producing virus	cells	Initial value: 0
R	Uninfected refractory cells	cells	Initial value: 0
F	Interferon	IFN fold change	Initial value:5.3
V	Viral load	$\text{TCID}_{50}/\text{ml}$	Initial value*: 3.5×10^{-1}
β	Rate of susceptible target cell infection	$(\text{TCID}_{50}/\text{ml})^{-1}.\text{d}^{-1}$	8.3×10^{-6}
ϕ	Rate of the IFN-induced antiviral efficacy	$(\text{IFN fold change})^{-1}.\text{d}^{-1}$	9×10^{-4}
ρ	Reversion rate from refractory to susceptible state	d^{-1}	1.5
k	$1/k$ is the average transition time from I_1 to I_2	d^{-1}	0.55
δ_I	Death rate of infected cells before the emergence of the adaptive immune response	d^{-1}	4
m	Killing rate of infected cells by NK cells activated by IFN	$(\text{IFN fold change})^{-1}.\text{d}^{-1}$	2.9×10^{-3}
p	Viral production rate	$(\text{TCID}_{50}/\text{ml}).\text{d}^{-1}$	4.8×10^{-3}
c	Clearance rate of virions	d^{-1}	11.5
q	Rate of IFN production	$(\text{IFN fold change})\ \text{cell}^{-1}$	1.1×10^{-5}
d	Rate of IFN decay	d^{-1}	0.72
σ	Speed of the death rate increase		4
Host-to-host transmission dynamic			
Δ	Duration of infection of each host	days	10
M	Host population size	hosts	$[10^3; 5 \times 10^7]$
λ	Contact rate		$[5 \times 10^{-4}; 1]$
τ	Transmission rate (Relative transmission bottleneck)		$[10^{-6}; 1]$
Z	Number of infected hosts	hosts	30
Sampling			
N	Sequencing depth	fragments	$[1; 500]$
L	Fragment size	nucleotide bases	$[30; 900]$
η	Rate of the exponential law used for defining sampling times		2/3

* Only for the first infected host.

2.5 Inference of transmission links

We use the SLAFEEL approach for inferring epidemiological links from deep sequencing data. Here, we only sketch this approach that is graphically represented and detailed in Alamil et al. (2019). SLAFEEL is grounded on statistical learning and a pseudo-evolutionary model. This model concisely describes transitions between sets of sequences sampled from different host units and is used to assess the probability of source-recipient pairs under epidemiological constraint. The model was designed as a regression function where the response variable is the set of sequences $\mathbf{S}$ observed from a recipient host and the explanatory variable $\mathbf{S}_0$ is the set of sequences observed from a putative source. The coefficients of the regression are weights accounting for the gain and loss of virus variants during within-host evolution and their loss during between-host transmission. These weights measure how much each sequence in $\mathbf{S}_0$ contributes to explain the sequences in $\mathbf{S}$.

Based on the pseudo-evolutionary model, we define a penalized pseudo-likelihood. One can consider different shapes for the penalization depending on the hypotheses that one makes and the available data. Here, we consider a penalization such that the distance between sequences in $\mathbf{S}$ and their contributing sequences in $\mathbf{S}_0$ is consistent with known features, namely with its expected value and its variance estimated from the training data set. This penalization corresponds to the H2-normal shape proposed by Alamil et al. (2019). The strength of the penalization is calibrated with the training data and, using the optimal penalization parameter value(s), we provide a quantitative assessment of the link intensity between any recipient host and any putative source.

In this approach, one has to *a priori* determine the putative sources for each infected host. Here, we assume that any host observed as infected up to 2 days after observation of the focal host h is a putative source for h.

Training data that are used to calibrate the penalization in the model consist of the knowledge of the source hosts for a set of six hosts at the most randomly drawn among the infected hosts. The number of transmissions that are known may be lower than 6 when some of the hosts that are drawn have very small viral loads at the sampling time and, consequently, no sequence can be sampled.

The performance of SLAFEEL in inferring epidemiological links is evaluated with an accuracy measure, namely the proportion of correct source identification, i.e. the proportion of hosts in each simulated outbreak for which the actual source coincide with the putative source with the highest link intensity.

2.6 Sensitivity analysis

We apply global sensitivity analysis to identify the key demographic, genetic and sampling factors impacting the reconstruction of epidemiological links. Schematically, a sensitivity analysis consists in: (i) determining the input parameters and assigning their respective variation ranges or their probability distributions; (ii) sampling the parameter space with a numerical experimental design; (iii) running the *code* yielding the output variable for each point in the parameter space; and (iv) assessing the influence of each input factor on the output variable by computing sensitivity indices (Faivre et al., 2016; Saltelli et al., 2000).

In our study, the input are demo-genetic factors whose impact on the within-host pathogen evolution has been demonstrated (Alamil et al., 2020), as well as factors related to host-to-host viral transmission and pathogen sampling. These factors are mentioned in Section 2.4 and are recalled here: host population size (M), mutation rate (μ), sequencing depth (N), transmission rate or relative transmission bottleneck (τ), contact rate between hosts (λ), shuffling parameter (γ_2), sequenced fragment size (L) and viral kinetic model quantifying the temporal viral load within a host through time ($\mathcal{K}$).

The ranges of parameters and the modalities of the kinetic model that are considered in the sensitivity analysis are given in Table 2.

Table 2: Variation ranges of the input parameters and modalities of the kinetic model.

Parameter	**Description**	**Set of values**
M	Population size	$[10^3; 5 \times 10^7]$
μ	Mutation rate	$[5 \times 10^{-7}; 5 \times 10^{-5}]$
N	Sequencing depth	$[1; 500]$
τ	Transmission rate (Relative bottleneck size)	$[10^{-6}; 1]$
λ	Contact rate	$[5 \times 10^{-4}; 1]$
γ	Shuffling parameter	$[0; 1]$
L	Sequenced fragment size	$[30; 900]$
$\mathcal{K}$	Viral kinetic model	$\{\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3, \mathcal{K}_{\text{mixed}}\}$

To assess the relative influence of evolution, transmission and sampling parameters on SLAFEEL performance, we compute the following variance-based sensitivity indices (or Sobol indices): the first-order sensitivity index ($\mathcal{I}_1$) and the total sensitivity index ($\mathcal{I}_{\text{tot}}$) (Saltelli et al., 2000, 2008; Monod et al., 2006). $\mathcal{I}_1$ is a measure of the main effect of the input factor of interest on the output variable. $\mathcal{I}_{\text{tot}}$ is a measure of the influence of the input factor of interest when one also takes into account its interactions with other input factors. These indices vary between 0 and 1 and $\mathcal{I}_{\text{tot}} \geq \mathcal{I}_1$. The larger the index, the larger the influence of the input variation (or uncertainty) on the output variation.

To reduce computational time, Sobol indices are computed using a fractional factorial design with resolution V (Droesbeke et al., 1997; Saltelli et al., 2000) to generate $4^4 = 256$ different parameter combinations. The input parameters are equally treated by fixing four levels for each parameter (the number of levels was fixed according to the number of modalities for the unique qualitative input that we considered, namely the number of options for the kinetic model $\mathcal{K}$). The levels of the quantitative variables were set following an OAT (one-at-time) study (Daniel, 1973; Saltelli et al., 2000, 2008).The OAT method allowed us to statistically explore how the accuracy of SLAFEEL evolves according to each input parameter. Varying one parameter at a time (and fixing the other parameters) helped us in identifying the values of this parameter at which eventual significant changes in SLAFEEL accuracy are observed; then, these eventual values were used to define the above-mentioned levels. For each parameter combination, 20 independent replications are made (a replication consists of simulating an outbreak, sampling data and inferring transmission links with SLAFEEL). Then, the relationship between SLAFEEL accuracy and input factors is modeled with a generalized linear meta-model taking into account interactions up to order 2 and assuming that the response variable is drawn from a beta distribution. Using this meta-model we predict SLAFEEL accuracy for $17,000$ new different parameter combinations generated with a Latin hypercube sampling (lhs), and we estimate Sobol indices using the Monte Carlo method described by Monod et al. (2006).

2.7 Total observed diversity

The *total observed diversity* quantifies the level of diversity between sequences sampled from infected hosts and accounts for intra- and inter-host diversity. It is calcu-

lated as follows:

$$D = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{S}\in\mathcal{S}}\frac{1}{|\mathbf{S}|(|\mathbf{S}|-1)}\sum_{\substack{S,S'\in\mathbf{S}\\ S\neq S'}} d(S,S') + \frac{1}{|\mathcal{S}|(|\mathcal{S}|-1)}\sum_{\substack{\mathbf{S},\mathbf{S}'\in\mathcal{S}\\ \mathbf{S}\neq\mathbf{S}'}}\frac{1}{|\mathbf{S}||\mathbf{S}'|}\sum_{(S,S')\in\mathbf{S}\times\mathbf{S}'} d(S,S'), \tag{1}$$

where the first term measures the average within-host genetic diversity and the second term measures the average between-host diversity. $d(S,S')$ is the Jukes-Cantor distance between two sequences S and S' measured by $-\frac{3}{4}\log(1-\frac{4}{3}n(S,S'))$; $n(S,S')$ is the number of different nucleotides between the two sequences S and S' (Jukes et al., 1969); $|\cdot|$ is the cardinal operator that gives the number of elements in the set under consideration; $\mathbf{S}$ (or $\mathbf{S}'$) is the set of genotypes observed from a given infected host in the sampled population; $\mathcal{S}$ is the set of all the sets of genotypes $\mathbf{S}$ observed from the population.

3 Results

3.1 Strong effect of kinetic and shuffling on SLAFEEL accuracy

The mean (over 20 replicates) of SLAFEEL accuracy computed for each of the initial 256 parameter combinations varies between 0.08 and 0.97. This range of values illustrates the significant variability in the accuracy with the input factors that we consider. The computation of Sobol indices shows that the kinetic model has the largest impact on SLAFEEL accuracy ($\mathcal{I}_{\text{tot}} = 0.78$; Figure 1). Then, one observes some influence of the shuffling parameter, the mutation rate, the sequencing depth and the transmission rate with $\mathcal{I}_{\text{tot}}$ equal to 0.21, 0.05, 0.045 and 0.035, respectively. In contrast, the variations of the size of the host population, the contact rate and the sequence fragment size, have a negligible impact on the accuracy. Interactions have a relatively small effect on the accuracy, and the principal effects represent more that 70% of the total effect for the kinetic model, the mutation rate and the sequencing depth. Note that, including the 2nd-order interactions, the meta-model explains 66% (generalized R^2) of the variability of SLAFEEL accuracy.

Figure 2 illustrates the relationship between each input factor and SLAFEEL accuracy. It materializes, beyond the Sobol indices, how the accuracy of SLAFEEL varies with diverse input factors. In particular, the reconstruction of transmissions is the poorly accurate when all the hosts have a viral kinetic corresponding to an acute infection with a delay in the production of virus ($\mathcal{K}_2$), whereas it is especially efficient

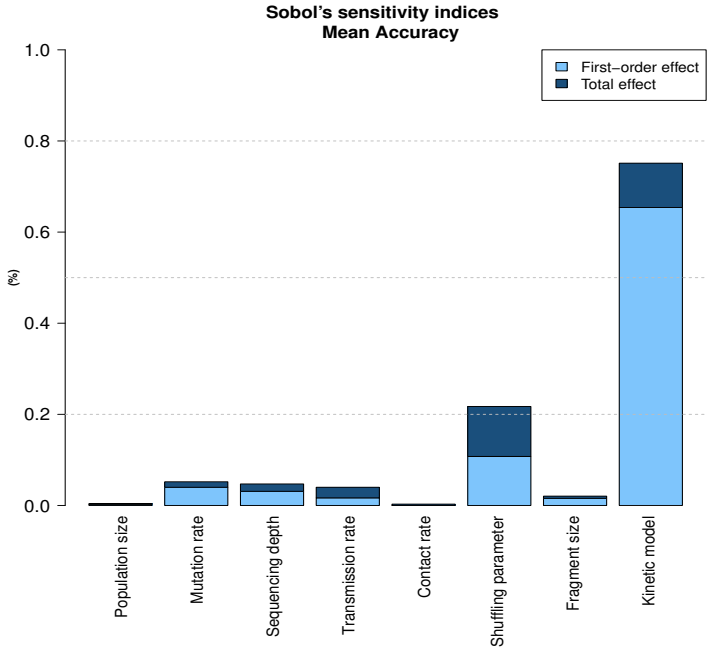


Figure 1: First-order and total Sobol indices for the seven input parameters and the kinetic model modality with respect to the mean accuracy of `SLAFEEL`.

when the kinetic includes an immune response ($\mathcal{K}_3$). Moreover, we observe a clear average increase in `SLAFEEL` accuracy when the shuffling parameter increases. More marginally, increasing the mutation rate, the sequencing depth and the sequence fragment size negatively affects the performance.

Supporting Figures S1 and S2 give the results of a sensitivity analysis performed for the standard deviation of `SLAFEEL` accuracy (i.e., the standard deviation that is observed among the 20 replicates for a given combination of parameters). Thus, one can evaluate how variations in the input factors impact the variability of `SLAFEEL` performance between replicates drawn from the same parameter combination. Based on the moderate values of Sobol indices and the moderate range of variation of the standard deviation, `SLAFEEL` performance is relatively stable across replicates. We however notice relatively large effects of the sequencing depth, the transmission rate, the shuffling parameter and the kinetic model, in particular when interactions are taken into account. In particular, the larger the sequencing depth (or the narrower the relative transmission bottleneck), the larger the variability between replicates.

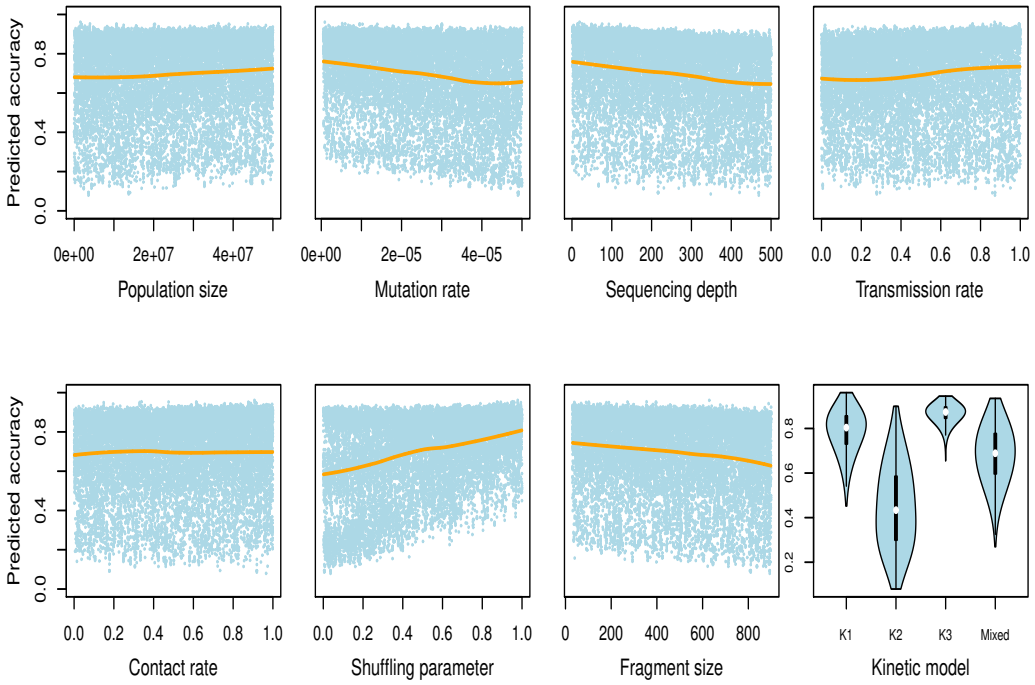


Figure 2: Relationship between input parameters and the predictions of `SLAFEEL` mean accuracy. Each point in the scatter plots represents the mean of accuracy predicted by the generalized linear meta-model. Solid orange line: local 2-degree polynomial regression (LOESS: LOcally weighted Scatterplot Smoother). Last panel: violin plot representing the effect of the kinetic model (qualitative variable) on the mean accuracy.

3.2 Identification of an optimal total observed diversity

Alamil et al. (2020) studied the link between parameters of the within-host pathogen evolution model and the within-host diversity of the pathogen. Here, we firstly explore the link between input factors considered in the sensitivity analysis (including those related to within-host pathogen evolution but also those related to host-to-

host transmission and sampling) and the observed diversity. The observed diversity is measured by the *total observed diversity* given by Equation (1). The sensitivity analysis shows that the observed diversity is especially influenced by the sequence fragment size ($\mathcal{I}_{\text{tot}} = 0.80$), the kinetic model ($\mathcal{I}_{\text{tot}} = 0.39$) and the shuffling parameter ($\mathcal{I}_{\text{tot}} = 0.11$), with a large contribution of interactions between input factors (Figure 3).

The presence of the kinetic model and the shuffling parameter among the most influential input for both the `SLAFEEL` accuracy and the observed diversity leads us to explore the relationship between the `SLAFEEL` accuracy and the observed diversity. We see a non-monotonous relationship between these variables (Figure 4), with an optimal total observed diversity (around 0.005) maximizing the average `SLAFEEL` accuracy. In addition, this optimal value correspond to a relatively low variability of `SLAFEEL` performance between replicates (Supporting Figure S3).

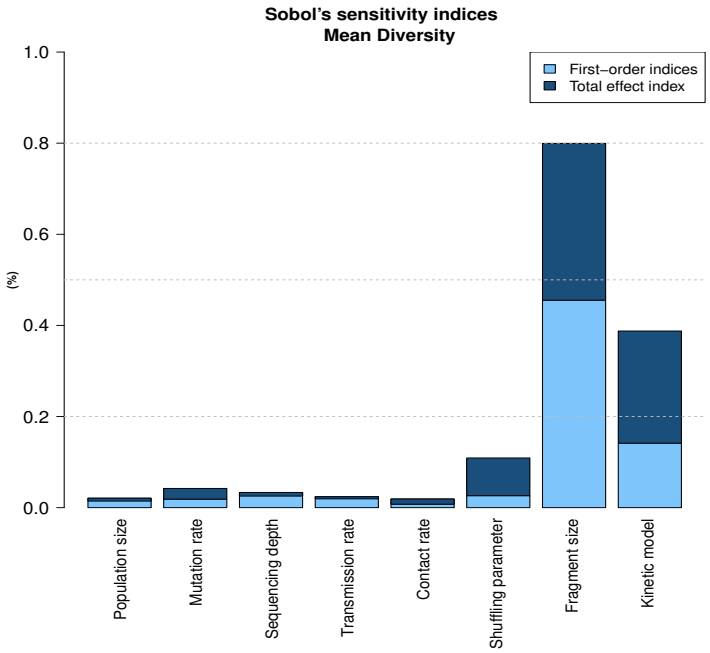


Figure 3: First-order and total Sobol indices for the seven input parameters and the kinetic model modality with respect to the mean total diversity.

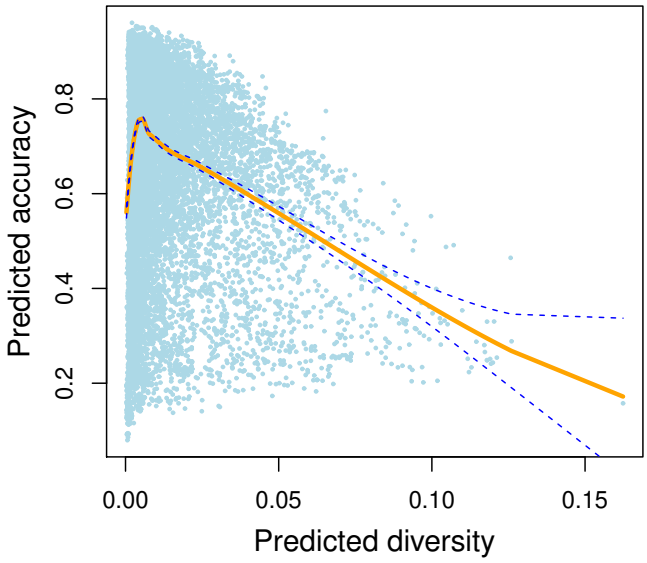


Figure 4: Predicted mean accuracy versus predicted total observed diversity. Solid orange line: LOESS smoother. Dashed blue lines: corresponding 95% pointwise prediction interval.

4 Discussion

The aim of this study was to investigate how demographic, genetic and sampling forces (related to the within-host pathogen evolution and viral load, the transmission between hosts and the observation) impact the ability of reconstructing outbreaks from pathogen sequence data informing the within-host diversity of the pathogen. We addressed this question by designing a sensitivity analysis to explore how `SLAFEEL` (statistical learning approach for estimating epidemiological links from sequence data; Alamil et al., 2019) performs with simulated data generated under various scenarios depending on demographic (kinetic model), evolutionary (mutation rate and shuffling parameter) and sampling (sequencing depth and sequenced fragment size) factors. Our results indicate that the viral kinetic model (characterizing the within-host pathogen demography) has the strongest effect on the accuracy of `SLAFEEL`: kinetics corresponding to acute infection with a delay in the production

of virus ($\mathcal{K}_2$) decrease the SLAFEEL performance, whereas kinetics including immune response ($\mathcal{K}_3$) increase it. Moreover, we observed that increasing the shuffling rate or decreasing the mutation rate increases SLAFEEL accuracy (increasing the shuffling parameter or decreasing the mutation rate reduces the number of variants but increases the number of 'variants with non-negligeable proportions' that are likely to be more easily *followed* throughout the transmission chains). These factors were all identified as influencing the within-host genetic diversity during the infection period (Alamil et al., 2020). Nevertheless, instead of observing a simple relationship between observed within-host diversity and SLAFEEL accuracy, we highlighted a non-monotonous relationship and even identified an optimal *total observed diversity* for an efficient reconstruction of transmission links.

De Maio et al. (2016) and De Maio et al. (2018) also investigated the impact of evolutionary, transmission and sequencing factors on the reconstruction of transmission links. From a methodological perspective, these investigations were carried out by assessing the performance of SCOTTI and BadTrIP with a battery of different scenarios drawn from a fixed base scenario viewed as a benchmark without taking into account interactions between evolutionary, transmission and sequencing factors, whereas we handle these interactions in the sensitivity analysis framework that we used.

Some of the results provided by De Maio et al. are consistent with our results. For instance, De Maio et al. (2016) demonstrated that reducing the efficiency of the transmission bottleneck and increasing the within-host effective population size (which both increase within-host genetic variation) lead to a decline of SCOTTI accuracy. This is consistent with our study showing that an increase in the mutation rate and a decrease in the shuffling parameter (which both lead to an augmentation of the within-host genetic diversity; Alamil et al., 2020) tend to globally reduce SLAFEEL accuracy.

We however establish other results that cast a different light on transmission reconstruction from genetic data and illustrate the complexity and challenging nature of the problem. For example, De Maio et al. 2016 show a 15% increase in SCOTTI accuracy by exploiting sequences 10 times longer than the sequences exploited in the base scenario. In contrast, we observe that longer fragment sizes tend to reduce SLAFEEL accuracy, certainly because they lead to a large number of variants including many rare variants (as large mutation rates do) and, consequently, some difficulty in *following* variants throughout the transmission chains for the considered sequencing depth. Even if SLAFEEL is constructed to detect and exploit the links between a rare variant and its close relative with non-negligible proportion (thanks to the smoothing term), we should improve it in this respect to better exploit information contained in observed rare variants (such an improvement should also certainly help in better

exploiting large sequencing depth with SLAFEEL).

Another example of discrepancy is the link between transmission bottleneck and the performance in reconstructing who infected whom. De Maio et al. (2018) stated that the stronger the transmission bottleneck, the more accurate BadTrIP. The interpretation of De Maio et al. is that less variants are shared between hosts with strong bottleneck and, hence, variants carrying specific mutations are more informative for inferring the epidemiological links. In contrast, with the experimental design that we used, SLAFEEL is more accurate with weak transmission bottleneck (i.e., τ). Thus, it seems that SLAFEEL better performs when they is a sufficiently large number of shared variants, but this property should be more precisely studied.

These similarities and dissimilarities have however to be considered with caution. Indeed, interactions between factors are not taken into account by De Maio et al. (2016, 2018), but also the demographic, evolutionary and sampling settings are quite different. Thus, De Maio et al. (2016) limited the number of SNPs (substitution) to 3-4 SNPs per sampled host, and hence limited viral diversity. De Maio et al. (2018) allowed the sampling of a maximum of 100 sequences. De Maio et al. (2016, 2018) used a constant within-host pathogen population size. We considered relatively short sequence fragments. Beyond the comparison between the different studies, the sensitivity analysis that we performed allows us to identify situations where SLAFEEL could be improved, namely and essentially, when the observed diversity is large.

As illustrated by Table 1, we vary only a limited part of the input factors in the sensitivity analysis. The impact of the other factors could be explored in further studies. In particular, a sensitivity analysis could be performed for exploring the tuning options in SLAFEEL, such as the penalization shape, the smoother, but also the training data. From this perspective, using the simulations performed in the sensitivity analysis presented in this article, we display the impact of the number of training hosts on SLAFEEL accuracy (see Supporting Figures S4 and S5). The sampling scheme was not designed to address this question and we have only a small number of cases with low numbers of training hosts. A dedicated sensitivity analysis should be carried out to confirm the impression given by Figure S4 that SLAFEEL is better with six training hosts than with two.

References

Alamil, M., K. Berthier, G. Thébaud, and S. Soubeyrand (2020). Characterizing viral within-host diversity in fast and non-equilibrium demo-genetic dynamics. *In progress*.

Alamil, M., J. Hughes, K. Berthier, C. Desbiez, G. Thébaud, and S. Soubeyrand (2019). Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. *Philosophical Transactions of the Royal Society B 374*(1775), 20180258.

Baccam, P., C. Beauchemin, C. A. Macken, F. G. Hayden, and A. S. Perelson (2006). Kinetics of influenza a virus infection in humans. *Journal of virology 80*(15), 7590–7599.

Campbell, F., A. Cori, N. Ferguson, and T. Jombart (2019). Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS computational biology 15*(3), e1006930.

Campbell, F., C. Strang, N. Ferguson, A. Cori, and T. Jombart (2018). When are pathogen genome sequences informative of transmission events? *PLoS pathogens 14*(2), e1006885.

Cottam, E. M., G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, and D. T. Haydon (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences 275*(1637), 887–895.

Daniel, C. (1973). One-at-a-time plans. *Journal of the American statistical association 68*(342), 353–360.

De Maio, N., C. J. Worby, D. J. Wilson, and N. Stoesser (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS computational biology 14*(4), e1006117.

De Maio, N., C.-H. Wu, and D. J. Wilson (2016). Scotti: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS computational biology 12*(9), e1005130.

Didelot, X., C. Fraser, J. Gardy, and C. Colijn (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution 34*(4), 997–1007.

Didelot, X., J. Gardy, and C. Colijn (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution 31*(7), 1869–1879.

Didelot, X., A. S. Walker, T. E. Peto, D. W. Crook, and D. J. Wilson (2016). Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology 14*(3), 150.

Droesbeke, J.-J., J. Fine, and G. Saporta (1997). *Plans d'expériences: applications à l'entreprise*. Editions technip.

Faivre, R., B. Iooss, S. Mahévas, D. Makowski, and H. Monod (2016). *Analyse de sensibilité et exploration de modèles: application aux sciences de la nature et de l'environnement*. Editions Quae.

Ferguson, N. M., C. A. Donnelly, and R. M. Anderson (2001). Transmission intensity and impact of control policies on the foot and mouth epidemic in great britain. *Nature 413*(6855), 542–548.

Frickel, J., P. G. Feulner, E. Karakoc, and L. Becks (2018). Population size changes and selection drive patterns of parallel evolution in a host–virus system. *Nature communications 9*(1), 1–10.

Gutiérrez, S., Y. Michalakis, and S. Blanc (2012). Virus population bottlenecks during within-host progression and host-to-host transmission. *Current opinion in virology 2*(5), 546–555.

Hang, D., E. Torng, C. Ofria, and T. M. Schmidt (2007). The effect of natural selection on the performance of maximum parsimony. *BMC Evolutionary Biology 7*(1), 94.

Jombart, T., A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol 10*(1), e1003457.

Jukes, T. H., C. R. Cantor, et al. (1969). Evolution of protein molecules. *Mammalian protein metabolism 3*(21), 132.

Klinkenberg, D., J. A. Backer, X. Didelot, C. Colijn, and J. Wallinga (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS computational biology 13*(5), e1005495.

Krzanowski, W. J. and D. J. Hand (2009). *ROC curves for continuous data*. Crc Press.

Kuo, C.-H., N. A. Moran, and H. Ochman (2009). The consequences of genetic drift for bacterial genome complexity. *Genome research 19*(8), 1450–1454.

LaBar, T. and C. Adami (2017). Evolution of drift robustness in small populations. *Nature communications 8*(1), 1–12.

Lau, M. S., G. Marion, G. Streftaris, and G. Gibson (2015). A systematic bayesian integration of epidemiological and genetic data. *PLoS computational biology 11*(11), e1004633.

Leitner, T. and E. Romero-Severson (2018). Phylogenetic patterns recover known hiv epidemiological relationships and reveal common transmission of multiple variants. *Nature microbiology 3*(9), 983–988.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz (2005). Superspreading and the effect of individual variation on disease emergence. *Nature 438*(7066), 355–359.

Mollentze, N., L. H. Nel, S. Townsend, K. Le Roux, K. Hampson, D. T. Haydon, and S. Soubeyrand (2014). A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B: Biological Sciences 281*(1782), 20133251.

Monod, H., C. Naud, and D. Makowski (2006). Uncertainty and sensitivity analysis for crop models. *Working with dynamic crop models: Evaluation, analysis, parameterization, and applications 4*, 55–100.

Morelli, M. J., G. Thébaud, J. Chadœuf, D. P. King, D. T. Haydon, and S. Soubeyrand (2012). A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol 8*(11), e1002768.

Pawelek, K. A., G. T. Huynh, M. Quinlivan, A. Cullinane, L. Rong, and A. S. Perelson (2012). Modeling within-host dynamics of influenza virus infection including immune responses. *PLoS Comput Biol 8*(6), e1002588.

Rousseau, E., B. Moury, L. Mailleret, R. Senoussi, A. Palloix, V. Simon, S. Valière, F. Grognard, and F. Fabre (2017). Estimating virus effective population size and selection without neutral markers. *PLoS pathogens 13*(11), e1006702.

Saltelli, A., K. Chan, and E. Scott (2000). *Sensitivity analysis*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Ltd.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Skums, P., A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, et al. (2018). Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics 34*(1), 163–170.

Sobol, I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki 7*(4), 784–802.

Sobol, I. M. (1976). Uniformly distributed sequences with an additional uniform property. *USSR Computational Mathematics and Mathematical Physics 16*(5), 236–242.

Soubeyrand, S. (2016). Construction of semi-Markov genetic-space-time SEIR models and inference. *Journal de la Société Française de Statistique 157*, 129–152.

Spada, E., L. Sagliocca, J. Sourdis, A. R. Garbuglia, V. Poggi, C. De Fusco, and A. Mele (2004). Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis c virus infection. *Journal of clinical microbiology 42*(9), 4230–4236.

Wallinga, J. and P. Teunis (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology 160*(6), 509–516.

Worby, C. J., M. Lipsitch, and W. P. Hanage (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput Biol 10*(3), e1003549.

Worby, C. J., P. D. O'Neill, T. Kypraios, J. V. Robotham, D. De Angelis, E. J. Cartwright, S. J. Peacock, and B. S. Cooper (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics 10*(1), 395.

Worby, C. J. and T. D. Read (2015). 'seedy'(simulation of evolutionary and epidemiological dynamics): An r package to follow accumulation of within-host mutation in pathogens. *PloS one 10*(6), e0129745.

Wymant, C., M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, and C. Fraser (2018). Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular biology and evolution 35*(3), 719–733.

4.4. Key points of Chapter 4

- ☞ Interactions between demographic and genetic processes within a host during an outbreak affect directly the within-host viral composition leaving uncertainty around the reconstruction of transmissions within outbreaks.
- ☞ An important step towards the accurate inference of epidemiological links was to exploit the within-host genetic data and examine the impact of the demo-genetic factors on the ability of our approach (presented in chapter 3) to infer epidemiological links.
- ☞ To obtain a more accurate estimation of transmission pathways from pathogen sequence data, we proposed to: exploit genetic samples with potentially multiple variants from each host, assess the genetic diversity within and between hosts before attempting the transmission inference and combine adequate temporal and contact data with the genetic data.

Chapter 5

Conclusion and perspectives

5.1. General conclusion

The main objective of this thesis was to develop an accurate, robust and rapid statistical approach adapted for inferring epidemiological links of infectious disease caused by fast-evolving pathogens, mainly viruses, from deep sequencing data reflecting the within-host diversity of the pathogen. To achieve this goal we have attempted to answer the questions addressed in Section 1.2 as follows.

5.1.1. Introducing a statistical learning approach for inferring epidemiological links

We proposed an exploratory statistical approach, called `SLAFEEL`, investigating epidemiological links between host units from deep sequencing data; hosts can designate, for example, individuals, households, farm premises and agricultural fields. `SLAFEEL` is based on a pseudo-mechanistic evolutionary model and grounded on statistical learning (Friedman et al., 2001; James et al., 2013), and was conceived as a rather weakly-demanding tool in terms of computer resources to offer the opportunity, in the future, to tackle big data sets with respect to host number, sequencing depth and sequence size (Pfeiffer and Stevens, 2015; Vatsavai et al., 2012; Ziegler and König, 2014). The pseudo-evolutionary model is inspired by (but does not stick to) a mechanistic vision of transmission and micro-evolutionary processes underlined by some assumptions that are expected to guide transmission inference. A statistical learning approach is proposed to learn

the similarities and differences in within-host genetic diversity that are expected between a recipient and its source, and, thus, intrinsically handle sequencing errors and avoid to rely on eventually misleading mechanistic assumptions. To sum up, the mechanistic assumptions are replaced by training data for constraining the inference of transmission links. The training data can consist of either classical contact information such as contact tracing, contact information based on a preliminary analysis applied to a subset of the genetic data, or proxies of contact information such as geographical distances between host units (see Chapter 3), social connections (Keeling and Eames, 2005; Skums et al., 2018b) and geographical connectivity via air masses for airbone pathogens (Choufany et al., 2019; Leyronas et al., 2018). Overall, the concept underlying `SLAFEEL` consists of learning the structure of the epidemiological links with a pseudo-evolutionary model applied to training data, and then to use this initial training stage for the inference of the links for the whole data set. In that respect, one can assume that whatever the way that training data are exploited, the closer the relationship between contact information and inferred transmission links, the more informative the training stage. The ability to use multiple types of contact information makes this approach a versatile tool to estimate transmission links in different contexts exploiting different data from animal, human and plant epidemics.

5.1.2. Evaluating how robust, effective and versatile this approach is

To test the versatility of `SLAFEEL`, we first applied it to investigate virus dynamics in human, animal and plant populations, by exploiting existing data sets obtained with different sequencing techniques, reflecting different levels of within- and between-hosts pathogen diversity, containing different levels of epidemiological data, and from which different types of training data could be drawn. This approach enabled us to estimate *who infected whom* in animal and human epidemics concerning respectively influenza A viruses sampled from animal populations and Ebola virus sampled from a human population. In the animal epidemic case study, transmissions were inferred for two experimental outbreaks in pigs (first analyzed by Murcia et al., 2012) with different immunological histories (naive and vaccinated) and with partially known chains of transmissions. We exploited a part of the partial knowledge of the transmission chains to draw contact information affected by some uncertainty (i.e., recipient hosts in the training data set generally had two possible sources); And we exploited the remaining part of the

partial knowledge of the transmission chains as validation data used to assess SLAFEEL performance. Thus, this data set had been used as a test for SLAFEEL, like simulated data. In the Ebola case study, whose data were previously published and analyzed by Gire et al. (2014), transmission links were inferred between patients originating from several chiefdoms. The inference was performed by calibrating the model with contact tracing published in a previous study (Senga et al., 2017). This case corresponds to a challenging situation since little pathogen diversity was observed and limited contact tracing information was used. In the case of a plant disease, namely a potyvirus of the wild salsify (Piry et al., 2017), we have interpreted the links between patches in a conservative way by inferring *who is closely related to whom* instead of *who infected whom*. We calibrated the penalization using geographical distances between patches assuming that close patches are *a priori* more likely infected by similar virus populations. A fourth data set was used to explore new specifications of SLAFEEL related to temporal constraints for the prior selection of putative sources for each recipient host and related to the penalization arising in the penalized pseudo-likelihood. For this data set, dealing with an equine influenza outbreak previously analyzed by Newton et al. (2006) and Hughes et al. (2012), we used training recipient-source pairs inferred from a prior analysis of a subset of genetic and epidemiological data and then we applied SLAFEEL to the whole data set. We also illustrated how the epidemiological temporal constraints can be varied. For each of these case studies, the results show a relatively large consistency between the inferred links and the contact information while dealing with diverse epidemiological situations, sequencing techniques and pathogen diversity levels. Ideally, this consistency must be measured by assessing how much inferred links and contact information match with cross-validation applied to training data.

A sensitivity analysis was then designed to evaluate the performance of SLAFEEL under multiple genetic, demographic and sequencing conditions related to the within-host pathogen evolution, inter-host transmission and observation. This analysis was performed by applying SLAFEEL to data simulated with an original model (grounded on the model proposed by Worby and Read, 2015) generating, in a forward approach, the transmission and the evolution of the virus. It has enabled us to identify that the most influential factors impacting the reconstruction of outbreaks through SLAFEEL are among those influencing the within-host genetic diversity during the infection period. This prompted us to examine to what extent the observed pathogen genetic diversity affects the performance of SLAFEEL. In other words, can we predict SLAFEEL accuracy given

the observed level of within- and between-hosts pathogen diversity. Instead of observing a simple monotonous relationship between SLAFEEL accuracy the *total observed genetic diversity*, we obtained a non-monotonous relationship and identified an optimal *total observed diversity* for efficiently reconstructing transmission links.

5.1.3. Generating and characterizing the within-host genetic diversity

Simulated data used to perform the sensitivity analysis were generated with an original model for the evolution and transmission of populations of pathogen sequences, under various demo-genetic and sampling assumptions. This model, inspired by the work of Worby and Read (2015), includes a within-host model allowing the generation of fast non-equilibrium demo-genetic dynamics and providing as output the evolution of genotypes and their frequencies across time. Demographic effects are considered mainly through the kinetic model quantifying the temporal variation in the viral load. This component of the model can be built from numerous viral kinetic models encountered in the literature (e.g., see Beauchemin and Handel, 2011). Genetic effects are considered through mutation and replication modes approximately mimicking natural selection and genetic drift. These modes are based, in particular, on the elimination of lethal genomes and the shuffling of genotype proportions generating over-dispersion with respect to multinomial draws. We characterized the behaviour of this model, in particular with respect to its ability to generate within-host genetic diversity, and highlighted the demographic and genetic factors shaping this diversity. This characterization was performed numerically with several classical diversity indices. Accounting for the temporal variation in the viral load, we observed a non-monotonous relationship between the within-host pathogen population size and the genetic diversity. This is in contradiction with some previous studies (Bailey et al., 2014; Golubchik et al., 2013; Nelson and Hughes, 2015) based on the estimation of the within-host effective population size and showing a positive correlation between the population size and the genetic diversity. Moreover, like other studies (Didelot et al., 2016; Elena et al., 2007; Kuo et al., 2009; LaBar and Adami, 2017), we pointed out the effect of the interaction between genetic and demographic forces (mainly between the viral kinetics and the forces causing the fast variation in genotype proportions) on the shape of the within-host genetic diversity.

5.2. Perspectives

My work carried out to design a robust, effective and versatile learning approach to infer transmission links has been based on hypotheses that have been viewed as relevant along the process of construction of the method. However, the diverse applications of SLAFEEL show that this is a promising *all-terrain* approach (as soon as contact information can be used as training data), but also highlight some situations where the tested versions of the method are not very effective (and therefore lack robustness). This section presents several ways for improving SLAFEEL as well as the model proposed to simulate outbreaks.

5.2.1. Improving SLAFEEL

The background idea of SLAFEEL consists of ***learning*** the structure of epidemiological links with a ***pseudo-evolutionary model*** calibrated with ***training data***, and using this initial stage to infer links for the whole ***dataset***. Based on this, an improvement plan could be considered by acting on three axes: the modeling (i.e., improving the construction of the pseudo-evolutionary model), the statistical learning, and the data.

Improving the modeling

The improvement of the penalized pseudo-evolutionary model essentially consist of modifying the pseudo-evolutionary model and the penalization shape.

In my thesis, to apply SLAFEEL to real and simulated data, I used a simple semi-parametric model (called pseudo-evolutionary model) describing the transition from an explanatory set of sequences observed from a putative source to a response set of sequences observed from the infected host. We considered that each sequence observed from the infected host is only explained by the closest sequence(s) observed in the putative source with respect to the number of different nucleotide. This specification avoided the possibility of multiple infections per host (several infectious hosts infect one susceptible host). Actually, SLAFEEL could relatively simply allow multiple infections by considering that the explanatory set of sequences can be made of several sets of sequences collected from different putative sources that can eventually be weighted according to some factors (temporal, spatial, environmental, etc).

Moreover, we use a kernel smoother for defining the pseudo-evolutionary model. Beyond the modification of the kernel shape, we could consider other semi-

parametric or non-parametric regression tools that could offer a larger flexibility, such as neuron networks (Beręsewicz et al., 2018; Buelens et al., 2015) and random-forest models (Tatem et al., 2014; Ziegler and König, 2014). The use of such regression models could however hamper to some extent the interpretability of the approach (indeed, the parameter of the kernel can be interpreted as a substitution rate in the current version of SLAFEEL).

Applying SLAFEEL to real cases in epidemiology, we introduced several penalization shapes corresponding to three different hypotheses, but the penalization could be adapted according to the available data by considering other hypotheses. Thus, we could propose a library of penalization shapes circumventing current SLAFEEL limitations. For instance, we can develop some penalization functions taking into account available temporal data, extending the proposal made for analyzing the equine influenza virus data set. We could also design some penalization functions forcing, to some extent, the algorithm to pair hosts sharing virus variants with specific genetic signals such as codon STOP (see Hughes et al., 2012, for a discussion about the same codon STOP observed from different hosts in the equine influenza virus data set).

Improving the statistical learning

For accurately inferring epidemiological links between hosts when applying SLAFEEL, we should adopt an adequate penalization shape such that the learning (or the calibration) of the penalization parameter from training data makes sense and is useful. Beyond the penalization parameter, we also suggested that the penalization shape but also the tolerance parameter (introduced in Chapter 3) can be learned from training data. This could be carried out as it is done for the penalization parameter, i.e., by estimating the sources for all the possible combinations 'penalization shape × penalization parameter × tolerance value' and retaining the combination(s) leading to the largest consistency between estimated sources and traced sources. Cross-validation, mentioned in Section 5.1.2 to measure this consistency, also deserves to be considered for tuning some components of SLAFEEL, e.g., the penalization shape, the tolerance value but also the temporal constraints for *a priori* selecting the putative sources for each recipient host. Whatever the solution that is chosen for calibrating the tuning parameters and functions, training data have to be sufficiently large to avoid overfitting or identifiability issues. In addition, one has to make the balance between tuning more and more components of SLAFEEL and the computational resources that this could require.

Improving the use of data

For limiting the computational burden, the pseudo-evolutionary model was designed as a regression function concisely describing transitions between two sets of sequences sampled at different times from a recipient host and its putative source. To implement this approach, we handle the row sequences. Alternatively, we could first transform sequence data using a technique of dimension reduction (Lareo and Acevedo, 1999; Pelé et al., 2012), i.e., projecting sequence data into a low-dimension space. This can be made, for example, with the multidimensional scaling of the pairwise distance matrix between sequences (Pelé et al., 2012). Projecting sequences could facilitate the scaling of SLAFEEL to big data.

At the learning stage, SLAFEEL requires a set of training data allowing to calibrate the penalization parameter (and other eventual tuning parameters and functions as proposed above). In Chapter 4, where we carried out the sensitivity analysis to determine the key factors impacting SLAFEEL accuracy, we briefly took a look at the performance of SLAFEEL with respect to the number of training recipient-source pairs. This factor was not explicitly included in the sensitivity analysis and, therefore, we could not draw a robust conclusion about it. However, it seems that SLAFEEL was more accurate with a larger number of training pairs. This preliminary result could be further explored with a dedicated sensitivity analysis. In practice, there may be training pairs resulting from accurate contact tracing for example, but also uncertain training pairs grounded on inaccurate information. In such a situation, the issue is not to consider as many training pairs as possible, but to make the balance between the number of training pairs that are used (favoring the most certain first), and the accuracy of the reconstruction of transmissions.

In the sensitivity analysis, we have seen that there may be an optimal level of total observed diversity enabling SLAFEEL to be relatively accurate. Hence, we have an indication about the type of sequencing data that can be adequate. However, in this sensitivity analysis, we only considered a subset of the model parameters as input factors. We could consider all the other parameters but also consider how variations in the knowledge of epidemiological data informing on the timing of the infection relate with SLAFEEL accuracy. Such an extended sensitivity analysis could give some indications about the relative importance of genetic data and epidemiological data in the reconstruction of transmission links.

5.2.2. Possible extensions of SLAFEEL

One of the extensions that could be considered consists of improving the way epidemiological data and genetic data enable the transmission and evolutionary processes to inform each other. In this thesis, epidemiological data arise in the *a priori* selection of putative source and in the penalized pseudo-likelihood, but are not actually included through a mechanistic framework. Instead, we could couple the pseudo-evolutionary model with a space-time SEIR-like (Susceptible-Exposed-Infectious-Removed) compartmental model. Like several approaches developed to estimate the transmission links based on a SEIR model and a micro-evolutionary model (Jombart et al., 2014; Mollentze et al., 2014; Morelli et al., 2012; Soubeyrand, 2016; Ypma et al., 2012, 2013a), we could couple the (penalized) pseudo-likelihood of SLAFEEL (described in Chapter 3) with a *temporal* likelihood, for example, based on sampling times (Jombart et al., 2014) or on incubation periods (Mollentze et al., 2014; Morelli et al., 2012; Soubeyrand, 2016). Moreover, when spatial information are available, we could also couple the pseudo-likelihood of SLAFEEL with a *spatial* likelihood derived from a diffusion kernel (Mollentze et al., 2014; Morelli et al., 2012; Soubeyrand, 2016). From a simple viewpoint, by making as if temporal, spatial and genetic data are independent, we could define a global likelihood as the product between the pseudo-likelihood of SLAFEEL, a temporal likeliood and a spatial likelihood like Ypma et al. (2012) did.

Another extension that deserves to be explored would offer the possibility to distinguish direct and indirect transmissions in the inference of transmission links. If the pseudo-evolutionary model of SLAFEEL is coupled with a SEIR model, the intermediate hosts could be implicitly handled within the joint model as proposed by Jombart et al. (2014), who introduce the probabilities for a recipient host to be separated from its (possibly indirect) source in the transmission chain by zero intermediate host, one intermediate host, two intermediate hosts, and so on. Intermediate hosts could also be inferred in a post-analysis, like in Mollentze et al. (2014), by identifying the inferred recipient-source pairs for which one get some forms of discrepancy in terms of timing and genetics.

5.2.3. Possible extensions of the outbreak simulator

There are several directions in which the simulation model could be extended in the future. At the within-host level, we can easily incorporate more advanced kinetic models governing the number of virions in the presence of: treatment with therapeutic antiviral agents (Beauchemin et al., 2008; Smith and Perelson,

2011), variation of virion infectivity over time (Beauchemin and Handel, 2011; Vaidya et al., 2010), decay of viral infectivity (Smith and Ribeiro, 2010), co-receptor switch (Alizon and Boldin, 2010) and virion loss due to cell entry (Beauchemin et al., 2008). In addition, it would be possible to use more realistic and complex mutation scenarios (Kimura, 2020; Tavaré, 1986), and to include relative fitness depending on the genetic sequence or on the frequency, which induces frequency-dependent selection (Alizon and Boldin, 2010; Sanjuán et al., 2004). We have however to say that, with the current code implemented with the `R` statistical software, the computation time and the memory load required by the within-host evolutionary model may be extremely large when both the number of virions and the number of genotypes are very high. Hence, an essential improvement of the model at the within-host level is the optimization of the simulation code.

The between-host transmission model was simulated with a simple SI dynamics. Obviously, more complex dynamics are of interest, especially if in the inference of transmissions one wants to take into account, for instance, information about the beginning of the infectious period. Thus, we could incorporate a recovery stage (SIR model), a return to the susceptible stage after recovery (SIRS) or a non-infectious (called exposed) stage (SEIR model) (Brauer, 2008). In the current version of the model, we suppose that each host can be infected once by a single host and that the transmission links are determined between hosts characterized by a single sample of pathogen sequences. We could easily allow for multiple infections per host as well as the possibility to exploit multiple pathogen sequence samples per host.

All these extensions can be implemented in a more or less easy way, and could help in evaluating the robustness of `SLAFEEL` and other methods for transmission reconstruction based on genetic and/or epidemiological data. The possibilities for drawing new demo-genetic scenarios are infinite (or nearly infinite!). Thus, it would be relevant to characterize the scenarios using measures such as the one that we identified in the sensitivity analysis (namely, the total observed diversity), which could be used to classify the diverse scenarios into a reduced number of clusters that would be informative about the expected efficiency of the method for reconstructing transmissions that is used. Other relevant measures could certainly be based on both observed epidemiological and genetic data.

Bibliography

H. Abbey. An examination of the reed-frost theory of epidemics. *Human biology*, 24(3):201, 1952.

S. Abel, P. A. zur Wiesch, B. M. Davis, and M. K. Waldor. Analysis of bottlenecks in experimental models of infection. *PLoS pathogens*, 11(6):e1004823, 2015.

M. Alamil, J. Hughes, K. Berthier, C. Desbiez, G. Thébaud, and S. Soubeyrand. Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. *Philosophical Transactions of the Royal Society B*, 374(1775):20180258, 2019.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Introduction to pathogens. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

W. C. Albrich and S. Harbarth. Health-care workers: source, vector, or victim of mrsa? *The Lancet infectious diseases*, 8(5):289–301, 2008.

M. Aldrin, T. Lyngstad, A. Kristoffersen, B. Storvik, Ø. Borgan, and P. Jansen. Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. *Journal of The Royal Society Interface*, 8(62): 1346–1356, 2011.

S. Alizon and B. Boldin. Within-host viral evolution in a heterogeneous environment: insights into the hiv co-receptor switch. *Journal of evolutionary biology*, 23(12):2625–2635, 2010.

S. Alizon, F. Luciani, and R. R. Regoes. Epidemiological and clinical consequences of within-host evolution. *Trends in microbiology*, 19(1):24–32, 2011.

H. K. Allen, U. Y. Levine, T. Looft, M. Bandrick, and T. A. Casey. Treatment, promotion, commotion: antibiotic alternatives in food-producing animals. *Trends in microbiology*, 21(3):114–119, 2013.

L. J. Allen. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(2):128–142, 2017.

R. Anderson and R. May. Population biology of infectious diseases springer verlag. *New York*, 1982.

R. Anderson, C. Donnelly, N. Ferguson, M. Woolhouse, C. Watt, H. Udy, S. MaWhinney, S. Dunstan, T. Southwood, J. Wilesmith, et al. Transmission dynamics and epidemiology of bse in british cattle. *Nature*, 382(6594):779–788, 1996.

R. M. Anderson, B. Anderson, and R. M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media, 2012.

V. Andreasen. Dynamics of annual influenza a epidemics with immuno-selection. *Journal of mathematical biology*, 46(6):504–536, 2003.

V. Andreasen, J. Lin, and S. A. Levin. The dynamics of cocirculating influenza strains conferring partial cross-immunity. *Journal of mathematical biology*, 35 (7):825–842, 1997.

A. Assiri, A. McGeer, T. M. Perl, C. S. Price, A. A. Al Rabeeah, D. A. Cummings, Z. N. Alabdullatif, M. Assad, A. Almulhim, H. Makhdoom, et al. Hospital outbreak of middle east respiratory syndrome coronavirus. *New England Journal of Medicine*, 369(5):407–416, 2013.

P. Baccam, C. Beauchemin, C. A. Macken, F. G. Hayden, and A. S. Perelson. Kinetics of influenza a virus infection in humans. *Journal of virology*, 80(15):7590–7599, 2006.

A. L. Bailey, M. Lauck, A. Weiler, S. D. Sibley, J. M. Dinis, Z. Bergman, C. W. Nelson, M. Correll, M. Gleicher, D. Hyeroba, et al. High genetic diversity and adaptive potential of two simian hemorrhagic fever viruses in a wild primate population. *PloS one*, 9(3):e90714, 2014.

N. T. Bailey. *The elements of stochastic processes with applications to the natural sciences*, volume 25. John Wiley & Sons, 1990.

N. T. Bailey et al. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

N. T. Bailey et al. *The biomathematics of malaria*. Charles Griffin & Company Ltd., 1982.

D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

F. Ball and P. Neal. Network epidemic models with two levels of mixing. *Mathematical biosciences*, 212(1):69–87, 2008.

F. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *The Annals of Applied Probability*, pages 46–89, 1997.

F. Ball, D. Sirl, and P. Trapman. Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Advances in Applied Probability*, 41(3):765–796, 2009.

A. Barbour and D. Mollison. Epidemics and random graphs in: Stochastic processes in epidemic theory, eds. jp gabriel, c lefèvre, p picard, pag. 86, 1990.

M. Bartlett. Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):211–229, 1949.

M. S. Bartlett. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 4, page 109, 1956.

C. A. Beauchemin and A. Handel. A review of mathematical models of influenza a infections within a host or cell culture: lessons learned and challenges ahead. *BMC public health*, 11(1):S7, 2011.

C. A. Beauchemin, J. J. McSharry, G. L. Drusano, J. T. Nguyen, G. T. Went, R. M. Ribeiro, and A. S. Perelson. Modeling amantadine treatment of influenza a virus in vitro. *Journal of theoretical biology*, 254(2):439–451, 2008.

N. Becker. The uses of epidemic models. *Biometrics*, pages 295–305, 1979.

N. G. Becker. *Analysis of infectious disease data*, volume 33. CRC Press, 1989.

R. Beckley, C. Weatherspoon, M. Alexander, M. Chandler, A. Johnson, and G. S. Bhatt. Modeling epidemics with differential equation, 2013.

M. Beręsewicz, R. Lehtonen, F. Reis, L. Di Consiglio, and M. Karlberg. An overview of methods for treating selectivity in big data sources. Technical report, Eurostat Statistical Working Paper. Doi: https://doi. org/10.2785/312232, 2018.

D. Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, pages 1–45, 1760.

V. Bloomfield. *Computer simulation and data analysis in molecular biology and biophysics: an introduction using R*. Springer Science & Business Media, 2009.

S. P. Blythe and C. Castillo-Chavez. Like-with-like preference and sexual mixing models. *Mathematical biosciences*, 96(2):221–238, 1989.

F. Brauer. Compartmental models in epidemiology. In *Mathematical epidemiology*, pages 19–79. Springer, 2008.

F. Brauer. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, 2(2):113–127, 2017.

T. Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225 (1):24–35, 2010.

T. Britton, T. House, A. L. Lloyd, D. Mollison, S. Riley, and P. Trapman. Five challenges for stochastic epidemic models involving global transmission. *Epidemics*, 10:54–57, 2015.

J. Bryan. It may not be perfect, but the influenza vaccine has saved many people's lives. *Evaluation*, 14(47):19, 2020.

W. Budd. *Typhoid fever: its nature, mode of spreading, and prevention*. Longmans, Green, 1873.

B. Buelens, J. Burger, and J. van den Brakel. *Predictive inference for non-probability samples: a simulation study*, volume 13. Statistics Netherlands The Hague, 2015.

D. S. Burke. Recombination in hiv: an important viral evolutionary strategy. *Emerging infectious diseases*, 3(3):253, 1997.

M. Cambra, N. Capote, A. Myrta, and G. Llácer. Plum pox virus and the estimated costs associated with sharka disease. *EPPO Bulletin*, 36(2):202–204, 2006.

F. Campbell, X. Didelot, R. Fitzjohn, N. Ferguson, A. Cori, and T. Jombart. outbreaker2: a modular platform for outbreak reconstruction. *Bmc Bioinformatics*, 19(11):1–8, 2018.

F. Campbell, A. Cori, N. Ferguson, and T. Jombart. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS computational biology*, 15(3):e1006930, 2019.

L. Canini and A. S. Perelson. Viral kinetic modeling: state of the art. *Journal of pharmacokinetics and pharmacodynamics*, 41(5):431–443, 2014.

M. Casewell, C. Friis, E. Marco, P. McMullin, and I. Phillips. The european ban on growth-promoting antibiotics and emerging consequences for human and animal health. *Journal of antimicrobial chemotherapy*, 52(2):159–161, 2003.

C. Castillo-Chavez. *Mathematical and statistical approaches to AIDS epidemiology*, volume 83. Springer Science & Business Media, 2013.

C. Castillo-Chavez, W. Huang, and J. Li. Competitive exclusion in gonorrhea models and other sexually transmitted diseases. *SIAM Journal on Applied Mathematics*, 56(2):494–508, 1996.

S. Cauchemez and N. M. Ferguson. Methods to infer transmission risk factors in complex outbreak data. *Journal of The Royal Society Interface*, 9(68):456–469, 2012.

S. Cauchemez, P.-Y. Boëlle, C. A. Donnelly, N. M. Ferguson, G. Thomas, G. M. Leung, A. J. Hedley, R. M. Anderson, and A.-J. Valleron. Real-time estimates in early detection of sars. *Emerging infectious diseases*, 12(1):110, 2006.

S. Cauchemez, A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. M. Ferguson, D. Swerdlow, S. V. Sodha, M. E. Moll, F. J. Angulo, et al. Role of social networks in shaping disease transmission during a community outbreak of 2009 h1n1 pandemic influenza. *Proceedings of the National Academy of Sciences*, 108 (7):2825–2830, 2011.

S. Cauchemez, P. Nouvellet, A. Cori, T. Jombart, T. Garske, H. Clapham, S. Moore, H. L. Mills, H. Salje, C. Collins, et al. Unraveling the drivers of mers-cov transmission. *Proceedings of the National Academy of Sciences*, 113(32):9081–9086, 2016.

D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini Jr. Flute, a publicly available stochastic influenza epidemic simulation model. *PLoS Comput Biol*, 6 (1):e1000656, 2010.

M. Choufany, D. Martinetti, R. Senoussi, C. E. Morris, and S. Soubeyrand. Spatiotemporal large-scale networks shaped by air mass movements. *arXiv preprint arXiv:1911.07007*, 2019.

J. Y. Chun, G. Baek, and Y. Kim. Transmission onset distribution of covid-19 in south korea. *medRxiv*, 2020.

A. Cori, C. A. Donnelly, I. Dorigatti, N. M. Ferguson, C. Fraser, T. Garske, T. Jombart, G. Nedjati-Gilani, P. Nouvellet, S. Riley, et al. Key data for outbreak evaluation: building on the ebola experience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1721):20160371, 2017.

E. M. Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, and D. T. Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society of London B: Biological Sciences*, 275(1637):887–895, 2008.

F. W. Crawford, L. S. T. Ho, and M. A. Suchard. Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10 (2):e1423, 2018.

J. M. Cuevas, R. Geller, R. Garijo, J. López-Aldeguer, and R. Sanjuán. Extremely high mutation rate of hiv-1 in vivo. *PLoS biology*, 13(9):e1002251, 2015.

D. J. Daley and J. Gani. *Epidemic modelling: an introduction*, volume 15. Cambridge University Press, 2001.

M. Das, P. L. Chu, G.-M. Santos, S. Scheer, E. Vittinghoff, W. McFarland, and G. N. Colfax. Decreases in community viral load are accompanied by reductions in new hiv infections in san francisco. *PloS one*, 5(6):e11068, 2010.

N. De Maio, C.-H. Wu, and D. J. Wilson. Scotti: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS computational biology*, 12(9):e1005130, 2016.

N. De Maio, C. J. Worby, D. J. Wilson, and N. Stoesser. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS computational biology*, 14(4):e1006117, 2018.

C. Desbiez, A. Schoeny, B. Maisonneuve, K. Berthier, I. Bornard, C. Chandeysson, F. Fabre, G. Girardot, P. Gognalons, H. Lecoq, et al. Molecular and biological characterization of two potyviruses infecting lettuce in southeastern france. *Plant pathology*, 66(6):970–979, 2017.

X. Didelot, J. Gardy, and C. Colijn. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution*, 31(7):1869–1879, 2014.

X. Didelot, A. S. Walker, T. E. Peto, D. W. Crook, and D. J. Wilson. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology*, 14(3):150, 2016.

X. Didelot, C. Fraser, J. Gardy, and C. Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution*, 34(4):997–1007, 2017.

O. Diekmann and J. Heesterbeek. Wiley series in mathematical and computational biology. mathematical epidemiology of infectious diseases: model building, analysis and interpretation, 2000.

O. Diekmann, J. A. P. Heesterbeek, and J. A. Metz. The legacy of kermack and mckendrick. *Publications of the Newton Institute*, 5:95–115, 1995.

K. Dietz. Epidemics and rumours: A survey. *Journal of the Royal Statistical Society: Series A (General)*, 130(4):505–528, 1967.

K. Dietz. The incidence of infectious diseases under the influence of seasonal fluctuations. In *Mathematical models in medicine*, pages 1–15. Springer, 1976.

K. Dietz. Density-dependence in parasite transmission dynamics. *Parasitology today*, 4(4):91–97, 1988.

K. Dietz and J. Heesterbeek. Daniel bernoulli's epidemiological model revisited. *Mathematical biosciences*, 180(1-2):1–21, 2002.

K. Dietz and D. Schenzle. Mathematical models for infectious disease statistics. In *A celebration of statistics*, pages 167–204. Springer, 1985.

S. F. Elena, C. O. Wilke, C. Ofria, and R. E. Lenski. Effects of population size and mutation rate on the evolution of mutational robustness. *Evolution*, 61(3): 666–674, 2007.

D. W. Eyre, M. L. Cule, D. J. Wilson, D. Griffiths, A. Vaughan, L. O'Connor, C. L. Ip, T. Golubchik, E. M. Batty, J. M. Finney, et al. Diverse sources of c. difficile infection identified on whole-genome sequencing. *N Engl J Med*, 369:1195–1205, 2013.

W. Farr. Progress of epidemics. *Second report of the Registrar General of England and Wales*, pages 16–20, 1840.

C. P. Farrington, M. N. Kanaan, and N. J. Gay. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):251–292, 2001.

A. S. Fauci. Emerging and re-emerging infectious diseases: influenza as a prototype of the host-pathogen balancing act. *Cell*, 124(4):665–670, 2006.

O. Faye, P.-Y. Boëlle, E. Heleze, O. Faye, C. Loucoubar, N. Magassouba, B. Soropogui, S. Keita, T. Gakou, L. Koivogui, et al. Chains of transmission and control of ebola virus disease in conakry, guinea, in 2014: an observational study. *The Lancet Infectious Diseases*, 15(3):320–326, 2015.

W. Feller. *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons, 2008.

J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

E. P. Fenichel, C. Castillo-Chavez, M. G. Ceddia, G. Chowell, P. A. G. Parra, G. J. Hickling, G. Holloway, R. Horan, B. Morin, C. Perrings, et al. Adaptive human behavior in epidemiological models. *Proceedings of the National Academy of Sciences*, 108(15):6306–6311, 2011.

N. M. Ferguson, C. A. Donnelly, and R. M. Anderson. Transmission intensity and impact of control policies on the foot and mouth epidemic in great britain. *Nature*, 413(6855):542–548, 2001.

B. Foxman and L. Riley. Molecular epidemiology: focus on infection. *American journal of epidemiology*, 153(12):1135–1141, 2001.

S. A. Frank. Coevolutionary genetics of hosts and parasites with quantitative inheritance. *Evolutionary ecology*, 8(1):74–94, 1994.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

T. Frische, S. Egerer, S. Matezki, C. Pickl, and J. Wogram. 5-point programme for sustainable plant protection. *Environmental Sciences Europe*, 30(1):8, 2018.

F. Galton. *Natural inheritance*. Macmillan and Company, 1894.

A. Gavryushkina, D. Welch, T. Stadler, and A. J. Drummond. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput Biol*, 10(12):e1003919, 2014.

M. A. Gilchrist and A. Sasaki. Modeling host-parasite coevolution: a nested approach based on mechanistic models. *Journal of Theoretical Biology*, 218(3): 289–308, 2002.

S. K. Gire, A. Goba, K. G. Andersen, R. S. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, et al. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *science*, page 1259657, 2014.

T. Golubchik, E. M. Batty, R. R. Miller, H. Farr, B. C. Young, H. Larner-Svensson, R. Fung, H. Godwin, K. Knox, A. Votintseva, et al. Within-host evolution of staphylococcus aureus during asymptomatic carriage. *PLoS One*, 8(5):e61319, 2013.

N. C. Grassly and C. Fraser. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, 6(6):477–487, 2008.

J. Graunt. Natural and political observations made upon the bills of mortality (1662). *Ed., WF Willcox. Baltimore*, 1939.

D. Greenhalgh and I. Moneim. Sirs epidemic model and simulations using different types of seasonal contact rate. *Systems Analysis Modelling Simulation*, 43 (5):573–600, 2003.

P. E. Greenwood and L. F. Gordillo. Stochastic epidemic modeling. In *Mathematical and statistical estimation approaches in epidemiology*, pages 31–52. Springer, 2009.

J. J. Grefenstette, S. T. Brown, R. Rosenfeld, J. DePasse, N. T. Stone, P. C. Cooley, W. D. Wheaton, A. Fyshe, D. D. Galloway, A. Sriram, et al. Fred (a framework for reconstructing epidemic dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC public health*, 13(1):1–14, 2013.

B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *science*, 303(5656):327–332, 2004.

S. Gubbins and C. A. Gilligan. Invasion thresholds for fungicide resistance: deterministic and stochastic analyses. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1437):2539–2549, 1999.

M. Hall, M. Woolhouse, and A. Rambaut. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS computational biology*, 11(12):e1004613, 2015.

R. J. Hall, S. Gubbins, and C. A. Gilligan. Invasion of drug and pesticide resistance is determined by a trade-off between treatment efficacy and relative fitness. *Bulletin of Mathematical Biology*, 66(4):825–840, 2004.

W. H. Hamer. *The Milroy Lectures on Epidemic Diseases in England: The Evidence of Variability and of Persistency of Type; Delivered Before the Royal College of Physicians of London, March 1st, 6th, and 8th, 1906.* Bedford Press, 1906.

A. Handel, I. M. Longini, and R. Antia. Towards a quantitative understanding of the within-host dynamics of influenza a infections. *Journal of the Royal Society Interface*, 7(42):35–47, 2010.

Y. Hayama, S. M. Firestone, M. A. Stevenson, T. Yamamoto, T. Nishi, Y. Shimizu, and T. Tsutsui. Reconstructing a transmission network and identifying risk factors of secondary transmissions in the 2010 foot-and-mouth disease outbreak in japan. *Transboundary and emerging diseases*, 66(5):2074–2086, 2019.

D. T. Haydon, M. Chase-Topping, D. Shaw, L. Matthews, J. Friar, J. Wilesmith, and M. Woolhouse. The construction and analysis of epidemic trees with reference to the 2001 uk foot–and–mouth outbreak. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1511):121–127, 2003.

J. C. Heijne, M. Rondy, L. Verhoef, J. Wallinga, M. Kretzschmar, N. Low, M. Koopmans, and P. F. Teunis. Quantifying transmission of norovirus during an outbreak. *Epidemiology*, 23(2):277–284, 2012.

J. T. Herbeck, J. E. Mittler, G. S. Gottlieb, and J. I. Mullins. An hiv epidemic model based on viral load dynamics: value in assessing empirical trends in hiv virulence and community viral load. *PLoS Comput Biol*, 10(6):e1003673, 2014.

H. W. Hethcote. A thousand and one epidemic models. In *Frontiers in mathematical biology*, pages 504–515. Springer, 1994.

H. W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4): 599–653, 2000.

H. W. Hethcote and S. A. Levin. Periodicity in epidemiological models. In *Applied mathematical ecology*, pages 193–211. Springer, 1989.

H. W. Hethcote and J. W. Van Ark. Modeling hiv transmission and aids in the united states. *Lecture notes in biomathematics*, 95, 1991.

H. W. Hethcote, H. W. Stech, and P. van den Driessche. Periodicity and stability in epidemic models: a survey. In *Differential equations and applications in ecology, epidemics, and population problems*, pages 65–82. Elsevier, 1981.

F. Hoppenstaedt. *Mathematical theories of populations: demographics, genetics and epidemics*. SIAM, 1975.

J. Hughes, R. C. Allen, M. Baguelin, K. Hampson, G. J. Baillie, D. Elton, J. R. Newton, P. Kellam, J. L. Wood, E. C. Holmes, et al. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS pathogens*, 8(12):e1003081, 2012.

A. Huppert and G. Katriel. Mathematical modelling and prediction in infectious disease epidemiology. *Clinical microbiology and infection*, 19(11):999–1005, 2013.

V. Isham and G. Medley. *Models for infectious human diseases: their structure and relation to data*, volume 6. Cambridge University Press, 1996.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

M. John. *A dictionary of epidemiology*. Oxford university press, 2001.

S. Johnson. *The ghost map: The story of London's most terrifying epidemic–and how it changed science, cities, and the modern world*. Penguin, 2006.

T. Jombart, R. Eggo, P. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383, 2011.

T. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS computational biology*, 10(1):e1003457, 2014.

R. R. Kao. The role of mathematical modelling in the control of the 2001 fmd epidemic in the uk. *Trends in microbiology*, 10(6):279–286, 2002.

S. Karlin. *A first course in stochastic processes*. Academic press, 2014.

M. J. Keeling and K. T. Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

M. J. Keeling, P. Rohani, and B. T. Grenfell. Seasonally forced disease dynamics explored as switching between attractors. *Physica D: Nonlinear Phenomena*, 148(3-4):317–335, 2001.

R. Kelatlhegile. A simply hiv/aids models with density-dependent demographics. *International Journal of Applied Mathematics*, 25(4):525–545, 2012.

E. Kenah, T. Britton, M. E. Halloran, and I. M. Longini Jr. Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees. *PLoS computational biology*, 12(4):e1004869, 2016.

D. G. Kendall. Deterministic and stochastic epidemics in closed populations. In *Proc. 3rd Berkeley Symp. Math. Statist. Prob*, volume 4, pages 149–165, 1956.

M. Kendall, D. Ayabina, Y. Xu, J. Stimson, C. Colijn, et al. Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees. *Statistical Science*, 33(1):70–85, 2018.

W. O. Kermack and A. G. McKendrick. 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, 1927.

M. Kimura. The neutral theory and molecular evolution. In *My Thoughts on Biological Evolution*, pages 119–138. Springer, 2020.

D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, and J. Wallinga. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS computational biology*, 13(5):e1005495, 2017.

K. Koelle, S. Cobey, B. Grenfell, and M. Pascual. Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. *Science*, 314 (5807):1898–1903, 2006.

A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The lancet infectious diseases*, 2020.

C.-H. Kuo, N. A. Moran, and H. Ochman. The consequences of genetic drift for bacterial genome complexity. *Genome research*, 19(8):1450–1454, 2009.

T. LaBar and C. Adami. Evolution of drift robustness in small populations. *Nature communications*, 8(1):1–12, 2017.

S. Lakhani. Early clinical pathologists: Edward jenner (1749-1823). *Journal of clinical pathology*, 45(9):756, 1992.

L. R. Lareo and O. E. Acevedo. Sequence mapping in a three-dimensional space by a numeric method and some of its applications. *Acta biotheoretica*, 47(2): 123–128, 1999.

M. S. Lau, G. Marion, G. Streftaris, and G. Gibson. A systematic bayesian integration of epidemiological and genetic data. *PLoS computational biology*, 11(11): e1004633, 2015.

M. Lauck, M. V. Alvarado-Mora, E. A. Becker, D. Bhattacharya, R. Striker, A. L. Hughes, F. J. Carrilho, D. H. O'Connor, and J. R. R. Pinho. Analysis of hepatitis c virus intrahost diversity across the coding region by ultradeep pyrosequencing. *Journal of virology*, 86(7):3952–3960, 2012.

S. V. Leavitt, R. S. Lee, P. Sebastiani, C. R. Horsburgh, H. E. Jenkins, and L. F. White. Estimating the relative probability of direct transmission between infectious disease patients. *International Journal of Epidemiology*, 2020.

T. Leitner and E. Romero-Severson. Phylogenetic patterns recover known hiv epidemiological relationships and reveal common transmission of multiple variants. *Nature microbiology*, 3(9):983–988, 2018.

P. Lemey, A. Rambaut, J. J. Welch, and M. A. Suchard. Phylogeography takes a relaxed random walk in continuous space and time. *Molecular biology and evolution*, 27(8):1877–1885, 2010.

Y. Leo, M. Chen, B. Heng, and C. Lee. Severe acute respiratory syndrome-singapore, 2003. *MMWR: Morbidity & Mortality Weekly Report*, 52(18):405–405, 2003.

C. Leyronas, C. E. Morris, M. Choufany, and S. Soubeyrand. Assessing the aerial interconnectivity of distant reservoirs of sclerotinia sclerotiorum. *Frontiers in microbiology*, 9:2257, 2018.

Y. Lin, D. Jiang, and T. Liu. Nontrivial periodic solution of a stochastic epidemic model with seasonal variation. *Applied Mathematics Letters*, 45:103–107, 2015.

I. M. Longini Jr and M. E. Halloran. Strategy for distribution of influenza vaccine to high-risk groups and children. *American journal of epidemiology*, 161(4): 303–306, 2005.

K.-J. Lui, W. W. Darrow, and G. W. Rutherford. A model-based estimate of the mean incubation period for aids in homosexual men. *Science*, 240(4857):1333–1335, 1988.

B. MacMahon, T. F. Pugh, J. Ipsen, et al. Epidemiologie methods. *Epidemiologie Methods.*, 1960.

S. Makintubee, J. Mallonee, and G. R. Istre. Shigellosis outbreak associated with swimming. *American journal of public health*, 77(2):166–168, 1987.

M. B. Mandary, M. Masomian, and C. L. Poh. Impact of rna virus evolution on quasispecies formation and virulence. *International journal of molecular sciences*, 20(18):4657, 2019.

F. A. Milner and R. Zhao. Sir model with directed spatial diffusion. *Mathematical Population Studies*, 15(3):160–181, 2008.

A. M'Kendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1925.

N. Mollentze, L. H. Nel, S. Townsend, K. Le Roux, K. Hampson, D. T. Haydon, and S. Soubeyrand. A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B: Biological Sciences*, 281(1782):20133251, 2014.

M. J. Morelli, G. Thébaud, J. Chadœuf, D. P. King, D. T. Haydon, and S. Soubeyrand. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol*, 8(11):e1002768, 2012.

J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5(3):e74, 2008.

P. R. Murcia, G. J. Baillie, J. Daly, D. Elton, C. Jervis, J. A. Mumford, R. Newton, C. R. Parrish, K. Hoelzer, G. Dougan, et al. Intra-and interhost evolutionary dynamics of equine influenza virus. *Journal of virology*, 84(14):6943–6954, 2010.

P. R. Murcia, J. Hughes, P. Battista, L. Lloyd, G. J. Baillie, R. H. Ramirez-Gonzalez, D. Ormond, K. Oliver, D. Elton, J. A. Mumford, et al. Evolution of an eurasian avian-like influenza virus in naive and vaccinated pigs. *PLoS Pathogens*, 8(5): e1002730, 2012.

J. Murray. Mathematical biology springer verlag berlin. *Heidelberg, New York*, 1989.

J. D. Murray, E. A. Stanley, and D. L. Brown. On the spatial spread of rabies among foxes. *Proceedings of the Royal society of London. Series B. Biological sciences*, 229(1255):111–150, 1986.

I. Nåsell. On the time to extinction in recurrent epidemics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):309–330, 1999.

I. Nåsell. Measles outbreaks are not chaotic. In *Mathematical approaches for emerging and reemerging infectious diseases: Models, methods, and theory*, pages 85–114. Springer, 2002.

C. W. Nelson and A. L. Hughes. Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. *Infection, Genetics and Evolution*, 30:1–7, 2015.

A. Neumaier. Mathematical model building. In *Modeling Languages in Mathematical Optimization*, pages 37–43. Springer, 2004.

J. Newton, J. Daly, L. Spencer, and J. Mumford. Description of the outbreak of equine influenza (h3n8) in the united kingdom in 2003, during which recently vaccinated horses in newmarket developed respiratory disease. *Veterinary Record*, 158(6):185–192, 2006.

M. Nowak and R. M. May. *Virus dynamics: mathematical principles of immunology and virology: mathematical principles of immunology and virology*. Oxford University Press, UK, 2000.

K. A. Pawelek, G. T. Huynh, M. Quinlivan, A. Cullinane, L. Rong, and A. S. Perelson. Modeling within-host dynamics of influenza virus infection including immune responses. *PLoS Comput Biol*, 8(6):e1002588, 2012.

J. Pelé, J.-M. Bécu, H. Abdi, and M. Chabbert. Bios2mds: an r package for comparing orthologous protein families by metric multidimensional scaling. *BMC bioinformatics*, 13(1):133, 2012.

S. J. Pethybridge and L. Madden. Analysis of spatiotemporal dynamics of virus spread in an australian hop garden by stochastic modeling. *Plant disease*, 87(1): 56–62, 2003.

D. U. Pfeiffer and K. B. Stevens. Spatial and temporal epidemiological analysis in the big data era. *Preventive veterinary medicine*, 122(1-2):213–220, 2015.

S. Piry, C. Wipf-Scheibel, J.-F. Martin, M. Galan, and K. Berthier. High throughput amplicon sequencing to assess within-and between-host genetic diversity in plant viruses. *bioRxiv*, page 168773, 2017.

E. Z. Poirier and M. Vignuzzi. Virus population dynamics during infection. *Current opinion in virology*, 23:82–87, 2017.

M. Porta. *A dictionary of epidemiology*. Oxford university press, 2014.

O. G. Pybus and A. Rambaut. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10(8):540, 2009.

O. G. Pybus, M. A. Suchard, P. Lemey, F. J. Bernardin, A. Rambaut, F. W. Crawford, R. R. Gray, N. Arinaminpathy, S. L. Stramer, M. P. Busch, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071, 2012.

D. A. Rasmussen, O. Ratmann, and K. Koelle. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol*, 7(8): e1002136, 2011.

L. Rimbaud, S. Dallot, T. Gottwald, V. Decroocq, E. Jacquot, S. Soubeyrand, and G. Thébaud. Sharka epidemiology and worldwide management strategies: learning lessons to optimize disease control in perennial plants. *Annual review of phytopathology*, 53:357–378, 2015.

L. Roques, O. Bonnefon, V. Baudrot, S. Soubeyrand, and H. Berestycki. A parsimonious model for spatial transmission and heterogeneity in the covid-19 propagation. *arXiv preprint arXiv:2007.08002*, 2020a.

L. Roques, E. K. Klein, J. Papaïx, A. Sar, and S. Soubeyrand. Impact of lockdown on the epidemic dynamics of covid-19 in france. *Frontiers in Medicine*, 7:274, 2020b.

J. V. Ross, T. House, and M. J. Keeling. Calculation of disease dynamics in a population of households. *PLoS One*, 5(3):e9666, 2010.

R. Ross. *The prevention of malaria*. John Murray, 1911.

R. Ross. An application of the theory of probabilities to the study of a priori pathometry.—part i. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 92(638):204–230, 1916.

R. Ross and H. P. Hudson. An application of the theory of probabilities to the study of a priori pathometry.—part ii. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 93(650): 212–225, 1917.

K. J. Rothman, S. Greenland, and T. L. Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

R. A. Saenz, M. Quinlivan, D. Elton, S. MacRae, A. S. Blunden, J. A. Mumford, J. M. Daly, P. Digard, A. Cullinane, B. T. Grenfell, et al. Dynamics of influenza virus infection and pathology. *Journal of virology*, 84(8):3974–3983, 2010.

J. Salt, P. Barnett, P. Dani, and L. Williams. Emergency vaccination of pigs against foot-and-mouth disease: protection against disease and reduction in contact transmission. *Vaccine*, 16(7):746–754, 1998.

A. Saltelli, K. Chan, M. Scott, et al. Sensitivity analysis. probability and statistics series, 2000.

A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

R. Sanjuán, A. Moya, and S. F. Elena. The distribution of fitness effects caused by single-nucleotide substitutions in an rna virus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8396–8401, 2004.

A. Sasaki and H. Godfray. A model for the coevolution of resistance and virulence in coupled host–parasitoid interactions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1418):455–463, 1999.

P. A. Schulte and F. P. Perera. *Molecular epidemiology: principles and practices*. Academic Press, 1998.

S. J. Schwager, C. Castillo-Chavez, and H. Hethcote. Statistical and mathematical approaches in hiv/aids modeling: a review. In *Mathematical and statistical approaches to AIDS epidemiology*, pages 2–35. Springer, 1989.

M. Senga, A. Koi, L. Moses, N. Wauquier, P. Barboza, M. D. Fernandez-Garcia, E. Engedashet, F. Kuti-George, A. D. Mitiku, M. Vandi, et al. Contact tracing

performance during the ebola virus disease outbreak in kenema district, sierra leone. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372 (1721):20160300, 2017.

Z. Shen, F. Ning, W. Zhou, X. He, C. Lin, D. P. Chin, Z. Zhu, and A. Schuchat. Superspreading sars events, beijing, 2003. *Emerging infectious diseases*, 10(2): 256, 2004.

H. Simmons, J. Dunham, J. Stack, B. Dickins, I. Pagan, E. Holmes, and A. Stephenson. Deep sequencing reveals persistence of intra-and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. *Journal of General Virology*, 93(8):1831–1840, 2012.

P. Skums, A. Zelikovsky, Z. Dimitrova, S. Ramachandran, D. Campo, L. Bunimovich, and Y. Khudyakov. Reconstruction of disease transmissions from viral quasispecies genomic data. 2018a.

P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, et al. Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, 34(1): 163–170, 2018b.

A. M. Smith and A. S. Perelson. Influenza a virus infection kinetics: quantitative data and models. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(4):429–445, 2011.

A. M. Smith and R. M. Ribeiro. Modeling the viral dynamics of influenza a virus infection. *Critical Reviews™ in Immunology*, 30(3), 2010.

E. S. Snitkin, A. M. Zelazny, P. J. Thomas, F. Stock, D. K. Henderson, T. N. Palmore, J. A. Segre, N. C. S. Program, et al. Tracking a hospital outbreak of carbapenem-resistant klebsiella pneumoniae with whole-genome sequencing. *Science translational medicine*, 4(148):148ra116–148ra116, 2012.

J. Snow. *On the mode of communication of cholera*. John Churchill, 1855.

S. Soubeyrand. Construction of semi-markov genetic-space-time seir models and inference. *Journal de la Société Française de Statistique*, 157(1):129–152, 2016.

T. Stadler and S. Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120198, 2013.

J. Steffensen. Deux problemes du calcul des probabilités. In *Annales de l'institut Henri Poincaré*, volume 3, pages 319–344, 1933.

J. F. Steffensen. Om sandsynligheden for at afkommet uddør. *Matematisk tidsskrift. B*, pages 19–23, 1930.

A. J. Tatem, Z. Huang, C. Narib, U. Kumar, D. Kandula, D. K. Pindolia, D. L. Smith, J. M. Cohen, B. Graupe, P. Uusiku, et al. Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malaria journal*, 13(1):52, 2014.

S. Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

J. Usher. Some mathematical models for cancer chemotherapy. *Computers & Mathematics with Applications*, 28(9):73–80, 1994.

N. K. Vaidya, R. M. Ribeiro, C. J. Miller, and A. S. Perelson. Viral dynamics during primary simian immunodeficiency virus infection: effect of time-dependent virus infectivity. *Journal of virology*, 84(9):4302–4310, 2010.

P. Van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical biosciences*, 180(1-2):29–48, 2002.

A. J. van Hoek, A. Melegaro, N. Gay, J. Bilcke, and W. J. Edmunds. The cost-effectiveness of varicella and combined varicella and herpes zoster vaccination programmes in the united kingdom. *Vaccine*, 30(6):1225–1234, 2012.

R. R. Vatsavai, A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, and S. Shekhar. Spatiotemporal data mining in the era of big spatial data: algorithms and applications. In *Proceedings of the 1st ACM SIGSPATIAL international workshop on analytics for big geospatial data*, pages 1–10, 2012.

T. M. Walker, C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, et al. Whole-genome sequencing to

delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases*, 13(2):137–146, 2013.

T. M. Walker, M. K. Lalor, A. Broda, L. S. Ortega, M. Morgan, L. Parker, S. Churchill, K. Bennett, T. Golubchik, A. P. Giess, et al. Assessment of mycobacterium tuberculosis transmission in oxfordshire, uk, 2007–12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine*, 2 (4):285–292, 2014.

J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1609):599–604, 2007.

J. Wallinga and P. Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 160(6):509–516, 2004.

X. Wang, I. K. Jordan, and L. W. Mayer. A phylogenetic perspective on molecular epidemiology. In *Molecular Medical Microbiology*, pages 517–536. Elsevier, 2015.

H. W. Watson and F. Galton. On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144, 1875.

K. Wickwire. Mathematical models for the control of pests and infectious diseases: a survey. *Theoretical population biology*, 11(2):182–238, 1977.

R. Williams and J. Wright. Epidemiological issues in health needs assessment. *Bmj*, 316(7141):1379, 1998.

E. B. Wilson and M. H. Burke. The epidemic curve. *Proceedings of the National Academy of Sciences of the United States of America*, 28(9):361, 1942.

C. J. Worby and T. D. Read. 'seedy'(simulation of evolutionary and epidemiological dynamics): An r package to follow accumulation of within-host mutation in pathogens. *PloS one*, 10(6):e0129745, 2015.

C. J. Worby, M. Lipsitch, and W. P. Hanage. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS computational biology*, 10(3):e1003549, 2014.

C. J. Worby, P. D. O'Neill, T. Kypraios, J. V. Robotham, D. De Angelis, E. J. Cartwright, S. J. Peacock, and B. S. Cooper. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics*, 10(1):395, 2016.

C. J. Worby, M. Lipsitch, and W. P. Hanage. Shared genomic variants: Identification of transmission routes using pathogen deep-sequence data. *American journal of epidemiology*, 186(10):1209–1216, 2017.

C. F. Wright, M. J. Morelli, G. Thébaud, N. J. Knowles, P. Herzyk, D. J. Paton, D. T. Haydon, and D. P. King. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *Journal of virology*, 85(5):2266–2275, 2011.

C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser, and T. B. Collaboration. Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular biology and evolution*, 35(3):719–733, 2018.

X.-K. Xu, X.-F. Liu, Y. Wu, S. T. Ali, Z. Du, P. Bosetti, E. H. Lau, B. J. Cowling, and L. Wang. Reconstruction of transmission pairs for novel coronavirus disease 2019 (covid-19) in mainland china: estimation of super-spreading events, serial interval, and hazard of infection. *Clinical Infectious Diseases*, 2020.

Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, et al. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of Thoracic Disease*, 12(3): 165, 2020.

R. J. Ypma, A. Bataille, A. Stegeman, G. Koch, J. Wallinga, and W. M. Van Ballegooijen. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1728):444–450, 2012.

R. J. Ypma, M. Jonges, A. Bataille, A. Stegeman, G. Koch, M. Van Boven, M. Koopmans, W. M. Van Ballegooijen, and J. Wallinga. Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *The Journal of infectious diseases*, 207(5):730–735, 2013a.

R. J. Ypma, W. M. van Ballegooijen, and J. Wallinga. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–1062, 2013b.

A. Ziegler and I. R. König. Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(1):55–63, 2014.

APPENDICES

A. Appendix of the article in Chapter 3

Electronic supplementary material

Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases

Alamil M., Hughes J., Berthier K., Desbiez C., Thébaud G. and Soubeyrand S.

March 18, 2019

Table S1: Statistics about data corresponding to the three case studies, namely Influenza in pigs, Ebola in humans and a potyvirus in salsifies.

Statistics	Influenza		Ebola	Potyvirus
	Naive chain	Vaccinated chain		
Number of host units	10	13	58	27
Number of sequence fragments	1	1	31	1
Fragment length°	939	939	885†	438
Mean (SD) sequencing depth	41.3 (16.2)	58.3 (14.8)	14300 (17200)†	1550 (930)
Number of different variants	331	623	16.1†	278
Mean (SD) number of different variants per host unit	18.6 (7.0)	26.1 (9.4)	1.37 (0.64)†	10.3 (7.6)
Mean (SD) distance b/n variants*	3.31	3.61 (1.34)	2.42 (1.01)†	25.9 (6.6)
Mean (SD) within-host distance b/n variants*	1.17	2.80 (1.00)	1.37 (0.56)†	23.6 (3.4)

° Obtained after the removal of sites with missing values.

† Average over the 31 available sequence fragments.

* The (genetic) distance between (b/n) two sequence fragments is the number of different nucleotides.

Table S2: Contact information used for the reconstruction of transmission chains of Influenza in pigs and Ebola in humans. Note that host 401 was alone in group 3 of the vaccinated chain.

Outbreak	Contact information	Training host	Contact
Swine influenza	For 2 hosts in the last group	106	105, 108, 112
Naive chain		112	105, 108, 106
	For 2 hosts in groups 3 and 4	111	104, 116, 109
		108	109, 111, 105
Swine influenza	For 2 hosts in the last group	400	412, 414, 413
Vaccinated chain		413	412, 414, 400
	For 2 hosts in groups 3 and 4	401	409, 417
		416	401,415
Ebola	For 5 hosts among 58	G3817	G3729
		G3820	G3729
		G3821	G3729
		G3823	G3729
		G3851	G3752

Table S3: Mean difference between each sequence fragment of each training host and the closest sequence fragment in its source identified by contact tracing. The last two lines give, for each host, the average and the standard deviation of the mean difference over all sequence fragments. Figures lower than 0.0005 are denoted by 0 to facilitate the identification of significant positive values.

Fragment	G3817	G3820	G3821	G3823	G3851
500-1500	0	0	0.020	0	0.031
1000-2000	0	0.012	0	0	0
1500-2500	0	0	0	0	0
2000-3000	0	0.002	0	0	0
2500-3500	0.009	0	0.024	0.020	0.013
3000-4000	0	0	0	0	0
3500-4500	0.002	0	0.074	0	0.012
4000-5000	0	0	1.081	0	0
4500-5500	0	0	0.029	0	0.014
5000-6000	0	0.018	0	0	0
5500-6500	0	0	0	0	0
6000-7000	0	0	0.011	0	0
6500-7500	0	0	0	0	0
7000-8000	0	0.005	0	0	0.010
7500-8500	1.047	0.073	0.078	0	0
8000-9000	1.010	0	0	0	0.037
8500-9500	0	0	0	0	0
9000-10000	0	0	0.050	0	0
9500-10500	2.000	0	1.000	0	0
10000-11000	2.000	0	1.000	0	0
10500-11500	0	0	0.057	0	0.002
11000-12000	0	0.073	0	0.005	0.003
11500-12500	0	0	0	0	0
13000-14000	1.000	0	0	0	0
13500-14500	1.000	0	0	0	0
14000-15000	0	0.002	0	0	0
14500-15500	0	0	0	0	0
15000-16000	0	0	0	0	0
15500-16500	0	0	0.034	0	0
16000-17000	0	0	0	1.001	0.001
16500-17500	0	0	0.043	0.936	0
17000-18000	0	0	0	0	0
Average	0.252	0.006	0.109	0.061	0.004
SD	0.570	0.018	0.301	0.238	0.009

Table S4: Potential donors for training host G3817 whose donor identified with contact tracing is G3729. Link intensities and ranks were obtained by cross-validation.

Recipient	Donor	Link intensity	Rank
G3817	EM111	0.0409	1
	G3713	0.0398	2
	G3788	0.0393	3
	G3724	0.0391	4
	EM113	0.0273	5
	EM115	0.0266	6
	G3735	0.0257	7
	G3771	0.0257	7
	G3809	0.0257	7
	EM112	0.0250	10
	EM110	0.0248	11
	G3816	0.0248	12
	G3821	0.0248	12
	EM106	0.0247	14
	EM124	0.0246	15
	EM119	0.0242	16
	G3707	0.0238	17
	EM104	0.0237	18
	NM042	0.0237	18
	G3752	0.0237	20
	G3729	0.0237	21
	G3820	0.0232	22
	G3750	0.0225	23
	G3734	0.0218	24
	EM096	0.0215	25
	G3677	0.0214	26
	G3758	0.0213	27
	G3770	0.0212	28
	G3787	0.0208	29
	G3679	0.0206	30
	G3818	0.0203	31
	EM121	0.0202	32
	G3682	0.0200	33
	EM120	0.0189	34
	G3823	0.0185	35
	G3800	0.0180	36
	G3769	0.0178	37
	G3676	0.0173	38
	G3683	0.0169	39
	G3670	0.0162	40
	G3680	0.0151	41
	G3686	0.0142	42
	G3805	0.0074	43
	G3789	0.0033	44

Table S5: Potential donors for training host G3820 whose donor identified with contact tracing is G3729. Link intensities and ranks were obtained by cross-validation.

Recipient	Donor	Link intensity	Rank
G3820	G3676	0.0525	1
	G3729	0.0277	2
	G3677	0.0270	3
	EM121	0.0269	4
	EM113	0.0269	5
	G3734	0.0268	6
	G3707	0.0267	7
	EM096	0.0263	8
	G3679	0.0254	9
	G3788	0.0249	10
	G3724	0.0244	11
	EM115	0.0244	12
	G3682	0.0239	13
	G3735	0.0238	14
	G3771	0.0238	14
	G3809	0.0238	14
	G3758	0.0238	17
	G3823	0.0236	18
	EM120	0.0231	19
	G3787	0.0229	20
	EM110	0.0229	21
	G3750	0.0229	22
	G3816	0.0229	22
	G3821	0.0229	22
	EM106	0.0227	25
	G3713	0.0226	26
	EM104	0.0223	27
	G3769	0.0223	28
	EM112	0.0220	29
	NM042	0.0219	30
	G3800	0.0218	31
	EM124	0.0216	32
	EM119	0.0215	33
	EM111	0.0211	34
	G3752	0.0205	35
	G3683	0.0204	36
	G3818	0.0192	37
	G3770	0.0191	38
	G3680	0.0188	39
	G3817	0.0179	40
	G3686	0.0178	41
	G3670	0.0170	42
	G3805	0.0057	43
	G3789	0.0033	44

Table S6: Potential donors for training host G3821 whose donor identified with contact tracing is G3729. Link intensities and ranks were obtained by cross-validation.

Recipient	Donor	Link intensity	Rank
G3821	G3816	0.0556	1
	G3729	0.0513	2
	EM104	0.0391	3
	G3771	0.0344	4
	EM106	0.0338	5
	G3752	0.0305	6
	G3788	0.0277	7
	G3787	0.0269	8
	EM096	0.0259	9
	G3734	0.0258	10
	EM113	0.0248	11
	G3735	0.0248	11
	EM115	0.0245	13
	EM111	0.0239	14
	EM112	0.0235	15
	G3670	0.0234	16
	G3809	0.0234	17
	G3683	0.0232	18
	EM110	0.0230	19
	EM124	0.0222	20
	G3707	0.0214	21
	G3770	0.0213	22
	G3724	0.0211	23
	G3713	0.0211	24
	G3818	0.0210	25
	EM119	0.0209	26
	EM121	0.0203	27
	G3677	0.0200	28
	G3758	0.0196	29
	NM042	0.0194	30
	G3820	0.0188	31
	EM120	0.0188	32
	G3750	0.0187	33
	G3682	0.0180	34
	G3679	0.0177	35
	G3817	0.0174	36
	G3823	0.0173	37
	G3769	0.0166	38
	G3800	0.0162	39
	G3676	0.0147	40
	G3680	0.0124	41
	G3686	0.0113	42
	G3805	0.0068	43
	G3789	0.0016	44

Table S7: Potential donors for training host G3823 whose donor identified with contact tracing is G3729. Link intensities and ranks were obtained by cross-validation.

Recipient	Donor	Link intensity	Rank
G3823	G3682	0.0300	1
	EM106	0.0293	2
	G3769	0.0290	3
	EM104	0.0283	4
	G3677	0.0271	5
	EM113	0.0270	6
	G3820	0.0263	7
	EM121	0.0262	8
	G3683	0.0256	9
	G3735	0.0254	10
	G3771	0.0254	10
	EM096	0.0253	12
	G3707	0.0253	13
	G3729	0.0247	14
	EM115	0.0245	15
	G3724	0.0245	15
	G3758	0.0244	17
	G3788	0.0244	18
	G3679	0.0244	19
	G3821	0.0244	19
	G3734	0.0244	21
	EM120	0.0243	22
	G3787	0.0238	23
	G3809	0.0238	24
	EM112	0.0236	25
	EM110	0.0236	26
	EM124	0.0235	27
	G3750	0.0229	28
	G3816	0.0228	29
	G3800	0.0224	30
	G3713	0.0219	31
	EM111	0.0212	32
	NM042	0.0210	33
	G3818	0.0200	34
	G3676	0.0197	35
	G3752	0.0196	36
	EM119	0.0195	37
	G3770	0.0192	38
	G3817	0.0184	39
	G3680	0.0181	40
	G3670	0.0179	41
	G3686	0.0166	42
	G3805	0.0067	43
	G3789	0.0033	44

Table S8: Potential donors for training host G3851 whose donor identified with contact tracing is G3752. Link intensities and ranks were obtained by cross-validation.

Recipient	Donor	Link intensity	Rank	Donor	Link intensity	Rank
G3851	G3769	0.0438	1	G3800	0.0139	45
	G3825	0.0430	2	G3838	0.0135	46
	G3724	0.0326	3	G3670	0.0132	47
	EM106	0.0274	4	G3682	0.0131	48
	G3771	0.0264	5	EM120	0.0128	49
	G3829	0.0262	6	G3817	0.0127	50
	EM104	0.0255	7	G3683	0.0119	51
	G3821	0.0250	8	G3823	0.0118	52
	G3752	0.0244	9	G3676	0.0112	53
	G3850	0.0211	10	G3680	0.0105	54
	EM113	0.0208	11	G3686	0.0098	55
	G3848	0.0208	11	G3805	0.0057	56
	EM115	0.0206	13	G3789	0.0025	57
	G3826	0.0200	14			
	G3856	0.0198	15			
	EM111	0.0187	16			
	G3735	0.0177	17			
	G3809	0.0177	17			
	G3840	0.0177	17			
	G3788	0.0175	20			
	EM110	0.0170	21			
	G3816	0.0170	22			
	NM042	0.0169	23			
	EM112	0.0168	24			
	G3845	0.0168	25			
	G3677	0.0165	26			
	G3707	0.0163	27			
	EM124	0.0161	28			
	G3713	0.0161	29			
	G3729	0.0159	30			
	G3787	0.0158	31			
	G3820	0.0156	32			
	EM119	0.0155	33			
	G3841	0.0147	34			
	G3734	0.0146	35			
	G3770	0.0146	36			
	EM096	0.0145	37			
	G3679	0.0145	37			
	G3750	0.0145	39			
	G3758	0.0144	40			
	G3846	0.0144	41			
	G3831	0.0143	42			
	EM121	0.0142	43			
	G3818	0.0140	44			

Table S9: Specification of the components of the inference procedure. Note that setting Δ_{ij} at the value 1 implies that the substitution parameter μ corresponds, for each inferred transmission, to the expected number of substitutions per nucleotide in the evolutionary duration separating the two samples.

Model component	Influenza	Ebola	Potyvirus
Duration Δ_{ij}	$\Delta_{ij} \equiv 1$	$\Delta_{ij} \equiv 1$	$\Delta_{ij} \equiv 1$
Shape for P_θ	H1-Normal (eq. (4.3))	H2-normal (eq. (4.5)) $(\bar{d}_{\text{obs}}, \sigma^2_{\text{obs}})$ estimated from training donor–recipient pairs	H1-Chi-squared (eq. (4.4))
Set Θ of values for the penalisation parameter θ	Naive chain: $\{0, 0.1, 0.2, \ldots, 4\}$ Vaccinated chain: $\{0, 0.25, 0.5, \ldots, 10\}$	$\{0, 10, 20, \ldots, 200\}$	$\{0, 1, 2, \ldots, 40\}$
Basis for the calibration of θ	Contact tracing (eq. (4.4))	Contact tracing (eq. (4.6))	Geographical distance (eq. (4.7))

Table S10: Discrepancy between inferred transmission graphs and reference graphs measured by the proportion of correct source identifications (CSI) and the Jeffreys discrepancy (JD) averaged over all hosts. For the Influenza case studies, the reference graph is the graph where the source for the hosts in the first group is an external source with probability 1, and the source for the hosts in the subsequent groups is any host in the preceding group with probability 0.5 (when the preceding group consists of 2 hosts) or probability 1 (when the preceding group consists of a single host; this occurs once in the vaccinated chain). For the Ebola case study, the reference is the graph obtained with BadTrIP by De Maio et al. (2018); in this case, the criteria were computed from recipient hosts that were in both analyses (BadTrIP and SLAFEEL). The proportion of CSI is computed as the proportion of hosts whose most likely source (based on the inferred graph) coincides with (one of) its source(s) in the reference graph (for the Ebola case study, the sources in the reference graph are only the most likely sources provided by BadTrIP; see Figure S14 for a less conservative definition). The JD (Chung et al., 1989; Jeffreys, 1946) measures the distance between two finite discrete probability distributions, say $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$, by the quantity $\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2$. In our applications, $\mathbf{p}$ gives for a given recipient host the estimated probability for any other host to be its donor, and $\mathbf{q}$ gives for the same recipient host the reference vector of probabilities built as described above. For each inferred transmission graph, the JD was computed for all observed recipient hosts and, then, averaged.

Case study	Method	Penal.	Training hosts	Figure	Prop. CSI	Mean JD	SE JD
Swine Influenza	SLAFEEL	Yes	106, 112	1 (A)	0.60	0.84	0.19
Naive chain	SLAFEEL	Yes	111, 108	1 (B)	0.60	0.78	0.14
	SLAFEEL	No	–	S3 (A)	0.20	1.48	0.18
	BadTrIP*	–	–	S12 (A)	0.30	0.90	0.14
Swine Influenza	SLAFEEL	Yes	400, 413	1 (C)	0.42	0.86	0.19
Vaccinated chain	SLAFEEL	Yes	401, 416	1 (D)	0.42	0.90	0.24
	SLAFEEL	Yes	400, 413, 401	S1 (A)	0.50	1.02	0.25
	SLAFEEL	Yes	400, 413, 415	S1 (B)	0.50	1.01	0.26
	SLAFEEL	Yes	400, 413, 416	S1 (C)	0.42	1.11	0.27
	SLAFEEL	Yes	401, 415, 416	S1 (D)	0.42	0.90	0.24
	SLAFEEL	No	–	S3 (B)	0.42	0.92	0.22
	BadTrIP*	–	–	S12 (B)	0.33	0.99	0.22
Ebola	SLAFEEL *vs* BadTrIP	Yes	G3817, G3829 G3821, G3823 G3851	3	0.08	0.80	0.04

*To fairly compare BadTrIP and SLAFEEL in the Influenza case studies, we *a posteriori* pruned impossible transmissions inferred by BadTrIP based on temporal information (as we *a priori* did with SLAFEEL), we reweighted the remaining inferred transmissions such that their probabilities sum to 1 for each infected host, and we computed the CSI, the mean JD and the SD of JD from the remaining transmissions and their updated probabilities.

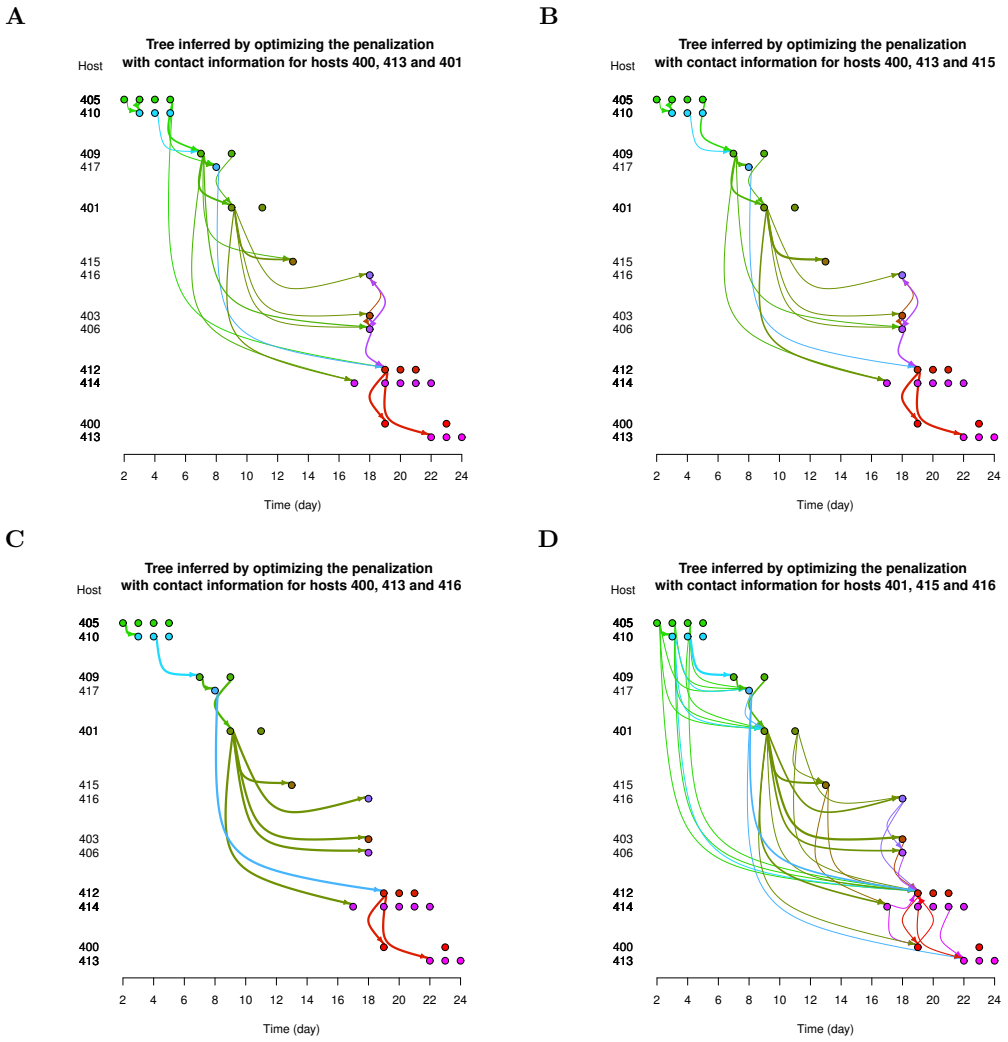


Figure S1: Transmissions inferred in the vaccinated chain with different sets of three training hosts for calibrating the penalisation. The thickness of each arrow is proportional to the intensity of the corresponding link.

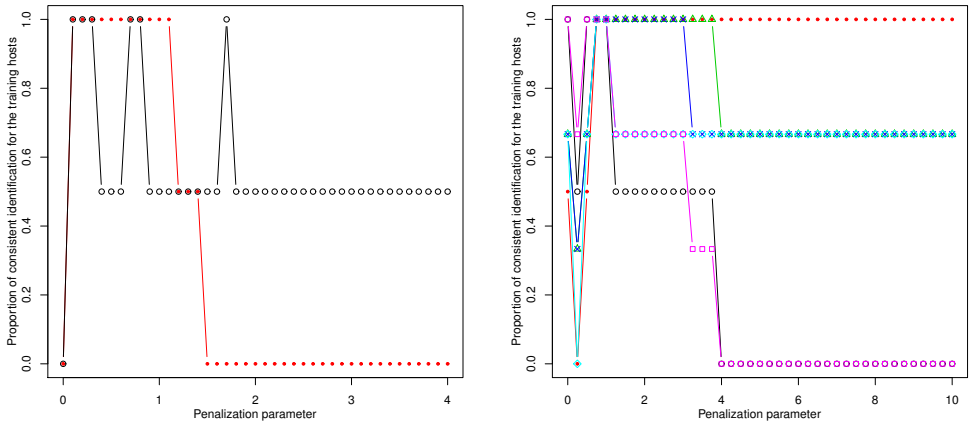


Figure S2: Proportion of source identifications that are consistent with contact information about the training hosts for the naive chain (left) and the vaccinated chain (right), as a function of the penalisation parameter. In each panel, the rate of consistent identifications is shown in red when the training hosts are the two pigs of the last group of the outbreak, and in black when the training hosts are two pigs selected from the 3rd and 4th groups of the outbreak; see details in Table 1 of the main text. In the right panel, the green curve corresponds to training hosts 400, 413 and 401; the dark blue curve to 400, 413 and 415, the light blue curve to 400, 413 and 416 and the pink curve to 401, 415 and 416. Adding a third host to training data allows us to reduce the range of optimal penalisation parameters.

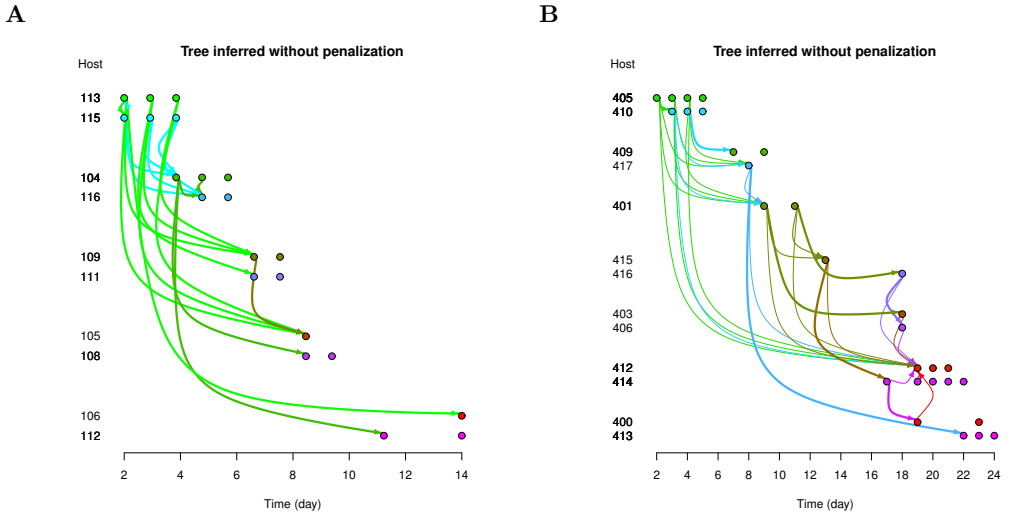


Figure S3: Transmissions inferred in the naive chain (left) and vaccinated chain (right) without including the penalisation and, therefore, without including training hosts.

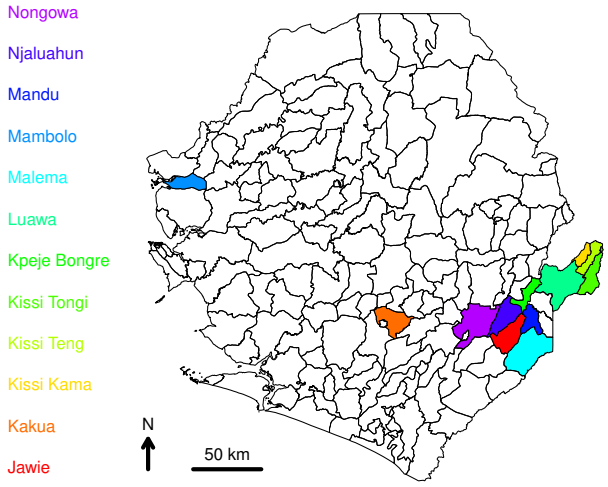


Figure S4: Map of Sierra Leone showing the locations of chiefdoms included in the analysis of Ebola data.

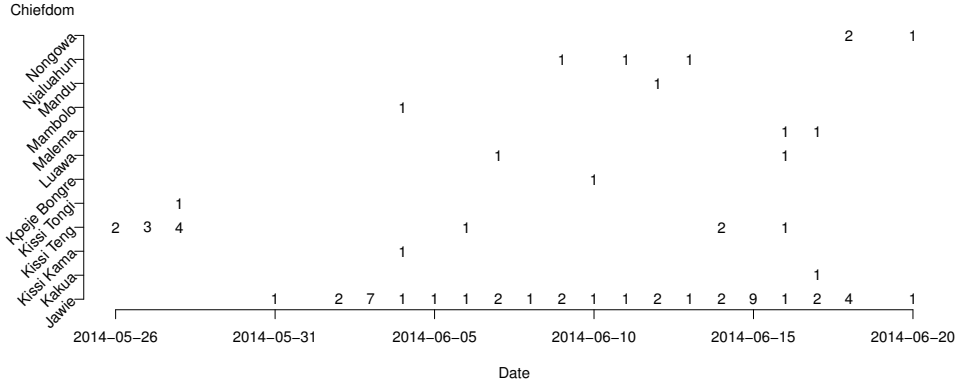


Figure S5: Number of Ebola patients included in the analysis as a function of collection date and chiefdom.

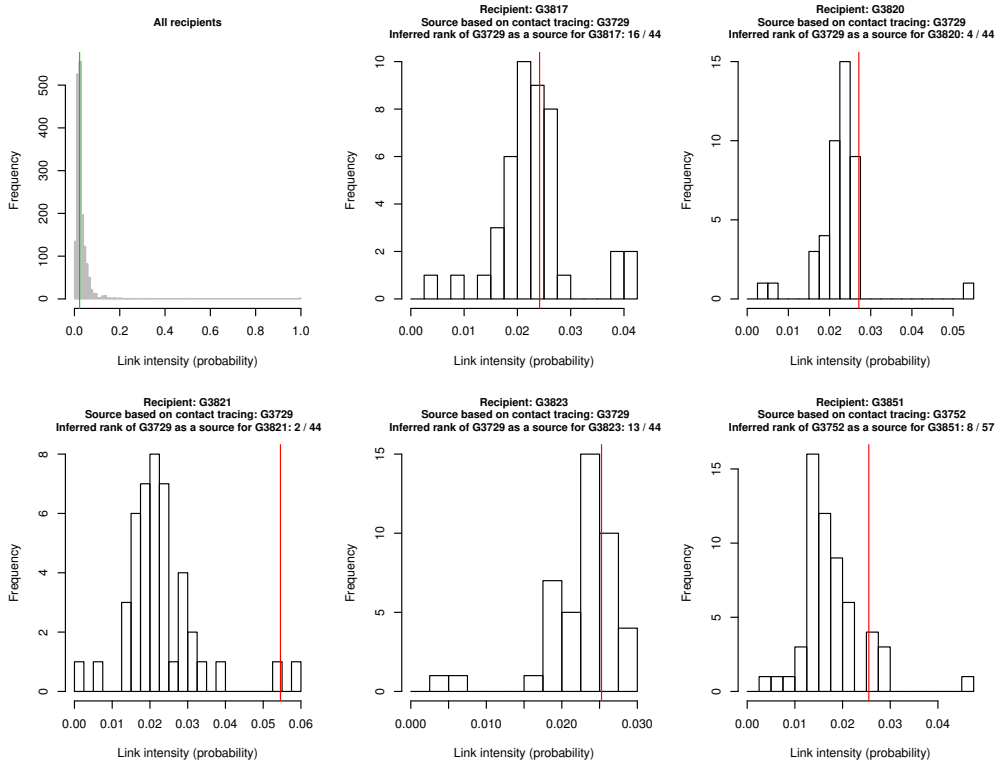


Figure S6: Estimated intensities of links in the Ebola dataset for all recipients (top left panel; green line: median intensity) and for each recipient in the training set of hosts (other panels; red line: intensity for the source identified with contact tracing). This figure was obtained from the combined analysis of 31 sequence fragments and without cross-validation.

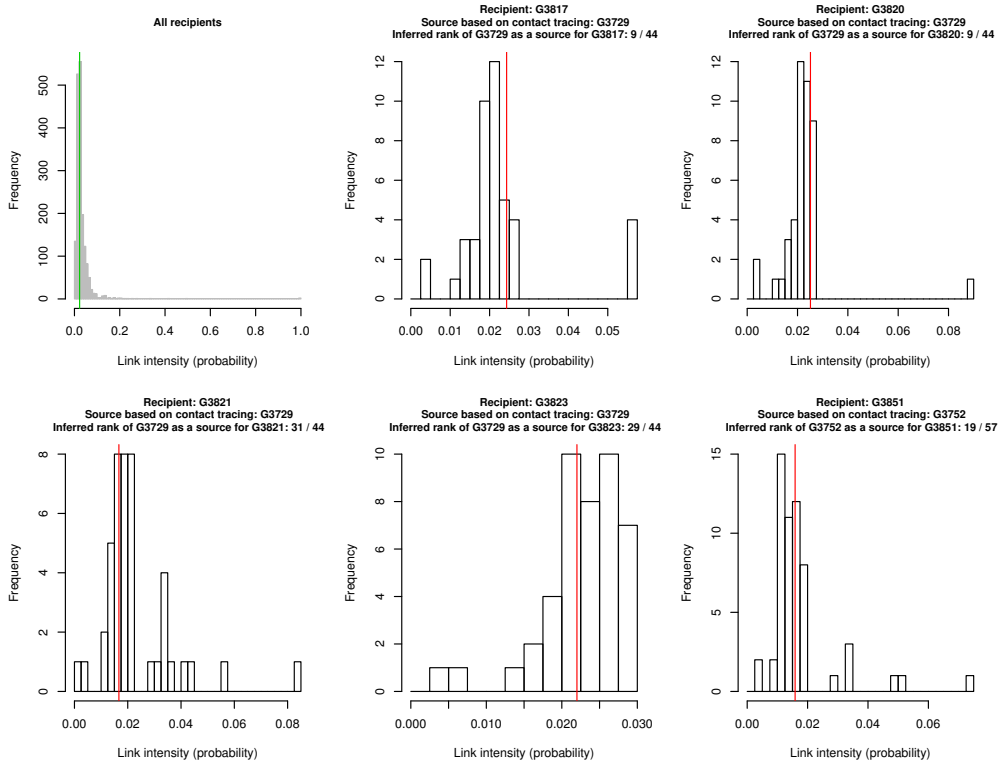


Figure S7: Estimated intensities of links for all recipients (top left panel; green line: median intensity) and for each recipient in the training set of hosts (other panels; red line: intensity for the source identified with contact tracing). This figure was obtained from the combined analysis of 15 sequence fragments from sequence site 500 to sequence site 9000, and with cross-validation.

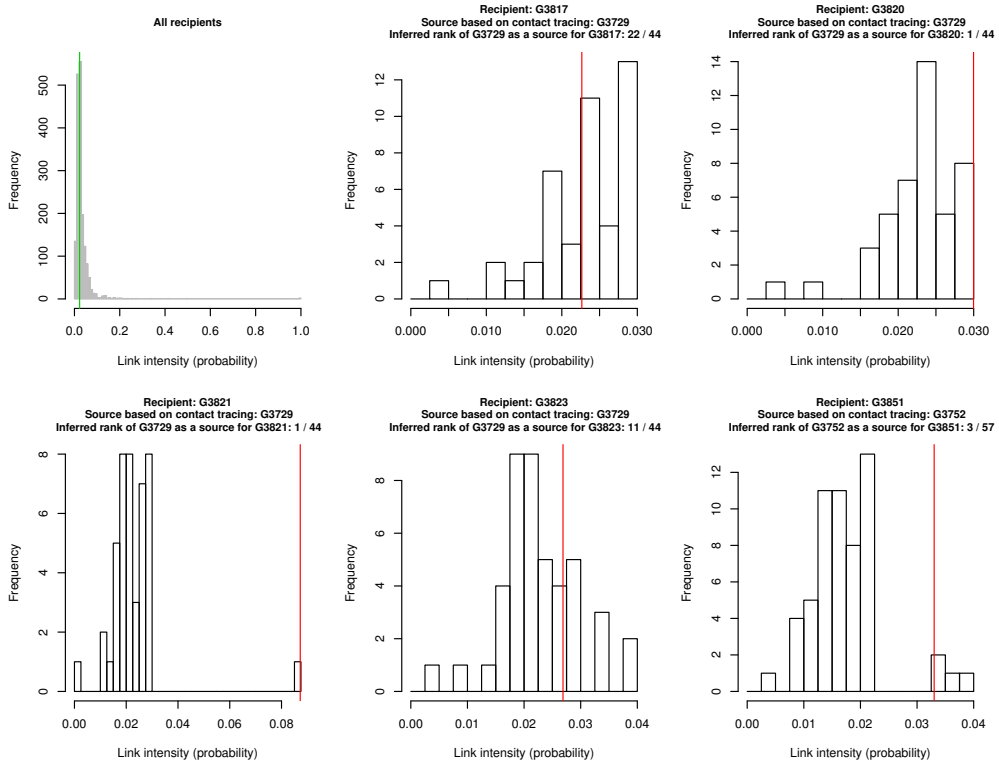


Figure S8: Estimated intensities of links for all recipients (top left panel; green line: median intensity) and for each recipient in the training set of hosts (other panels; red line: intensity for the source identified with contact tracing). This figure was obtained from the combined analysis of 15 sequence fragments from sequence site 9000 to sequence site 18000, and with cross-validation.

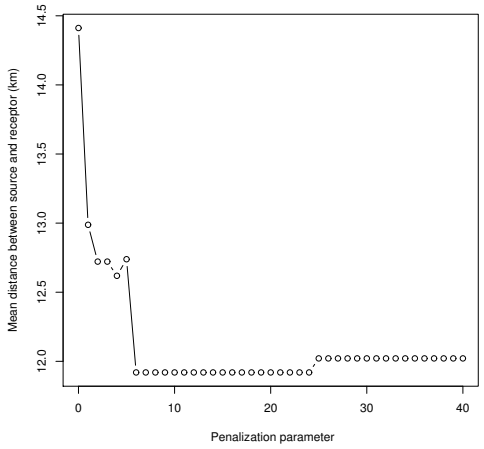


Figure S9: Mean distance between connected salsify patches with respect to the penalisation parameter.

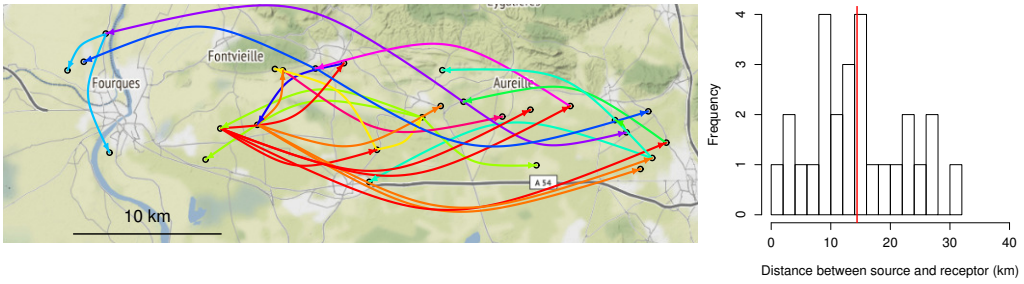


Figure S10: Links inferred without penalisation between salsify populations based on sampled sets of potyvirus sequences (left; links from the same source have the same color) and distribution of link distances (right; the vertical red line indicates the mean distance).

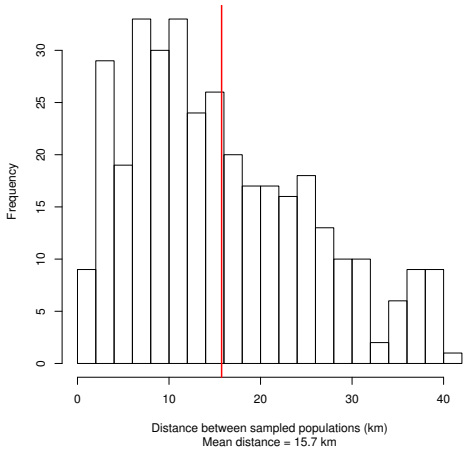


Figure S11: Distribution of distances between salsify patches. The vertical red line indicates the mean distance.

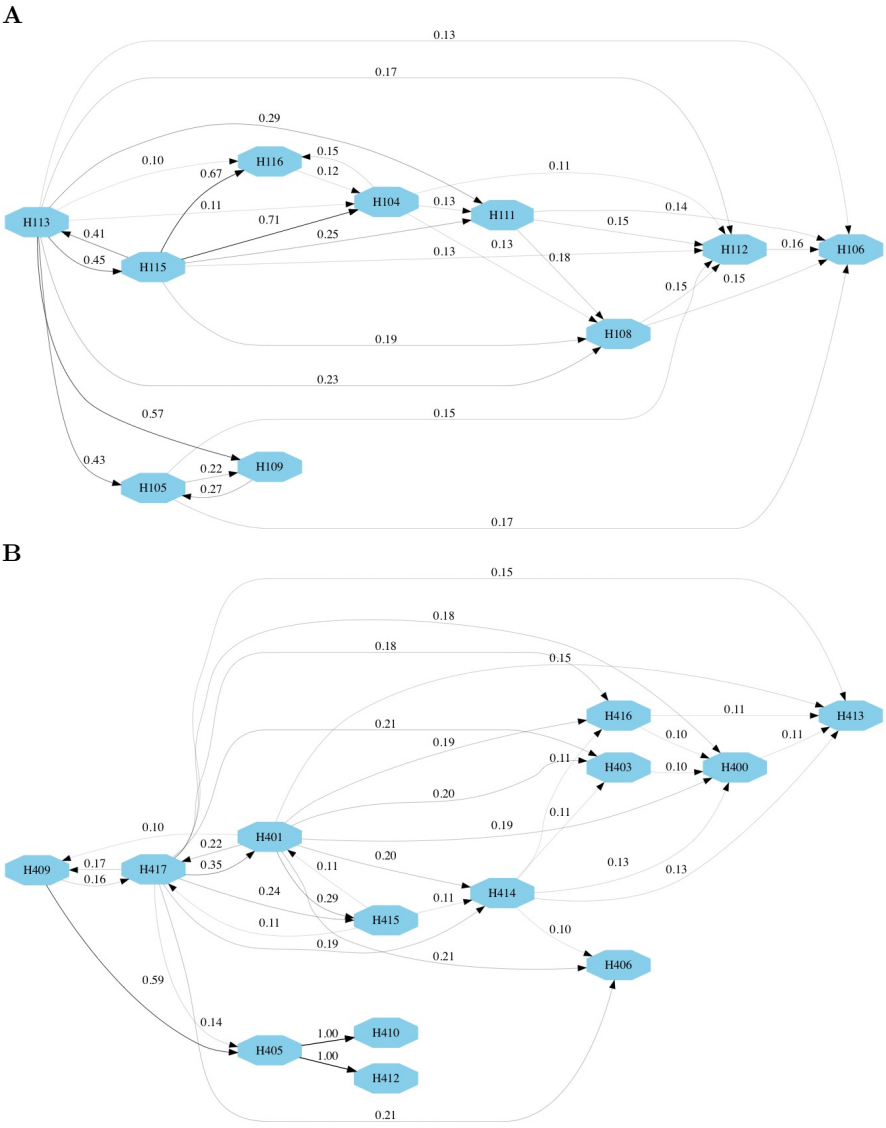


Figure S12: Inference of transmissions in the naive (A) and vaccinated (B) Swine influenza transmission chains. Transmission events with posterior probability higher than 0.10 as inferred by BadTrIP are shown. Hexagons represent hosts, while arrows are transmission events between hosts. The posterior probability of transmissions are shown next to the arrows and higher values are shown with thicker arrows. — For both datasets, the sequences for each sample were re-coded for use in the BadTrIP package (De Maio et al., 2018) embedded in BEAST2 (Bouckaert et al., 2014). BadTrIP uses the PoMo model (De Maio et al., 2015) that describes how a population evolves along the branches of a population tree. We allowed each host in the Swine influenza transmission chain to be infectious for the whole period of the experiment. We ran the BadTrIP MCMC for approximately 4 million independent steps, which provided an effective sample size of 20 and took one week of computation (on one CPU of an iMac 4 GHZ Intel Core i7).

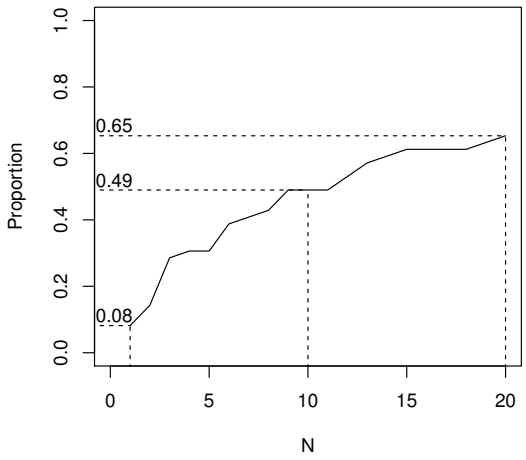


Figure S13: Proportion of recipient hosts whose SLAFEEL-based most likely sources are among the N BadTrIP-based most likely sources. The proportion obtained when N=1 corresponds to the proportion of correct source identifications (CSI) provided in Table S10.

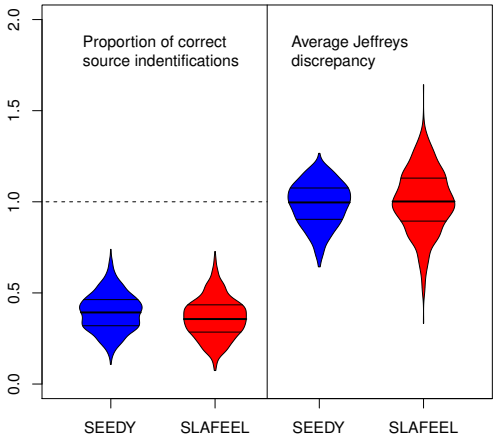


Figure S14: Comparison between SLAFEEL and SEEDY (Worby and Read, 2015) in their ability to identify transmission trees simulated with SEEDY. The comparison was made by assessing the discrepancy between inferred transmission graphs and the simulated graphs, using two criteria: the proportion of correct source identifications and the average Jeffreys discrepancy (see their definitions in table S10). These criteria were computed for 1000 data sets generated with SEEDY by using parameter values chosen by Worby and Read to generate their 4th figure (see details below). The mean epidemic size of simulated outbreaks was: 26.6 infected hosts (SD=2.3). For the application of SLAFEEL to each simulated outbreak, we randomly drew 4 training hosts whose sources were supposed to be known, we chose the H1-normal penalisation, we set $\Delta_{ij} \equiv 1$ and $\Theta = \{0, 1, 2, \ldots, 10\}$. — Outbreaks were simulated with the following parameter values. Number of susceptibles in population: 30. Rate of infection: 0.02. Rate of removal/recovery: 0.001. Mutation rate per sequence per generation: 0.001. Equilibrium population size within host: 1000. Transmission bottleneck size: 10. Samples taken per time point: 10 (1 time point per host, randomly and uniformly drawn between 1 and 300 time steps after host infection). Minimum number of cases before returning (retries until fulfilled): 20. Genome length: 10^5.

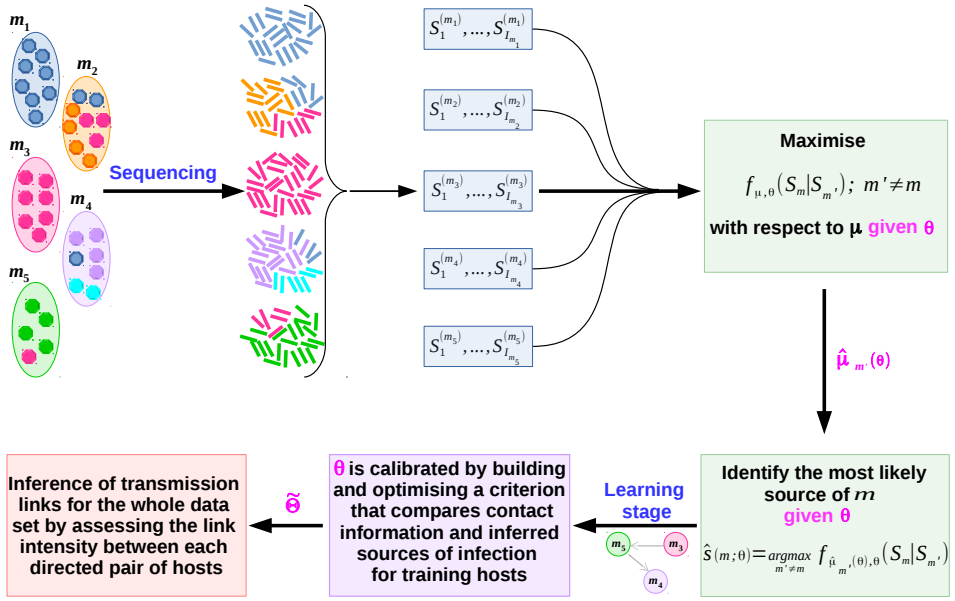


Figure S15: Graphical representation of SLAFEEL. Virus sequences are collected from several hosts $m_1, m_2, \ldots$. In a first step, the penalised pseudo-likelihood $f_{\mu,\theta}(\mathbf{S}_m \mid \mathbf{S}'_m)$ is maximised for each possible donor–recipient pair (m', m) and a set of values for the penalisation parameter θ. This maximisation provides an estimate $\hat{\mu}_{m'}(\theta)$ of the evolutionary parameter μ given θ and the putative source m'. Then, given θ, the most likely source of the recipient m, say $s(m;\theta)$, is identified by maximising $f_{\hat{\mu}_{m'}(\theta),\theta}(\mathbf{S}_m \mid \mathbf{S}'_m)$ with respect to m'. In a second step, by using contact information about training hosts (e.g., m_3 possibly infected m_5 and m_5 possibly infected m_4), the penalization parameter θ is calibrated with a learning approach by building and optimising a criterion that compares contact information and sources of infection $\hat{s}(m_4;\theta)$ and $\hat{s}(m_5;\theta)$ inferred for training hosts m_4 and m_5, respectively. $\tilde{\Theta}$ is the set of penalisation values for which the criterion is optimal. In a third step, the link intensity is used to assess the likelihood of the link between a donor and a recipient.

References

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology 10*, e1003537.

Chung, J. K., P. L. Kannappan, C. T. Ng, and P. K. Sahoo (1989). Measures of distance between probability distributions. *Journal of Mathematical Analysis and Applications 138*, 280–292.

De Maio, N., D. Schrempf, and C. Kosiol (2015). PoMo: An allele frequency-based approach for species tree estimation. *Systematic Biology 64*, 1018–1031.

De Maio, N., C. J. Worby, D. J. Wilson, and N. Stoesser (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Computational Biology 14*, e1006117.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A 186*, 453–461.

Worby, C. J. and T. D. Read (2015). 'SEEDY'(simulation of evolutionary and epidemiological dynamics): An R package to follow accumulation of within-host mutation in pathogens. *PLoS One 10*(6), e0129745.

B. Appendix of the application of SLAFEEL to the Equine influenza virus in Chapter 3

,

Supporting material

Varying the penalization shape and the temporal constraints in SLAFEEL: Illustration with Equine Influenza virus data

Note: This is a preliminary work presented in a rather concise way. In particular, concerning the description of the methodological background, we only highlight the changes with respect to the ***SLAFEEL*** *method described in the previous chapter*

S1 Construction of a joint penalization under H3

Let $\mathbf{S}$ and $\mathbf{S}^{(0)}$ denote the sets of sequences observed from a recipient host and its putative source, respectively. Let $\bar{d} = \frac{1}{J}\sum_{j=1}^{J} \tilde{d}_j$ denote the genetic distance between $\mathbf{S}$ and $\mathbf{S}^{(0)}$, where $\tilde{d}_j$ is the minimum of the genetic distances (measured by the number of different nucleotides) between the j-th sequence in $\mathbf{S}$ and every sequences in $\mathbf{S}^{(0)}$, and J is the number of sequences in $\mathbf{S}$. Let $\Delta_{\text{obs}} = T_{\text{obs}} - T_{\text{obs}}^{(0)}$ the difference between the times of observation of the infection of the recipient and the source.

Note that, even if there is no direct link between $\bar{d}$ and Δ_{obs} (because Δ_{obs} is not the evolutionary duration $\Delta = (T_{\text{obs}} - T_{\text{inf}}) + (T_{\text{obs}}^{(0)} - T_{\text{inf}})$ separating $\mathbf{S}$ and $\mathbf{S}^{(0)}$, where T_{inf} is the infection time of the recipient), there may be a significant link between these two variables.

Here, we want to build a penalization (i) which favors recipient-source pairs whose genetic distances are consistent with training data (like in hypothesis H2), and (ii) which also favors recipient-source pairs whose lags in terms of observation times make sense with respect to the observed genetic distances. Hence, we aim to build a joint penalization over $\bar{d}$ and Δ_{obs} that corresponds to the following hypothesis, say H3:

H3: The distances between sequences in the recipient and their contributing sequences in the source are consistent with some known features, and the relationship between these distances and the lag in the observation of the source and the recipient is consistent with some known features.

As for H2, the *known features* mentioned in H3 are learned from training data. Below, we make explicit these known features as well as the penalization.

The penalization P_θ is the joint distribution of $(\bar{d}, \Delta_{\text{obs}})$ to the power $\theta \geq 0$:

$$\begin{aligned} P_\theta(\mathbf{S}, \mathbf{S}^{(0)}, T_{\text{obs}}, T_{\text{obs}}^{(0)}) &= f(\bar{d}, \Delta_{\text{obs}})^\theta \\ &= \{f(\Delta_{\text{obs}} \mid \bar{d}) f(\bar{d})\}^\theta, \end{aligned}$$

where f denotes the probability distribution function (p.d.f.) of the variable of interest (or the conditional p.d.f. of the variable given another variable). We first assume that $\bar{d}$ is drawn from a gamma distribution with shape and scale parameters (α_1, α_2) that are estimated from training data:

$$\bar{d} \sim \text{Gamma}(\alpha_1, \alpha_2). \tag{S1}$$

For deriving the conditional distribution of $\Delta_{\text{obs}} \mid \bar{d}$, we start from the Jukes–Cantor micro-evolutionary process considered over the duration Δ with the substitution parameter ν (for the sake of simplification, ν does not coincide with μ parameterizing

the kernel smoother arising in the pseudo-likelihood). This process leads to a number of substitutions that is binomial with size L (the size of the sequence fragment) and with success probability $p = (3/4)(1 - \exp(-4\nu\Delta))$. As indicated above,

$$\begin{aligned}\Delta &= (T_{\text{obs}} - T_{\text{inf}}) + (T_{\text{obs}}^{(0)} - T_{\text{inf}}) \\ &= \Delta_{\text{obs}} + 2(T_{\text{obs}}^{(0)} - T_{\text{inf}}) \\ &= \Delta_{\text{obs}} + 2\delta,\end{aligned}$$

where $\delta = (T_{\text{obs}}^{(0)} - T_{\text{inf}})$ is the signed duration of the period between the observation time of the source and the infection time of the recipient. Therefore, we get:

$$\Delta_{\text{obs}} = -2\delta - \frac{1}{4\nu}\log\left(1 - \frac{4}{3}p\right). \tag{S2}$$

By plugging in this equation the approximation $\bar{d}/L$ of the probability p, and by adding a noise term to account for this replacement, Equation (S2) becomes:

$$\Delta_{\text{obs}} = -2\delta - \frac{1}{4\nu}\log\left(1 - \frac{4\bar{d}}{3L}\right) + \epsilon, \tag{S3}$$

where we assume that the noise ϵ follows a centered Gaussian distribution with variance σ^2. Equation (S3) corresponds to a linear regression of the response variable Δ_{obs} with respect to the explanatory variable $\bar{d}$, and parameters (δ, ν, σ) are simply estimated from training data (like (α_1, α_2)) in the framework of linear regression.

Thus, we obtain the following expression for the penalization:

$$P_\theta(\mathbf{S}, \mathbf{S}^{(0)}, T_{\text{obs}}, T_{\text{obs}}^{(0)}) = \left\{\phi\left(\Delta_{\text{obs}} \mid \bar{d}\,; \delta, \nu, \sigma^2\right)\gamma\left(\bar{d}\,; \alpha_1, \alpha_2\right)\right\}^\theta,$$

where ϕ is the Gaussian p.d.f. corresponding to Equation (S3), γ is the Gamma p.d.f. corresponding to Equation (S1) and, in practice, $(\delta, \nu, \sigma, \alpha_1, \alpha_2)$ are replaced by their estimates obtained from training data.

Horse's ID	Sampling date	Number of sampled sequences	Time of the start of infectious period	Time of the end of infectious period
A01	8	12	5	20
N34	37	9	34	43
B03	20	5	17	26
B02	20	5	17	26
C04	21	12	18	27
E07	22	4	19	28
D05	22	7	19	28
D06	22	2	19	28
E10	23	25	20	29
F11	23	4	20	29
E09	23	10	20	29
G08	23	11	20	29
E13	26	11	23	32
D12	26	6	23	32
E14	26	2	23	32
E15	26	3	23	32
I17	27	8	24	33
E18	27	7	24	33
H16	27	8	24	33
E19	27	6	24	33
D20	28	9	25	34
K22	28	14	25	34
J21	28	2	25	34
H24	30	15	27	36
H23	30	11	27	36
L25	34	10	31	40
M29	34	9	31	40
L27	34	11	31	40
N28	34	11	31	40
M31	35	7	32	41
M32	36	18	33	42
O33	37	9	34	43
Q36	40	13	37	46
P35	40	3	37	46
L30	34	1	31	57
N37	40	19	37	46
R38	40	14	37	46
L39	41	5	38	47
L40	42	7	39	64
L43	45	3	42	51
L44	45	1	42	51
L42	45	26	42	51
L47	47	1	42	51
V46	45	17	42	51
T48	48	11	45	54
N45	45	8	42	51
U49	54	8	51	60
W50	64	11	61	70

Table S1: Information about the 48 confirmed Equine influenza horses.

Penalization shape	Temporal constraint 1	2	3
H1-χ^2	5	6	6
H2-normal	5	6	6
H3-gamma.LM	6	6	6

Table S2: Number of recipient-source pairs in the training data set inferred as the most likely pairs, with the three penalization shapes, the three temporal constraints and a tolerance $\eta = 1$.

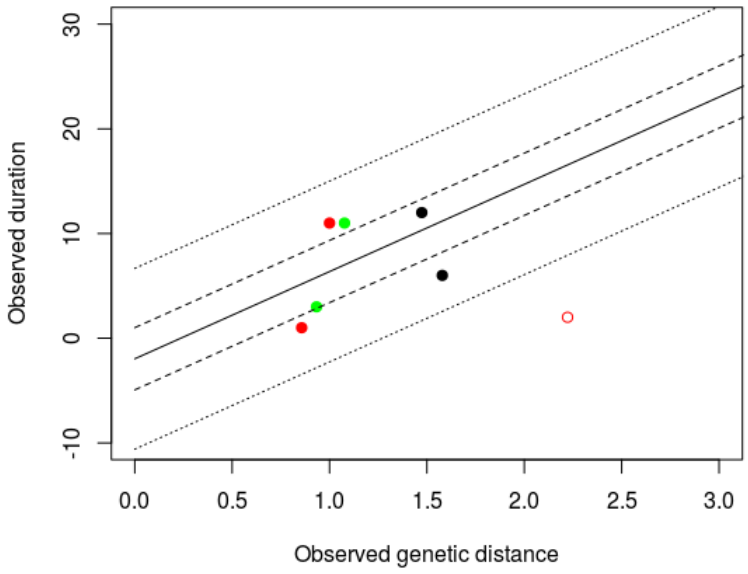


Figure S1: Linear regression given by Equation S3 describing the relationship between the duration Δ_{obs} between observation times and the genetic distance $\bar{d}$ between sets of sequences collected from the source and the recipient. The regression was fitted to all training pairs except (M32,L25). Solid line: prediction; Dashed and dotted lines: pointwise 25% and 95% confidence intervals, respectively. Green and black points: training recipient-source pairs ranked first (black points correspond to the training host N37 having two traced sources). Red points: training pairs inferred as likely by `SLAFEEL` but not ranked first. Red circle: training pair (M32,L25) not likely according to `SLAFEEL`.

C. Appendix of the first article of Chapter 4

,

Supporting material

Characterizing viral within-host diversity in fast and non-equilibrium demo-genetic dynamics

Alamil M.[1,*], Thébaud G.[2], Berthier K.[3], and Soubeyrand S.[1]

[1]INRAE, BioSP, 84914 Avignon, France
[2]BGPI, Univ Montpellier, INRAE, CIRAD, Institut Agro, Montpellier, France
[3]INRAE, Pathologie Végétale, 84140, Montfavet, France
[*]Corresponding author: maryam.alamil@inrae.fr

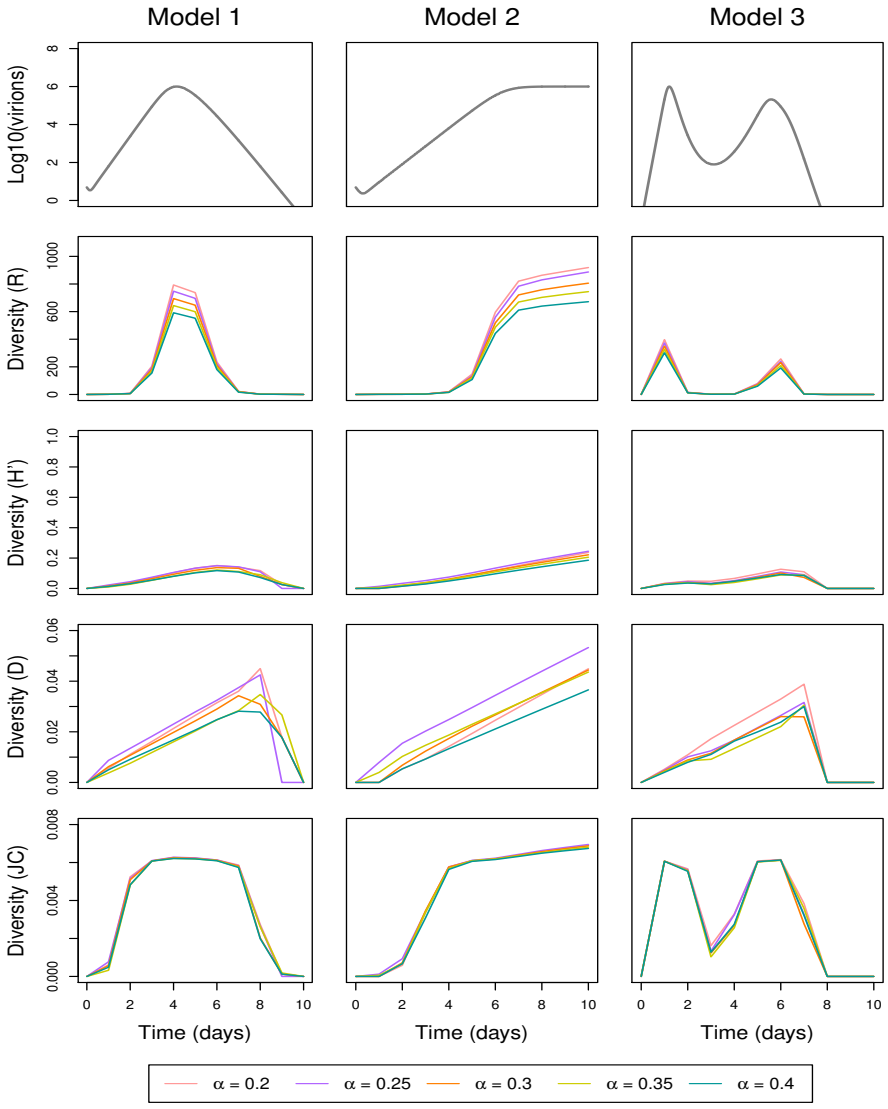


Figure S1: Simulated effects of the proportion α of lethal mutations on the levels of within-host genetic diversity. Simulations performed without the shuffling process. Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2-5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the proportion α of lethal mutations.

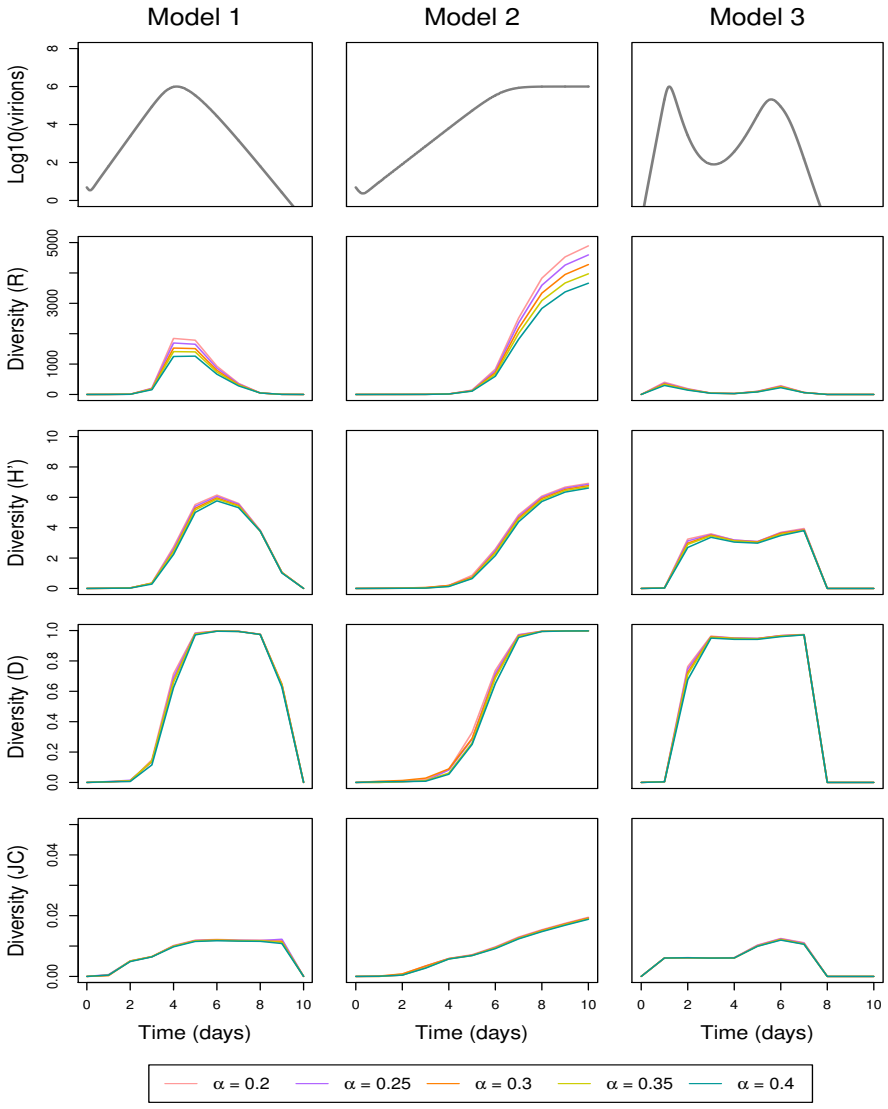


Figure S2: Simulated effects of the proportion α of lethal mutations on the levels of within-host genetic diversity. Simulations performed with the shuffling process. Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2-5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the proportion α of lethal mutations.

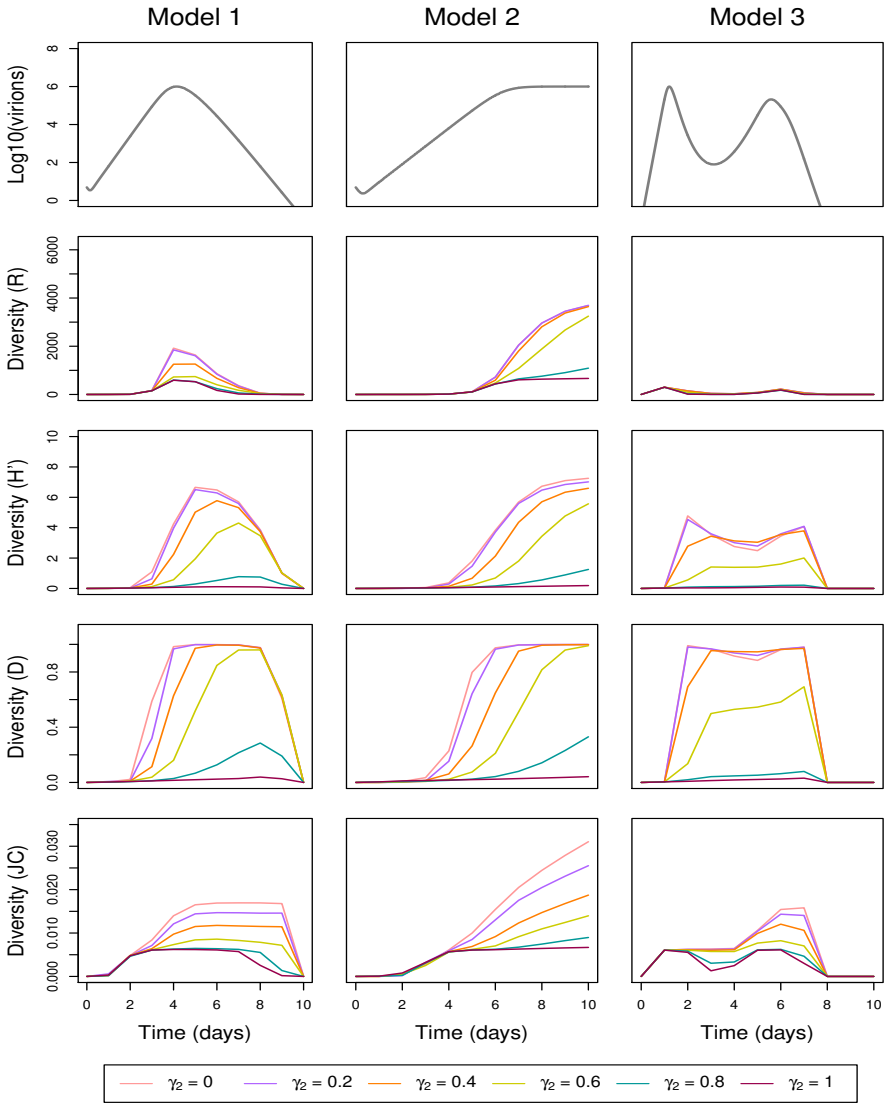


Figure S3: Simulated effects of the shuffling parameter γ_2 on the levels of within-host genetic diversity. Simulations performed with the elimination of lethal genomes ($\alpha = 0.4$). Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2-5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the shuffling parameter γ_2.

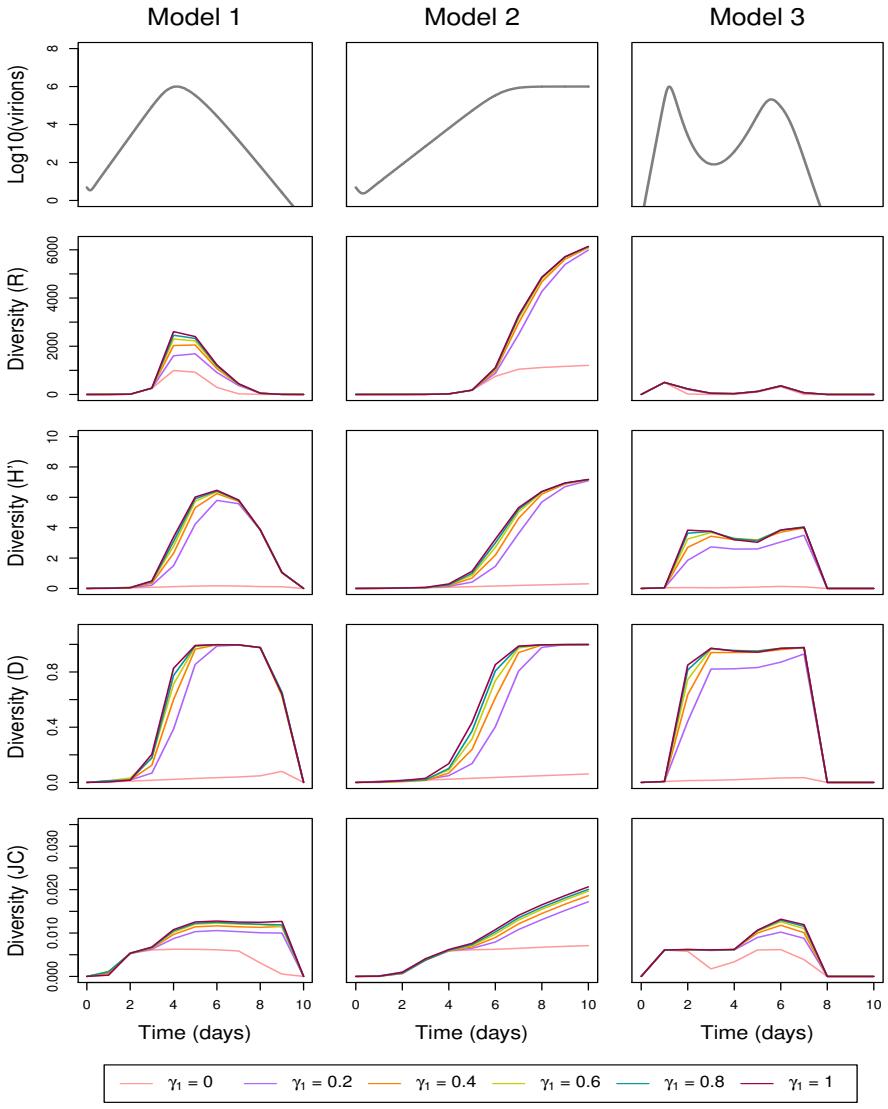


Figure S4: Simulated effects of the shuffling parameter γ_1 on the levels of within-host genetic diversity. Simulations performed without the elimination of lethal genomes. Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2-5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the shuffling parameter γ_1.

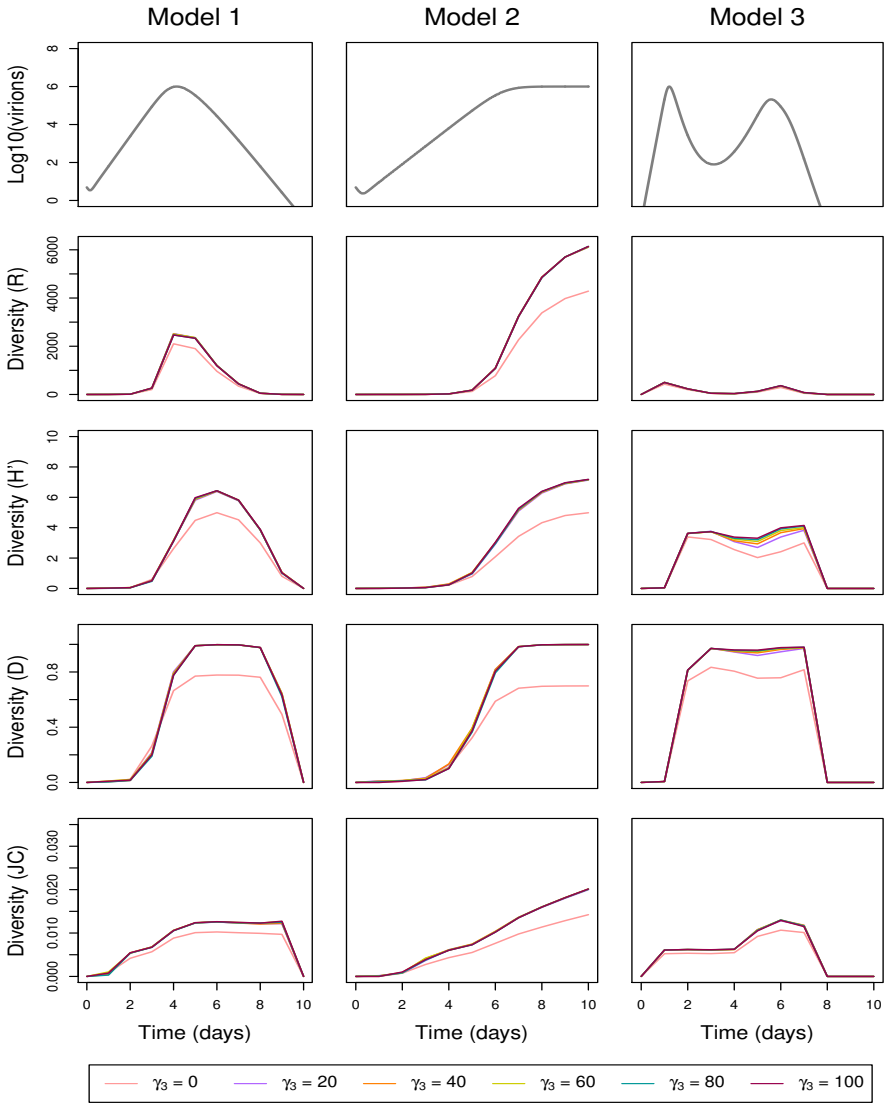


Figure S5: Simulated effects of the shuffling parameter γ_3 on the levels of within-host genetic diversity. Simulations performed without the elimination of lethal genomes. Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2-5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the shuffling parameter γ_3.

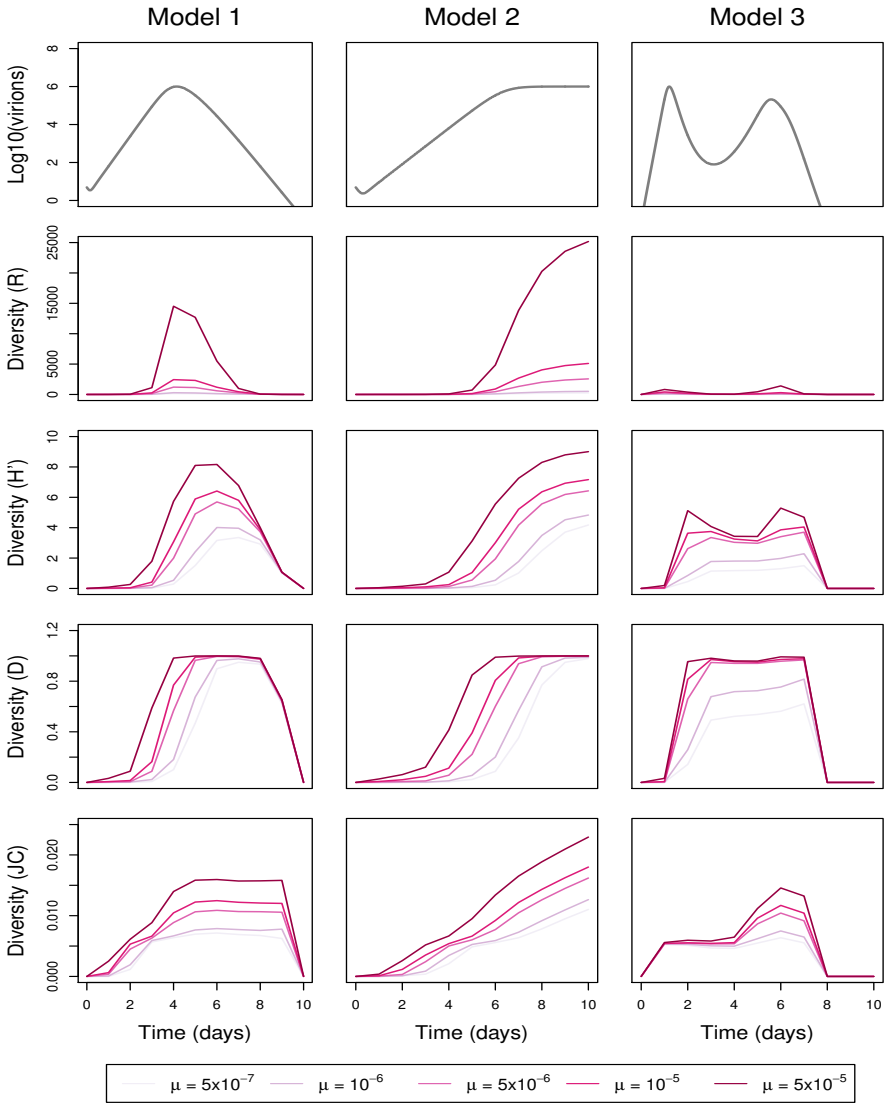


Figure S6: Simulated effects of the mutation rate μ on the levels of within-host genetic diversity. Simulations performed with the shuffling process but without the elimination of lethal genomes (with $\gamma_2 = 0.4$ and $\alpha = 0$). Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2-5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the mutation rate μ.

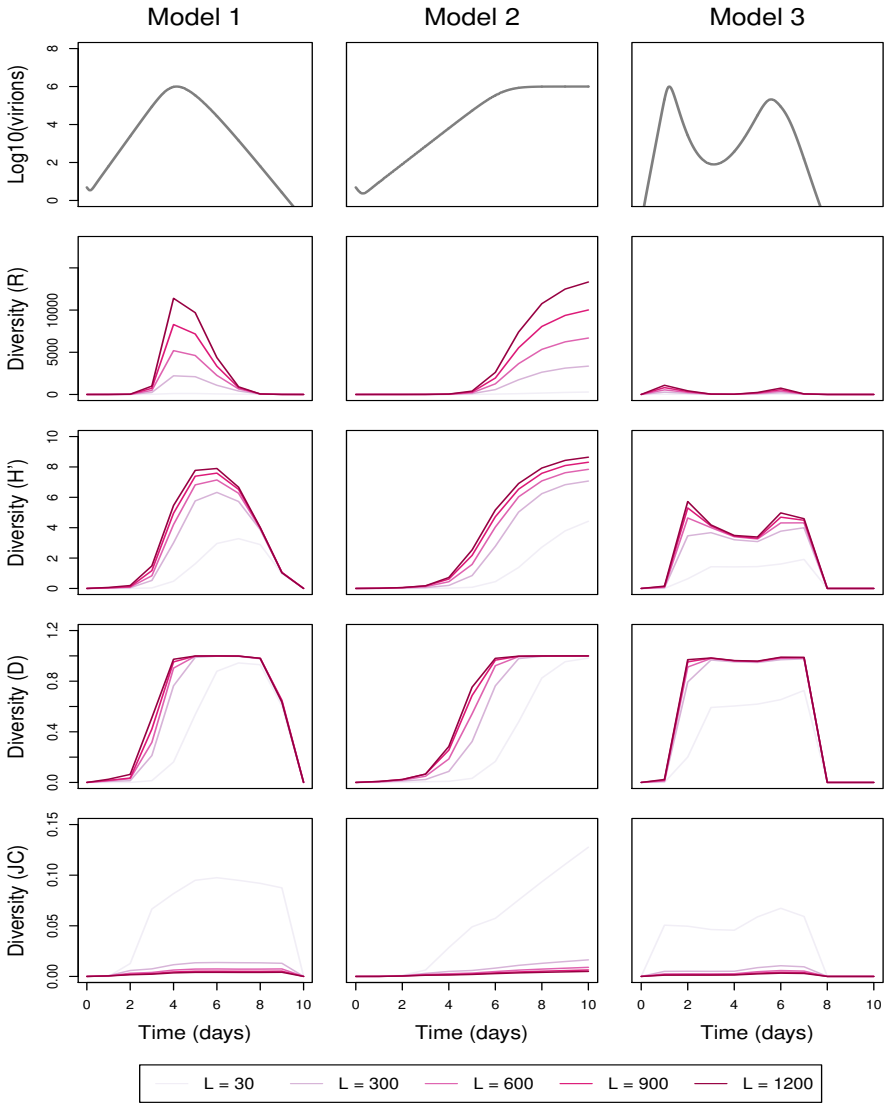


Figure S7: Simulated effects of the genetic sequence size L on the levels of within-host genetic diversity. Simulations performed with the shuffling process but without the elimination of lethal genomes (with $\gamma_2 = 0.4$ and $\alpha = 0$). Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2-5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of of the genetic sequence size L.

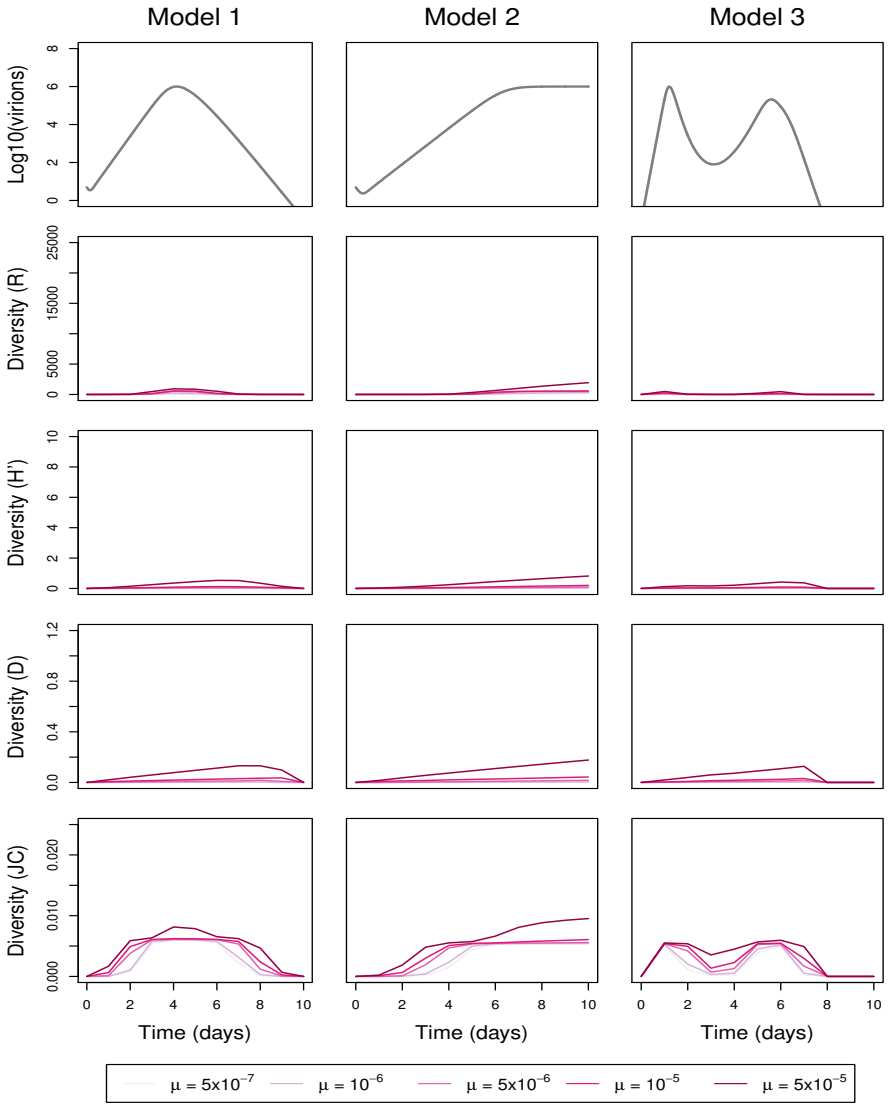


Figure S8: Simulated effects of the mutation rate μ on the levels of within-host genetic diversity. Simulations performed without the shuffling process but with the elimination of lethal genomes (with $\alpha = 0.4$). Row 1: changes in within-host virion quantity predicted respectively by the three different kinetic models mentioned in Section 2.1. Rows 2-5: variation in within-host genetic diversity assessed by the four indices presented in Section 2.3 during 10 days for various values of the mutation rate μ.

S1 Supporting text: Specific study of the shuffling process

Our model includes a component that accounts for episodes of natural selection and strong genetic drift by enhancing, e.g., the growth of low-frequency variants. This component, called shuffling process (Section 2.2.3), is constructed by noising the variant proportions P with the Gaussian distribution $\mathcal{N}(P, \gamma_1 \times P^{\gamma_2} \times (1-P)^{\gamma_3})$ of mean P and variance depending on P and three non-negative parameters γ_1, γ_2 and γ_3, by applying a two-side cut-off $\min\{1, \max\{0, \cdot\}\}$, and then by dividing the resulting variables by their sum to get a vector of probabilities. To gain insight into the impact of the shuffling parameters on the variant proportions, we apply the noising process to a set of 100 probabilities (summing to one) obtained by drawing a vector from the Dirichlet distribution with rate vector $\alpha = (\alpha_1, ..., \alpha_{100})$, using $\alpha_1 = ... = \alpha_{80} = 0.02$ and $\alpha_{81} = ... = \alpha_{100} = 0.06$.

Figure S9 displays the realization of the Dirichlet draw and shows how the proportions of the dominant variants decrease and the weights of a few low-proportion variants increase by noising the vector of probabilities with $(\gamma_1, \gamma_2, \gamma_3) = (0.8, 0.4, 70)$. Figure S10 displays the evolution of the noise variance with respect to the initial probability values generated with the Dirichlet distribution. The variance is maximum for intermediate values around 0.004. Then, we applied the shuffling process 10000 times to the same vector of Dirichlet probabilities (by taking into account the two-side cut-off and the division by the sum) for assessing the distributions of the noisy versions of the maximum probability ($p_{\max}$; i.e. maximum of the set of probabilities generated with Dirichlet distribution), the minimum positive probability ($p_{\min}$; i.e. minimum of the set of probabilities generated with Dirichlet distribution) and the probability corresponding to the maximum variance ($p_{\mathrm{var}_{\max}}$; i.e. the probability of the maximum variance value in Figure S10). These distributions are displayed in Figure S11 and allow us to visualize how variable the probabilities are in one step (i.e., one generation). For evaluating the potential evolution of probabilities throughout 10 generations, we sequentially repeated the noising process 10 times. In other words, starting with the initial set of probabilities generated with Dirichlet distribution, at each generation we applied the shuffling process to the noisy probabilities of the previous generation (by taking into account the two-side cut-off and the division by the sum). This process was repeated 10000. Figure S12 gives the evolution of the maximum probability, the minimum positive probability and the probability that has undergone the greatest variation. We noticed a decrease of the maximum probability value. In contrast, an increase in the minimum and and the maximum variation probability values was observed.

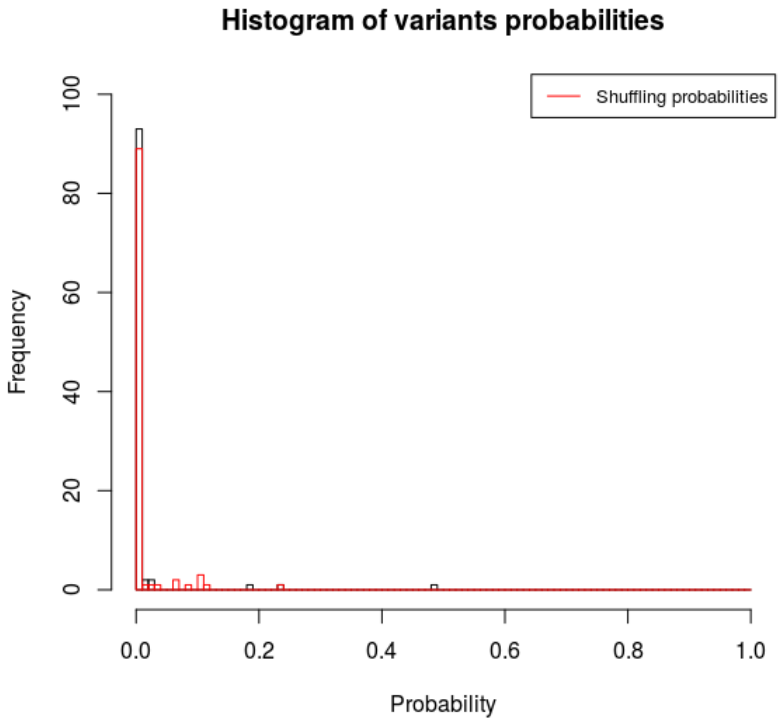


Figure S9: Distribution of initial and noisy probabilities.

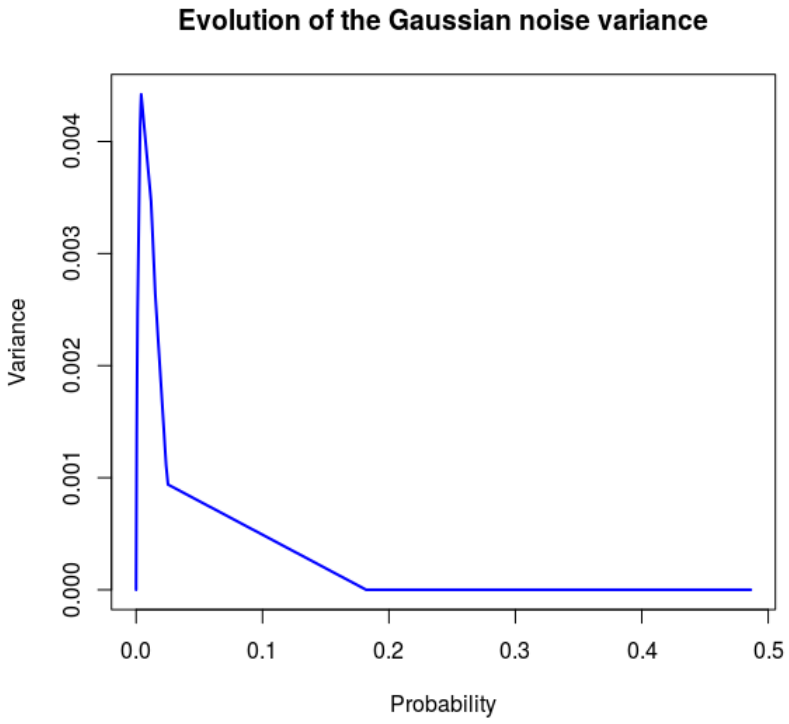


Figure S10: Evolution of the Gaussian noise variance with respect to the initial probability values generated with the Dirichlet distribution.

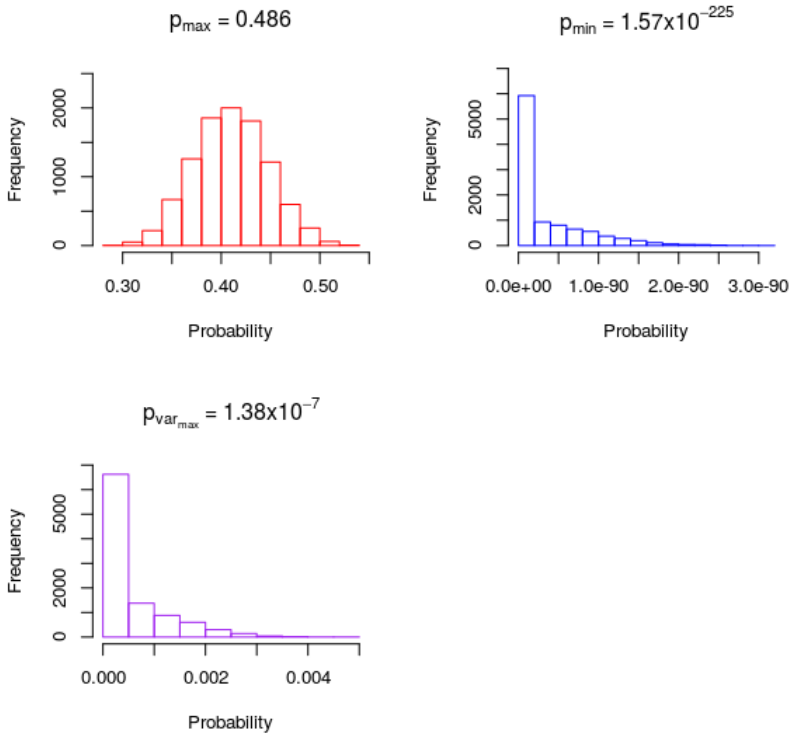


Figure S11: Distribution of the noisy version of the maximum probability distribution (red), the minimum positive probability (blue), and the probability corresponding to the maximum noise variance (purple).

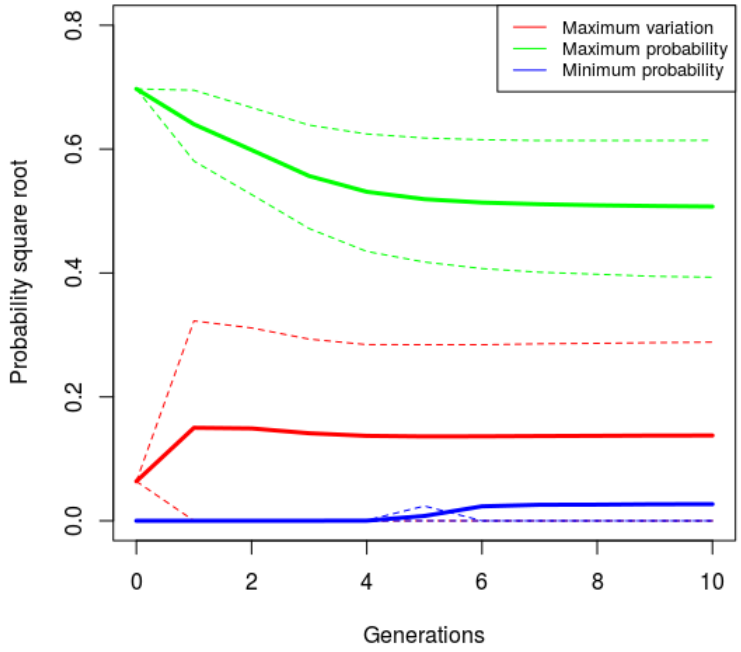


Figure S12: Evolution of the probabilities during 10 generations by applying the shuffling process. The red, green and blue lines show, respectively, the average evolution of the maximum probability, the minimum positive probability and the probability that has undergone the greatest variation. The dashed lines give the pointwise 95% confidence envelopes.

D. Appendix of the second article of Chapter 4

,

Supporting material

Factors influencing the inference of transmission events in disease outbreaks

Alamil M.[1,2,*], Bruchou C.[1], Ribaud M.[1,], Thébaud G.[3], and Soubeyrand S.[1]

[1]INRAE, BioSP, 84914 Avignon, France
[2]LMA, Université d'Avignon, 84140, AVIGNON
[3]BGPI, INRAE, Univ Montpellier, Cirad, Institut Agro, Montpellier, France
[*]Corresponding author: maryam.alamil@inrae.fr

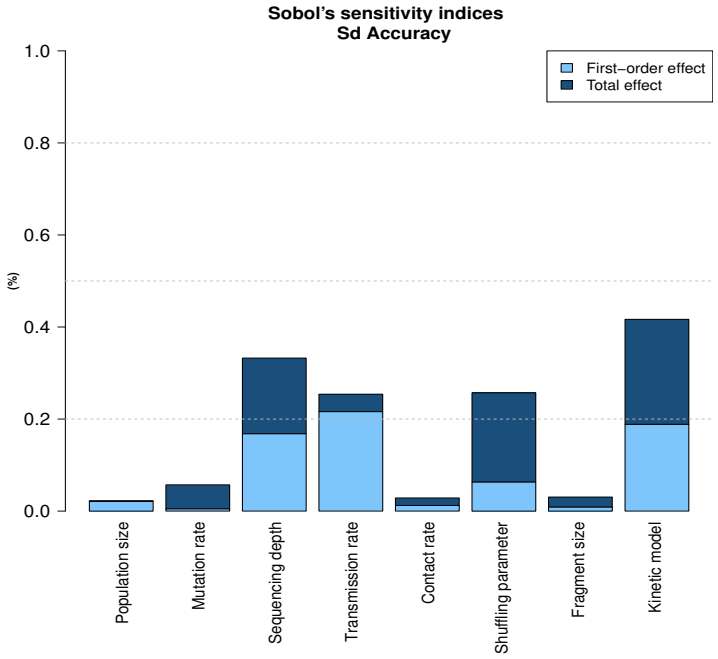


Figure S1: First-order and total Sobol indices for the seven input parameters and the kinetic model modality with respect to the standard deviation of `SLAFEEL`.

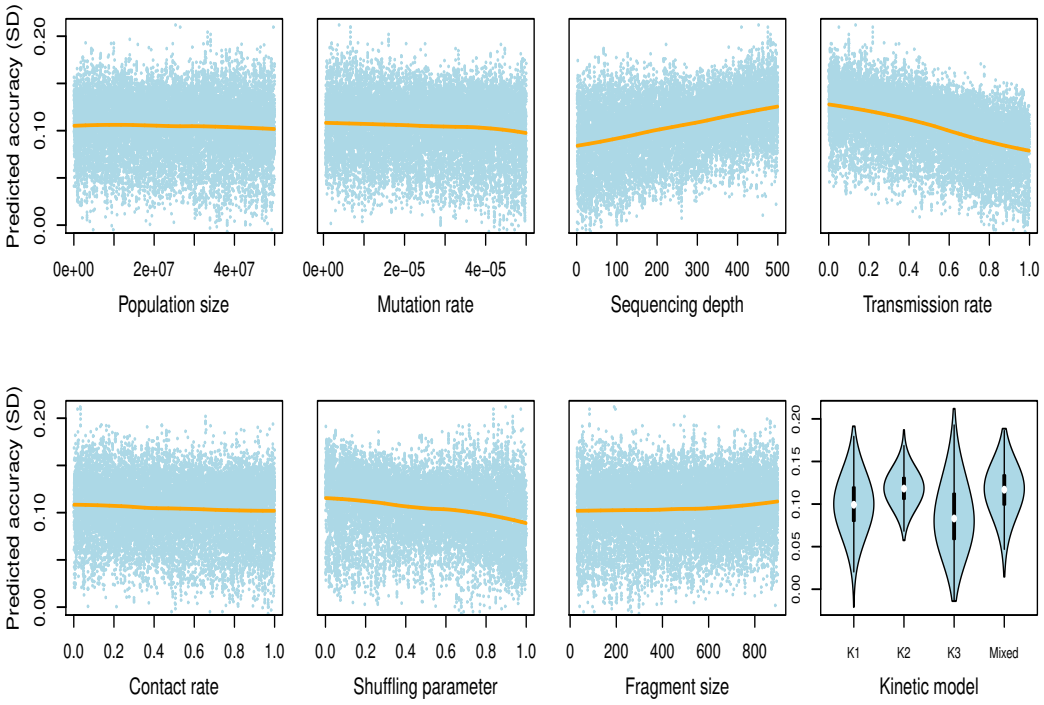


Figure S2: Relationship between input parameters and the predictions of the standard deviation of `SLAFEEL` accuracy. Each point in the scatter plots represents the standard deviation of accuracy predicted by the generalized linear meta-model. Solid orange line: local 2-degree polynomial regression (LOESS: LOcally weighted Scatterplot Smoother). Last panel: violin plot representing the effect of the kinetic model (qualitative variable) on the standard deviation of mean accuracy.

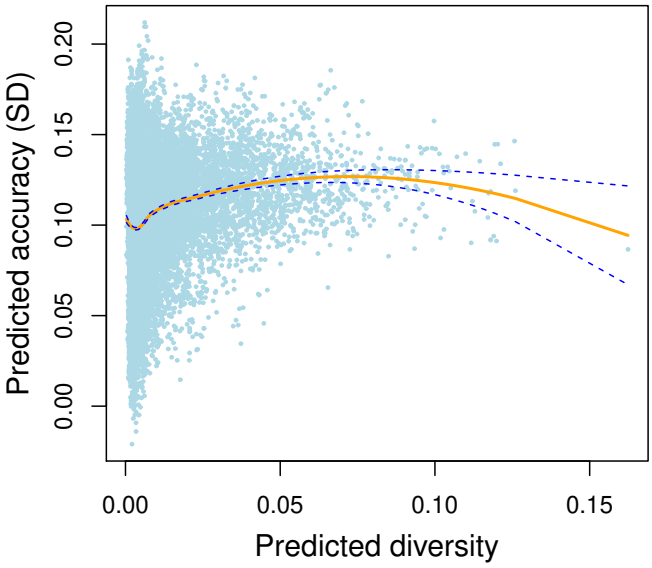


Figure S3: Predicted standard deviation of `SLAFEEL` accuracy versus predicted total Jukes Cantor diversity. The solid orange line is a smoother and the dotted blue lines represent its corresponding 95% prediction interval.

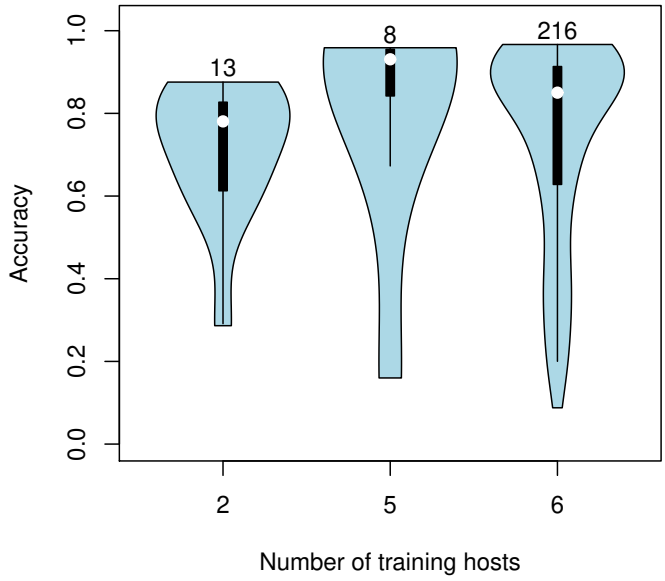


Figure S4: Variation in the performance of `SLAFEEL` against the rounded average number of training hosts used for the penalization calibration (the average is taken over the 20 replicates for each parameter combination). The overhead numbers are the numbers of parameter combinations constituting the different violin distributions.

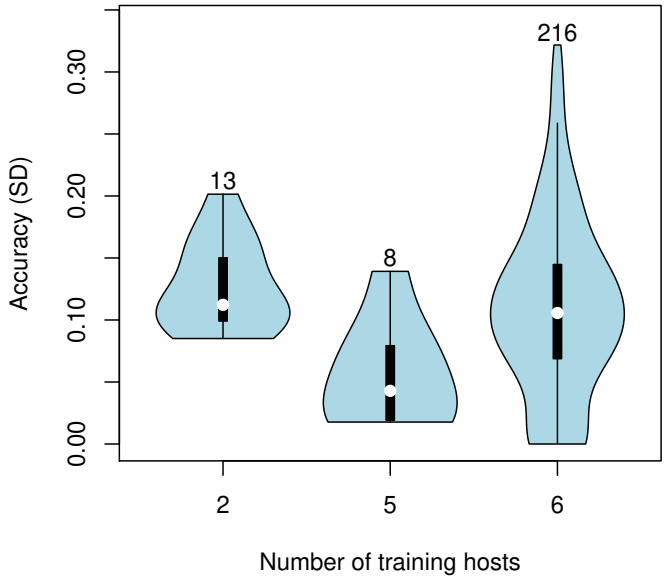


Figure S5: Variation in the the standard deviation of `SLAFEEL` accuracy against the rounded average number of training hosts used for the penalization calibration (the average is taken over the 20 replicates for each parameter combination). The overhead numbers are the numbers of parameter combinations constituting the different violin distributions.