

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

Maryam Alamil, Gaël Thébaud, Karine Berthier, Samuel Soubeyrand

▶ **To cite this version:**

Maryam Alamil, Gaël Thébaud, Karine Berthier, Samuel Soubeyrand. Characterizing viral within-host diversity in fast and non-equilibrium demo-genetic dynamics. Frontiers in Microbiology, 2022, 13 (983938), 10.3389/fmicb.2022.983938 . hal-03831203

HAL Id: hal-03831203
https://hal.inrae.fr/hal-03831203

Submitted on 27 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 05 October 2022
DOI 10.3389/fmicb.2022.983938

Check for updates

OPEN ACCESS

EDITED BY
Esteban A. Hernandez-Vargas,
University of Idaho, United States

REVIEWED BY
Van Kinh Nguyen,
Imperial College London,
United Kingdom
Lubna Pinky,
Eastern Virginia Medical School,
United States

*CORRESPONDENCE
Maryam Alamil
maryam.alamil@hotmail.com;
malamil@alfaisal.edu
Samuel Soubeyrand
samuel.soubeyrand@inrae.fr

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 04 July 2022
ACCEPTED 08 September 2022
PUBLISHED 05 October 2022

CITATION
Alamil M, Thébaud G, Berthier K and Soubeyrand S (2022) Characterizing viral within-host diversity in fast and non-equilibrium demo-genetic dynamics. *Front. Microbiol.* 13:983938. doi: 10.3389/fmicb.2022.983938

COPYRIGHT
© 2022 Alamil, Thébaud, Berthier and Soubeyrand. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

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

Maryam Alamil[1,2]*, Gaël Thébaud[3], Karine Berthier[4] and Samuel Soubeyrand[1]*

[1]INRAE, BioSP, Avignon, France, [2]Department of Mathematics and Computer Science, Alfaisal University, Riyadh, Saudi Arabia, [3]PHIM Plant Health Institute, INRAE, Univ Montpellier, CIRAD, Institut Agro, IRD, Montpellier, France, [4]INRAE, Pathologie Végétale, Montfavet, France

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. The performance of such characterizations and inferences cannot be analytically assessed in general and are often grounded on computer-intensive evaluations. Then, being able to simulate within-host genetic diversity across time under various demo-genetic assumptions is paramount to assess the performance of the approaches of interest. In this context, we built an original model that can be simulated to investigate the temporal evolution of genotypes and their frequencies under various demo-genetic assumptions. The model describes the growth and the mutation of genotypes at the nucleotide resolution conditional on an overall within-host viral kinetics, and can be tuned to generate fast non-equilibrium demo-genetic dynamics. We ran simulations of this model and computed classic diversity indices to characterize the temporal variation of within-host genetic diversity (from high-throughput amplicon sequences) of virus populations under three demographic kinetic models of viral infection. Our results highlight how demographic (viral load) and genetic (mutation, selection, or drift) factors drive variations in within-host diversity during the course of an infection. In particular, we observed a non-monotonic relationship between pathogen population size and genetic diversity, and a reduction of the impact of mutation on diversity when a non-specific host immune response is activated. The large variation in the diversity patterns generated in our simulations suggests that the underlying model provides a flexible basis to produce very diverse demo-genetic scenarios and test, for instance, methods for the inference of transmission links during outbreaks.

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 (Biek et al., 2015; Nelson and Hughes, 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 contributed to unravel the level of genetic diversity within a single host and how it varies spatially and temporally 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 temporal changes in within-host genetic diversity can be achieved using whole genome high-throughput sequencing (HTS) approaches on serial samples from infected hosts. However, and although most RNA viruses have relatively small genome sizes, accurate whole genome sequencing of numerous samples still remains costly and time consuming (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, e.g., using the R package `dada2` (Callahan et al., 2016).

Within-host genetic diversity of viruses is of particular interest for inferring (potentially indirect) epidemiological links between hosts and even reconstructing transmission chains 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 (Brunker et al., 2012; Picard et al., 2017). Typically, these approaches used genetic-space-time relations at the multi-host level to reconstruct transmission links during outbreaks (Cottam et al., 2008; Morelli et al., 2012; Ypma et al., 2012, 2013; Jombart et al., 2014; Mollentze et al., 2014; Hall et al., 2015; 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 homogeneous environment, within which the viral population at a given time was represented by a unique sequence, such as the consensus sequence 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 (Hughes et al., 2012; Morelli et al., 2012; Murcia et al., 2012; Walker et al., 2013; Didelot et al., 2014; Jombart et al., 2014; Worby et al., 2014; De Maio et al., 2018; Leitner and Romero-Severson, 2018; Wymant et al., 2018; Alamil et al., 2019). To evaluate the performance of these approaches in numerous diverse and challenging settings, we need simulation models of viral within-host genetic diversity and tools to characterize this diversity. Here, we propose such a framework, based on the work of Worby and Read (2015) on the simulation of evolutionary and epidemiological dynamics, as well as classical viral kinetic models and widely used diversity indices. This framework was designed to possibly generate non-equilibrium fast evolutionary dynamics. Briefly, the "non-equilibrium" feature means that the system can bifurcate into new dynamic steady states (Chaisson, 2004), and the adjective "fast" indicates that such bifurcations may arise quite frequently. In the context of virology considered here, a "frequent bifurcation" is typically manifested by a change in the dominant viral genotype during the infection of a host, as observed, e.g., by Hughes et al. (2012). Multiple mechanisms related to selection and drift can drive such changes and our approach is to account for them implicitly by going beyond the binomial or multinomial draws intuitively applied to modeling genotype replication.

In our approach, the within-host virus population is simulated by generating genotypes (i.e., sequence fragments) and their proportions conditional on a demographic kinetics to be specified. 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 time points 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 may be relatively strong (Abel et al., 2015; Poirier and Vignuzzi, 2017), and a demographic kinetics described by a set of differential equations and quantifying the variation of the viral load during the course of the infection, which is represented by a set of differential equations any other mathematical formalism may be used for the demographic kinetics as soon as it provides a quantity of virions across time; e.g., see Yuan and Allen (2011), for models based on stochastic differential equations and continuous-time Markov chains. We consider three examples of kinetic models, all including a latent period, and respectively, representing an acute infection, a persistent infection and an infection mitigated by an immune response. These examples were chosen more for their ability to produce contrasting viral load dynamics than for their applicability to a specific case study.

Genetic effects incorporated into the model 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 extreme 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 relative fitness of the genotypes, which can vary during the course of the infection *via* the over-dispersion of the multinomial draws. This over-dispersion is governed by a 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). 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-genetic 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 characterization 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 influence of the parameters on temporal variations 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. The results especially highlight the major impact of the shuffling process, the non-monotonic relationship between pathogen population size and genetic diversity, and the reduction of the impact of mutation on diversity when a host immune response is activated. These elements are discussed in the last section of this article. An `R` code called `MoWPP` (Model of Within-host Pathogen Population dynamics) allowing the user to run the model and compute the diversity indices is provided at https://doi.org/10.5281/zenodo.6783246.

2. Materials and methods

The following three subsections detail the modeling framework using the mathematical formalism. The last subsection provides a concise overview of the model *via* an algorithmic description.

2.1. Kinetic models

We consider that the size of the within-host pathogen population varies over time. To quantify this temporal variation, one can use a wide range of kinetic models that were developed to study within-host dynamics of many pathogens (Perelson and Nelson, 1999; Nowak and May, 2000; Baccam et al., 2006; Beauchemin et al., 2008; Handel et al., 2010; Saenz et al., 2010; Beauchemin and Handel, 2011; Smith and Perelson, 2011; Pawelek et al., 2012; Canini and Perelson, 2014; Hernandez-Vargas, 2019), including SARS-CoV-2 (Du and Yuan, 2020; Gonçalves et al., 2020; Goyal et al., 2020; Hernandez-Vargas and Velasco-Hernandez, 2020; Pinky and Dobrovolny, 2020; Wang et al., 2020; Blanco-Rodriguez et al., 2021; Ghosh, 2021). These models are grounded on sets of ordinary differential equations (ODE) basically governing the numbers of susceptible target cells, infected cells, and virions.

For this article primarily focused on a new modeling framework coupling viral kinetics and micro-evolution, we considered three kinetic models that were chosen because they generate clearly contrasting viral-load patterns (with one peak, one plateau, and two peaks), but many other kinetic models may be considered (see Section 4.2). The three models include a latent period modeled by considering two compartments of infected cells: those not producing virions yet and those actively producing virions. The first model corresponds to acute infections (Baccam et al., 2006) and has been recently used to describe the within-host kinetics of SARS-CoV-2 with or without latent period (Abuin et al., 2020; Hernandez-Vargas and Velasco-Hernandez, 2020). The second model is a direct derivation of the first model, that allows us to transform the acute infection model into a persistent infection model (i.e., presenting a plateau) over the 10-day study period by setting the death rate of infectious cells at zero (it is viewed as a toy model). The third model is a hybrid between the model of Baccam et al. (2006) (allowing a latent period in virion production) and the model of Pawelek et al. (2012) introducing an immune response.

In this section, the state variables (S, I, V, ...) are functions of continuous-time $t \geq 0$.

2.1.1. Acute and persistent 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 acute infection model:

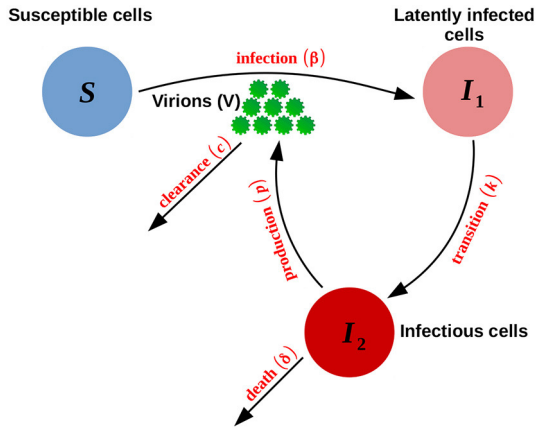


FIGURE 1
Schematic representation of the acute and persistent models defined by Equation (1). The persistent model is obtained by setting the death rate of infectious cells δ to zero.

$$\begin{cases} \frac{dS}{dt} = -\beta SV \\ \frac{dI_1}{dt} = \beta SV - kI_1 \\ \frac{dI_2}{dt} = kI_1 - \delta I_2 \\ \frac{dV}{dt} = pI_2 - cV \end{cases} \quad (1)$$

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 persistent infection (over a relatively short-time period, i.e., 10 days in our simulation study), we use the acute model of Equation (1) and we set the death rate δ of infectious cells I_2 to zero. In that respect, we assume that the infectious cells I_2 are not removed. This corresponds to both the absence of cytotoxic effects of the virus and a delay/lack of activation of the immune response against the infectious cells, resulting in negligible damage to these cells (Boldogh et al., 1996), at least over the short time period considered in our work.

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

2.1.2. Model with immune system reactions

A third model accounts for an immune response. Innate immunity through interferon (IFN) induction is modeled by adding two compartments to the acute-infection model defined by Equation (1): the IFNs (F) and the refractory uninfected cells (R). The rising adaptive immune response is modeled as an increase in the death rate of the infectious cells, δ, after an initial delay. This model, illustrated in Figure 2, is defined by:

$$\begin{cases} \frac{dS}{dt} = -\beta SV - \phi SF + \rho R \\ \frac{dI_1}{dt} = \beta SV - kI_1 - mI_1F \\ \frac{dI_2}{dt} = kI_1 - \delta I_2 - mI_2F \\ \frac{dR}{dt} = \phi SF - \rho R \\ \frac{dV}{dt} = pI_2 - cV \\ \frac{dF}{dt} = qI_2 - dF \end{cases} \quad (2)$$

where IFNs are secreted only by infectious cells I_2 at rate q and decay at rate d; upon exposure to these signaling 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:

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

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

2.1.3. Kinetic parameter values and model solving

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, 2 for the three kinetic models. Many of these values are taken from previous studies (Baccam et al., 2006; Pawelek et al., 2012), in which parameters were either fixed by the authors or 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, say V_{max}, is the same for the three different models (we use $V_{max} =$

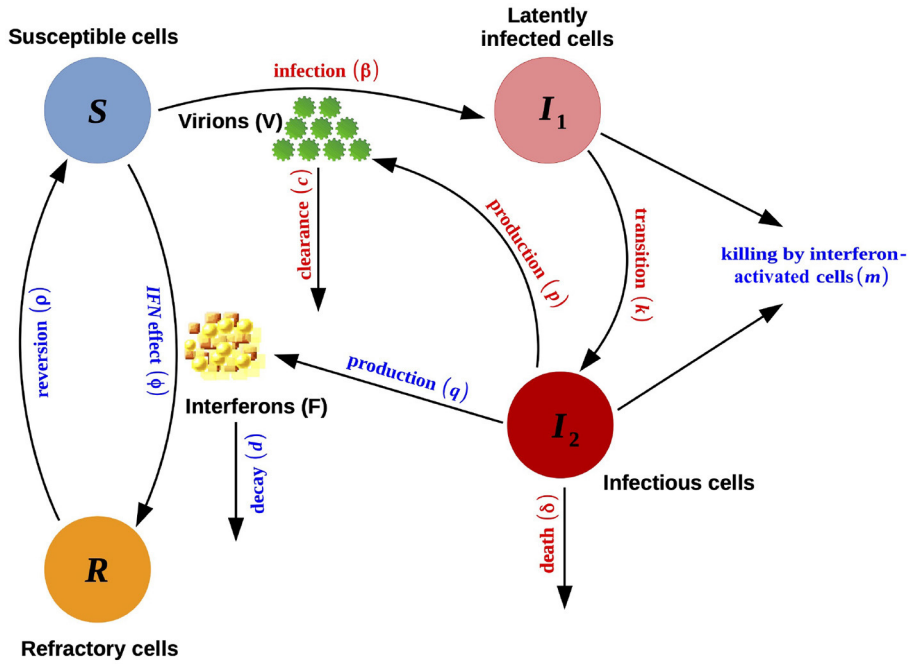


FIGURE 2
Schematic representation of the model with immune system reaction defined in Equation (2). This scheme is an edited version (with permission) of the original scheme presented by Pawelek et al. (2012).

TABLE 1 Description and values of variables and parameters used in the acute and persistent infection models described by Equation (1).

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	$TCID_{50}$/ml	Initial value: 4.9
β	Rate of susceptible target cell infection	$(TCID_{50}/ml)^{-1}.d^{-1}$	5.3×10^{-6}
k	$1/k$: 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 | 0*
p	Viral production rate	$(TCID_{50}/ml).d^{-1}$	0.05 | 0.0095*
c	Clearance rate of virions	d^{-1}	3.8

*First value for the acute infection model, second value for the persistent infection model.
In the acute infection model, initial values of S, I_1, I_2, and V and values of β, k, δ, and c are those obtained for patient 4 in Baccam et al. (2006, Table 3). The same values were used for the persistent infection model, except δ that was set to zero. In both models, p was chosen such that the maximum viral load over the 10-days study period is $V_{max} = 10^6$ virions.
d stands for day in the unit column.

10^6 virions). For each model, parameter p is computed by minimizing (with respect to p) the squared difference, $\Delta_p = (V_{max} - \bar{V}_p)^2$, 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.

The systems of ODEs can be numerically solved with the `ode` function of the `deSolve` package in R. We used a 0.001 day resolution in the applications for defining the time sequence for which values of states variables (S, I, V, ...) were computed.

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

TABLE 2 Description and values of variables and parameters used in the model with immune response described by Equation (2).

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	$TCID_{50}$/ml	Initial value: 3.5×10^{-1}
β	Rate of susceptible target cell infection	$(TCID_{50}/ml)^{-1}.d^{-1}$	8.3×10^{-6}
ϕ	Rate of the IFN-induced antiviral efficacy	$(\text{IFN fold change})^{-1}.d^{-1}$	9×10^{-4}
ρ	Reversion rate from refractory to susceptible state	d^{-1}	1.5
k	$1/k$: average transition time from I_1 to I_2	d^{-1}	0.55
δ_1	Death rate of infected cells before onset of the adaptive immune response	d^{-1}	4
t_1	Time at which the adaptive immune response starts	d	6
σ_1	Speed of the death rate increase		4
m	Killing rate of infected cells by IFN-activated NK cells	$(\text{IFN fold change})^{-1}.d^{-1}$	2.9×10^{-3}
p	Viral production rate	$(TCID_{50}/ml).d^{-1}$	4.8×10^{-3}
c	Clearance rate of virions	d^{-1}	11.5
q	Rate of IFN production	$(\text{IFN fold change}).cell^{-1}$	1.1×10^{-5}
d	Rate of IFN decay	d^{-1}	0.72

Initial values of the number of susceptible, infected and refractory cells and the value of β were taken from Pawelek et al. (2012, Tables 1, 2). The other values were chosen such that the viral load has two clearly distinct peaks over 10 days.
d stands for day in the unit column.

fast-evolving dynamics and consequently to significant changes in the viral composition.

Our model integrates several demographic and genetic factors, namely the kinetic model, the growth of genotypes, the mutations of genomes and two fitness components, namely the shuffling process and the elimination of lethal genome. In what follows, we first describe the growth and mutation stages that form the skeleton of the demo-genetic model (Sections 2.2.1, 2.2.2). Then, we present the shuffling process in Section 2.2.3, which is an "option" in our model for noising the proportions of genotypes during the growth stage (it can be used for favoring the growth of minor variants, for example, and hence favoring "frequent bifurcations"). Finally, we present the elimination of lethal genomes, which is also an "option" in our model that can be applied after the mutation stage.

In the model, 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 a viral kinetic model such as those 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 (in the literature, the reproductive cycle of viruses ranges from 8 to more than 72 h; Roizman, 1996); time t takes integer values corresponding to the generation and coinciding with integer values of time in the definition of the kinetic models provided in Section 2.1.

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 $F(t) = \{f_1(t), ..., f_{n(t)}(t)\}$ ($g_i(t)$ is the i-th genotype at time t; $f_i(t)$ is the frequency of the i-th genotype at time 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

At each time t, 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^*):

$$F'(t) \mid P^*(t-1), V(t) \sim \text{Multinomial}\left(V(t)P^*(t-1)\right) \quad (3)$$

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

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

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

the set of non-zero 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

After the growth of genotypes and the removal of those with zero-frequencies, genomes undergo a mutation stage. At this stage, the number $N_v(t)$ of mutations occurring in the genome $v \in \{1, \ldots, V(t)\}$, whose genotype is denoted by $\aleph_v \in G^*(t)$, follows a binomial distribution with size L (which is the genome length) and probability μ (which is the mutation rate 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 $v \in \mathcal{V}(t)$, $N_v(t)$ indices, noted $j_1, \ldots, j_{N_v(t)}$, are selected uniformly with replacement from $\{1, \ldots, L\}$ (drawing mutated positions with replacement allows us to take into account multiple mutations on the same nucleotide site, which means that the effective mutation rate is slightly lower than μ; note however that given the parameter values that we use in the results section, this event is extremely rare). Then, for j from j_1 to $j_{N_v(t)}$, the j-th nucleotide $\aleph_v(j)$ of genome v whose genotype is written $\aleph_v = \{\aleph_v(1), \ldots, \aleph_v(L)\}$ is updated by randomly and uniformly drawing a new nucleotide from the set {A,C,G,T}, excluding the current value of $\aleph_v(j)$. Let $\tilde{\aleph}_v$ denote the genotype obtained using this iterative procedure.

Let $\tilde{\mathcal{V}}(t)$ designate the set of genomes in $\mathcal{V}(t)$ which remain after the elimination of possible lethal genomes [see Section 2.2.4; $\tilde{\mathcal{V}}(t) = \mathcal{V}(t)$ if none of the genomes are lethal]. 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 $\{\tilde{g}_{m(t)+1}(t), \ldots, \tilde{g}_{m(t)+q(t)}(t)\} = \{\tilde{\aleph}_v : v \in \tilde{\mathcal{V}}(t)\}$ and $\tilde{g}_i(t) = g^*_i(t)$ for each $i \in \{1, \ldots, m(t)\}$, the genotype set is henceforth:

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

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

$$\tilde{F}(t) = \tilde{F^*}(t) \cup \{\tilde{f}_{\tilde{\aleph}_v} : v \in \tilde{\mathcal{V}}(t)\}$$
$$= \{\tilde{f}_1(t), ..., \tilde{f}_{m(t)+q(t)}(t)\}$$

where $\tilde{F^*}$ is the set of frequencies F^* updated 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; $\forall k = m(t)+1, ..., m(t)+q(t), f_k(t) = 1$.

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 $P(t-1)$, introduced in Equation (3). 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 eventually 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:

$$\tilde{P} \mid P \sim \mathcal{N}\left(P, \sigma^2\right) \tag{4}$$

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 re-scaling $\hat{P}$:

$$P^* = \frac{1}{\sum_{i=1}^{n} \hat{p}_i} \hat{P} \tag{5}$$

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 Supplementary Text 1. 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

The proportion of lethal mutations incurred by a viral genome lies between 0.2 and 0.4 for viruses infecting hosts from

different kingdoms (animal, plant, bacteria; Sanjuán, 2010). For vesicular stomatitis virus (VSV), an animal virus for which this value is known, the proportion of lethal mutations is 0.4 (Sanjuán et al., 2004). Hence, we account for a reference 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 re-scaled after the lethal-genome elimination step, 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:

$$H'(t) = -\sum_{i=1}^{R(t)} p_i(t) \log(p_i(t)) \tag{6}$$

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:

$$D(t) = 1 - \sum_{i=1}^{R(t)} p_i^2(t) \tag{7}$$

where $R(t)$ and $p_i(t)$ are defined as for Equation (6). This index is sensitive to abundant genotypes.

2.3.4. Jukes-Cantor distance (JC)

Pairwise indices are grounded on the comparison of the sequences of each pair of sequences. Here, we used the Jukes-Cantor distance (Jukes and Cantor, 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:

$$\bar{d}(t) = \mathbb{E}_{ij}[d(g_i(t), g_j(t))] \tag{8}$$

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 $\hat{p}_{i,j}(t) = n_{i,j}(t)/L$, $n_{i,j}(t)$ being the number of nucleotide differences between $g_i(t)$ and $g_j(t)$.

2.4. Simulation setting

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, as default values. These populations differ in the demo-genetic characteristics that are included through the kinetic model, the shuffling process and the elimination of lethal genomes. We remind that kinetic parameters are specified in Tables 1, 2. Default values of genetic parameters are specified in Table 3. This table also indicates how default values are varied for each figure displaying diversity dynamics. For each demo-genetic scenario, we performed 100 independent simulations of the temporal dynamics of the within-host population.

2.5. Backbone of the algorithmic description of the model

Here, the model is summarized *via* an algorithmic description. Details and justifications are provided in the previous subsections.

1. Set parameter values;

TABLE 3 List of parameters included in the genetic component of the model and their default values (top part of the table); values actually taken by these parameters, set of kinetic parameters and number of generations per day (G/d) for each figure displaying diversity dynamics (bottom part of the table).

Symbol	Definition	Default value
L	Length of genome fragment	330
μ	Mutation rate per nucleotide per generation	10^{-5}
γ_1	Scale parameter in the noise variance of the shuffling process	0.8
γ_2	First shape parameter in the noise variance of the shuffling process	0.4
γ_3	Second shape parameter in the noise variance of the shuffling process	70
α	Proportion of lethal mutations	0.4

Figure	L	μ	γ_1	γ_2	γ_3	α	Kinetic parameters	G/d
Figure 3	330	10^{-5}	{0.8, NA}	{0.4, NA}	{70, NA}	{0, 0.4}	Tables 1, 2	1
Supplementary Figure 1	330	10^{-5}	NA	NA	NA	(0.2, 0.4)	Tables 1, 2	1
Supplementary Figure 2	330	10^{-5}	0.8	0.4	70	(0.2, 0.4)		1
Figure 4	330	10^{-5}	0.8	(0, 1)	70	0	Tables 1, 2	1
Supplementary Figure 3	330	10^{-5}	0.8	(0,1)	70	0.4	Tables 1, 2	1
Supplementary Figure 4	330	10^{-5}	(0,1)	0.4	70	0	Tables 1, 2	1
Supplementary Figure 5	330	10^{-5}	0.8	0.4	(0,100)	0	Tables 1, 2	1
Figure 5	330	$(5 \times 10^{-7}, 5 \times 10^{-5})$	0.8	0.4	70	0.4	Tables 1, 2	1
Figure 6	(30, 1,200)	10^{-5}	0.8	0.4	70	0.4	Tables 1, 2	1
Supplementary Figure 6	330	$(5 \times 10^{-7}, 5 \times 10^{-5})$	0.8	0.4	70	0	Tables 1, 2	1
Supplementary Figure 7	(30, 1,200)	10^{-5}	0.8	0.4	70	0	Tables 1, 2	1
Supplementary Figure 8	330	$(5 \times 10^{-7}, 5 \times 10^{-5})$	NA	NA	NA	0.4	Tables 1, 2	1
Supplementary Figure 9	330	10^{-5}	{0.8, NA}	{0.4, NA}	{70, NA}	{0, 0.4}	Tables 1, 2	1
Supplementary Figure 10	330	10^{-5}	0.8	0.4	70	0.4	Tables 1, 2 + scaling to vary V_{max}	1
Supplementary Figure 11	330	10^{-6}	{0.8, NA}	{0.4, NA}	{70, NA}	{0, 0.4}	Table 1 for shared parameters	1
Supplementary Figure 12	330	10^{-5}	0.8	0.4	70	0.4	Tables 1, 2	{1, 2}

For Supplementary Figure 11, kinetic parameters that are not shared (i.e., those specific to models with persistent infection and immune response) are given in Supplementary Figure 11 caption, and μ was set at 10^{-6} to keep the computation time reasonable for the model with persistent infection whose maximum viral load was larger than 5×10^6. Unique value when the parameter is fixed; values within brackets when two values are used; values within parentheses when a range of values is used, the two values given providing the minimum and the maximum of the range.

2. Set initial states (at time, or generation, $t = 0$) of the number of virions $V(0)$, the family of genotypes $G(0)$ and the genotype frequencies $F(0)$;
3. Compute the number of virions $V(t)$ from $t = 1$ to $t = 10$ with the kinetic model;
4. For time t from 1 to 10,

Growth of genotypes conditional on the kinetics

(a) Compute previous proportions of genotypes $P(t-1) = V(t-1)^{-1}F(t-1)$;

(b) If the shuffling process is applied, add noise to $P(t-1)$ with a centered Gaussian distribution: $\tilde{P}(t-1) \mid P(t-1) \sim \mathcal{N}\left\{P(t-1), \sigma^2\right\}$, cut $\tilde{P}(t-1)$ off: $\hat{P}(t-1) = \min 1, \max\left\{0, \tilde{P}(t-1)\right\}$; and rescale $\hat{P}(t-1)$: $P^*(t-1) = (\sum_{i=1}^{n} \hat{p}_i(t-1))^{-1}\hat{P}(t-1)$;
Else keep P unchanged: $P^*(t-1) = P(t-1)$;

(c) Draw the frequency vector of the $n(t-1)$ genotypes constituting the $G(t-1)$ family after the growth stage: $F'(t) \mid P^*(t-1), V(t) \sim \text{Multinomial}\left[V(t), P^*(t-1)\right]$;

(d) Remove genotypes with zero-frequencies and update genotype frequencies accordingly; updated genotypes and genotype frequencies are denoted $G^*(t)$ and $F^*(t)$;

Mutations of genomes

(e) For genome, or virion, v from 1 to $V(t)$, draw the number of mutations $N_v(t) \sim \text{Binomial}(L, \mu)$ and let $\mathcal{V}(t) = \{v = 1, \ldots, V(t) : N_v(t) > 0\}$ denote the set of genomes undergoing at least one mutation;

(f) For genome $v \in \mathcal{V}(t)$, draw $N_v(t)$ indices from $\{1, \ldots, L\}$ and update the nucleotide corresponding to each index by drawing a new nucleotide from the set {A,C,G,T} excluding the current value of the nucleotide;

(g) If the elimination of lethal genomes is applied, discard the genomes with mutations in the first 40% of the nucleotide positions along the sequence and update genotype frequencies accordingly;
(h) Remove genotypes whose frequencies are zero, aggregate identical genotypes and update genotype frequencies accordingly; updated genotypes and their frequencies are denoted $G(t)$ and $F(t)$, respectively;

5. For time t from 0 to 10, compute the four diversity indices based on variables derived from $\{G(t), F(t)\}$.

3. Results

Table 3 provides the model parametrizations that were used for all the sets of simulations corresponding to all the figures mentioned in the results section, the parameters that vary in each figure, and the ranges of variation of these parameters.

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 and each model configuration (i.e., with/without shuffling process; with/without lethal genome elimination). 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 persistent infection model and the model with an immune response. The simulations are performed with default parameter values, namely the kinetic parameters given in Tables 1, 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.

The four 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, column 3).

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 presence of the shuffling process results in major qualitative changes in the within-host diversity measured by the H' and D indices, and to a lesser extent by the JC index. This statement can be observed in more details by comparing Supplementary Figures 1, 2, which show the temporal changes in the four diversity indices when the proportion of lethal mutations α varies between 0.2 and 0.4, either in the absence of the shuffling process (Supplementary Text 1) or in its presence (Supplementary Text 2).

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. Supplementary Figures 1, 2 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 some low-proportion genotypes may reach large proportions. Figure 4 and Supplementary Figure 3 (both displaying the effect of the variation of γ_2 but corresponding respectively to 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 indices (see Supplementary Figures 4, 5, which display the effect of the variation of γ_1 and γ_3, respectively, both in the absence of lethal-genome elimination).

3.3. Changes in the number of mutations

The proportion of mutated genomes increases with the mutation rate μ and the genome size L. By applying the shuffling process and eliminating the lethal genomes or not, Figures 5, 6 and Supplementary Figures 6–8 show that the three diversity indices R, H', and D are affected in a qualitatively similar manner by increasing μ or L (Figures 5, 6, respectively display the effect of the variation of μ and L, in the presence of the shuffling process and the elimination of lethal genomes; Supplementary Figures 6, 7 are analogous to Figures 5, 6, but do not include lethal-genome elimination; Supplementary Figure 8 is analogous to Figure 5, but does not include the shuffling process). These effects are similar to the one obtained

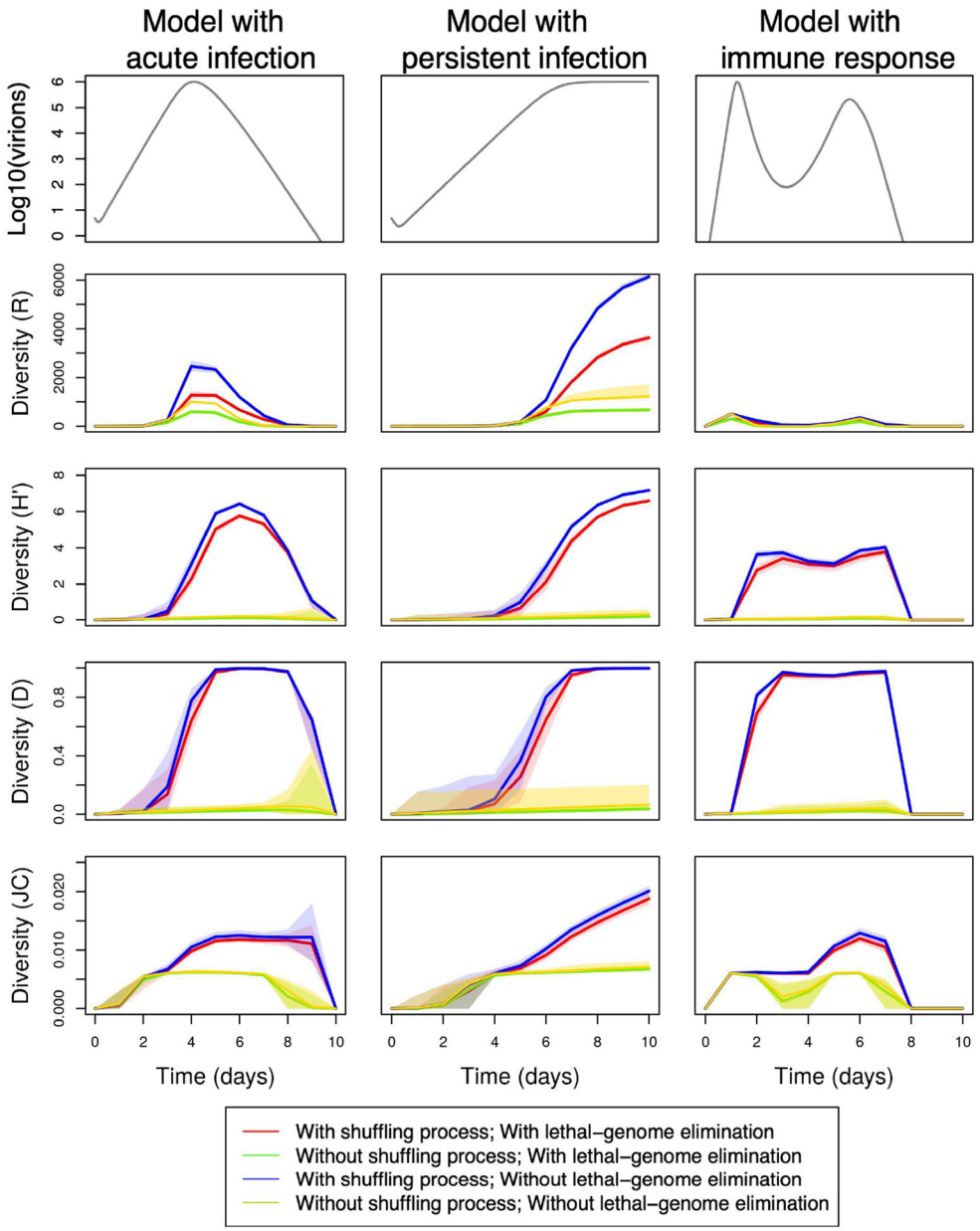


FIGURE 3
Temporal variations in within-host genetic diversity under various demo-genetic conditions, namely with or without the shuffling process [$(\gamma_1, \gamma_2, \gamma_3) = (0.8, 0.4, 70)$ or (NA, NA, NA)] and with or without the elimination of lethal genomes ($\alpha = 0.4$ or 0); See Table 3 for full parameter specification. Row 1: variation in within-host virion quantity under the models with acute infection (column 1), persistent infection (column 2), and immune response (column 3). Rows 2–5: variation in within-host genetic diversity measured by richness (R), Shannon (H'), Gini-Simpson (D), and Jukes-Cantor (JC) indices, respectively. Shaded areas delimit the 95% pointwise confidence envelopes of the diversity curves under all demo-genetic conditions.

by decreasing γ_2 (Figure 4). To refine this observation, the increased diversity of genotypes obtained for higher values of μ or L is reflected by an increased richness (R) and a faster increase of the Gini-Simpson diversity (D) up to its maximum. Even for high values of mutation rate and genome size, the immune response mitigates the replication of new genotypes, leading to a very low richness (R) and a low evenness (H').

As expected, the proportion of nucleotide differences (JC) increases with the mutation rate μ (Figure 5), while it decreases when the sequence size L increases (Figure 6), even in the absence of removal of lethal genomes (Supplementary

10.3389/fmicb.2022.983938

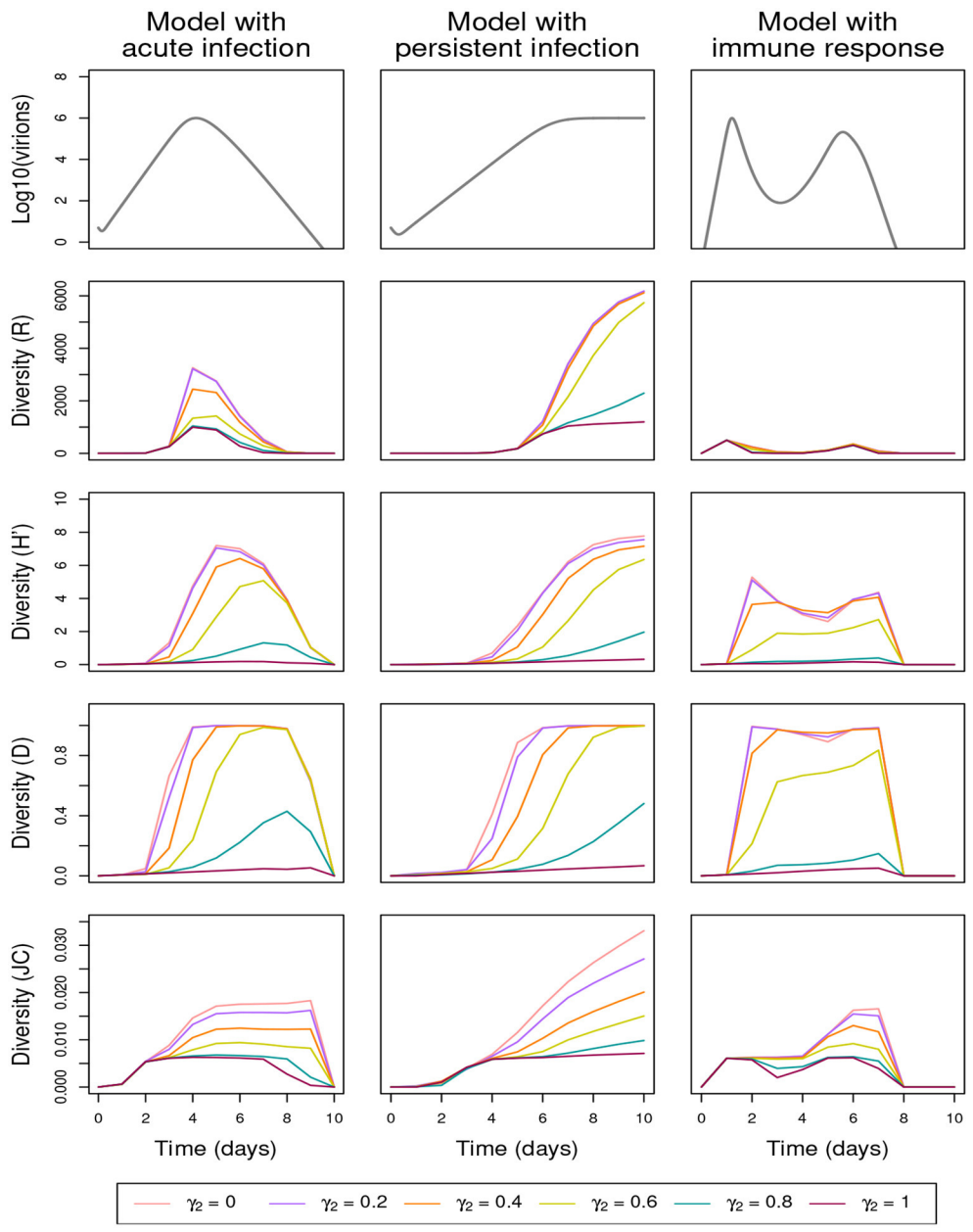


FIGURE 4
Effect of the shuffling parameter γ_2 on within-host genetic diversity in the absence of lethal-genome elimination ($\alpha = 0$), the other shuffling parameters being equal to their default values $(\gamma_1, \gamma_3) = (0.8, 70)$; See Table 3 for full parameter specification. Row 1: variation in within-host virion quantity under the models with acute infection (column 1), persistent infection (column 2), and immune response (column 3). Rows 2–5: variation in within-host genetic diversity measured by richness (R), Shannon (H'), Gini-Simpson (D), and Jukes-Cantor (JC) indices, respectively.

Figure 7). The JC index is computed among unique sequences which, for a given mutation rate, are less numerous when sequences are shorter. Thus, the JC curves corresponding to $L = 30$ reach significantly higher values than the curves obtained for longer sequences, and the following curves gradually and slightly lower from $L = 300$ to $L = 1,200$ nucleotides.

4. Discussion

4.1. Discussion of the results

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

10.3389/fmicb.2022.983938

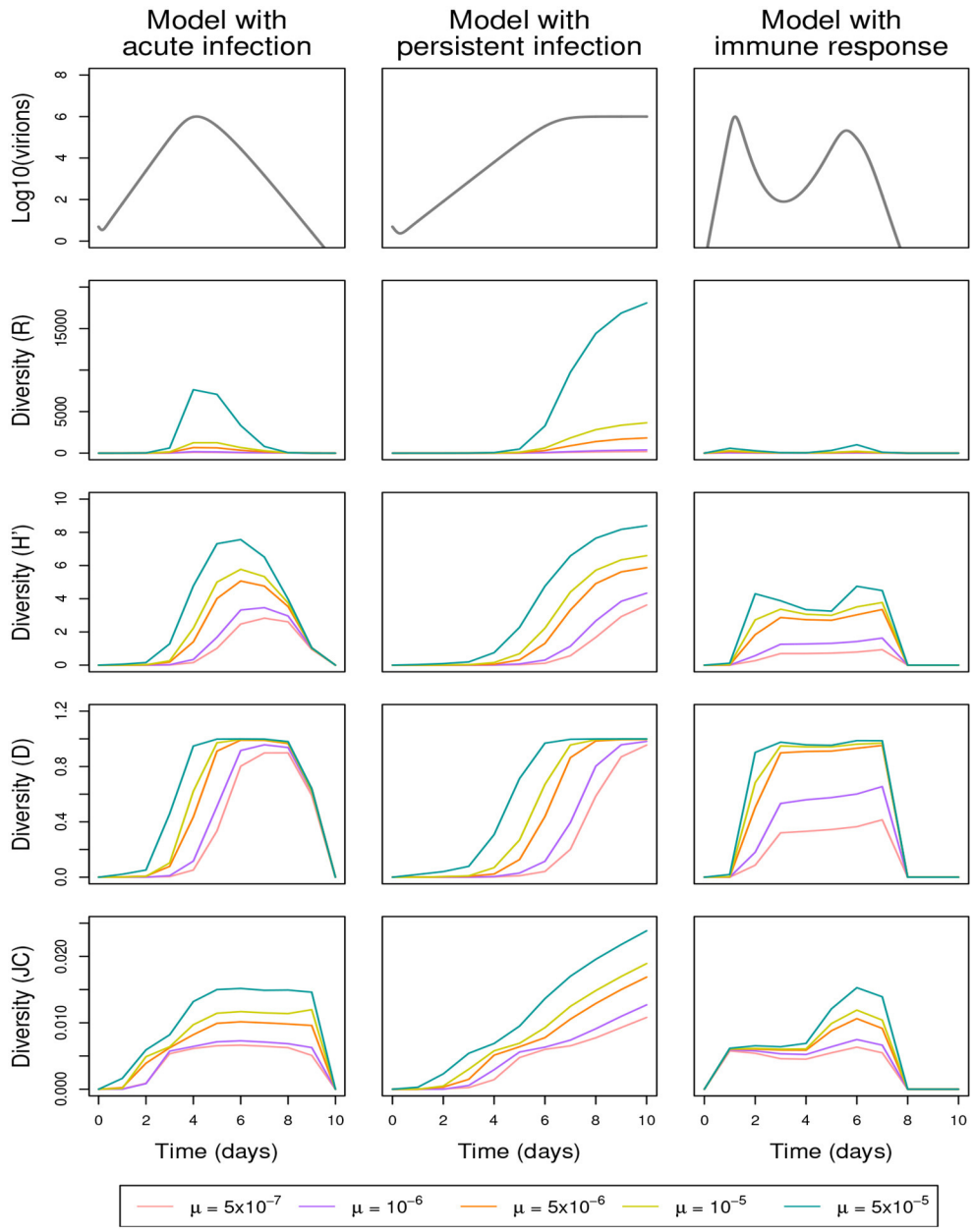


FIGURE 5
Effect of the mutation rate μ on within-host genetic diversity in the presence of the shuffling process [$(\gamma_1, \gamma_2, \gamma_3) = (0.8, 0.4, 70)$] and the elimination of lethal genomes ($\alpha = 0.4$); See Table 3 for full parameter specification. Row 1: variation in within-host virion quantity under the models with acute infection (column 1), persistent infection (column 2), and immune response (column 3). Rows 2–5: variation in within-host genetic diversity measured by richness (*R*), Shannon (*H'*), Gini-Simpson (*D*) and Jukes-Cantor (*JC*) indices, respectively. See Table 3 for full parameter specification.

within-host genetic diversity. Our explicit model developed in a forward framework allows us to monitor temporal changes (i.e., across viral 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 illustrated in Section 3. 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 processes approximately mimicking natural selection and genetic drift. These processes are based, in particular, on the elimination of lethal genomes (leading to negative selection)

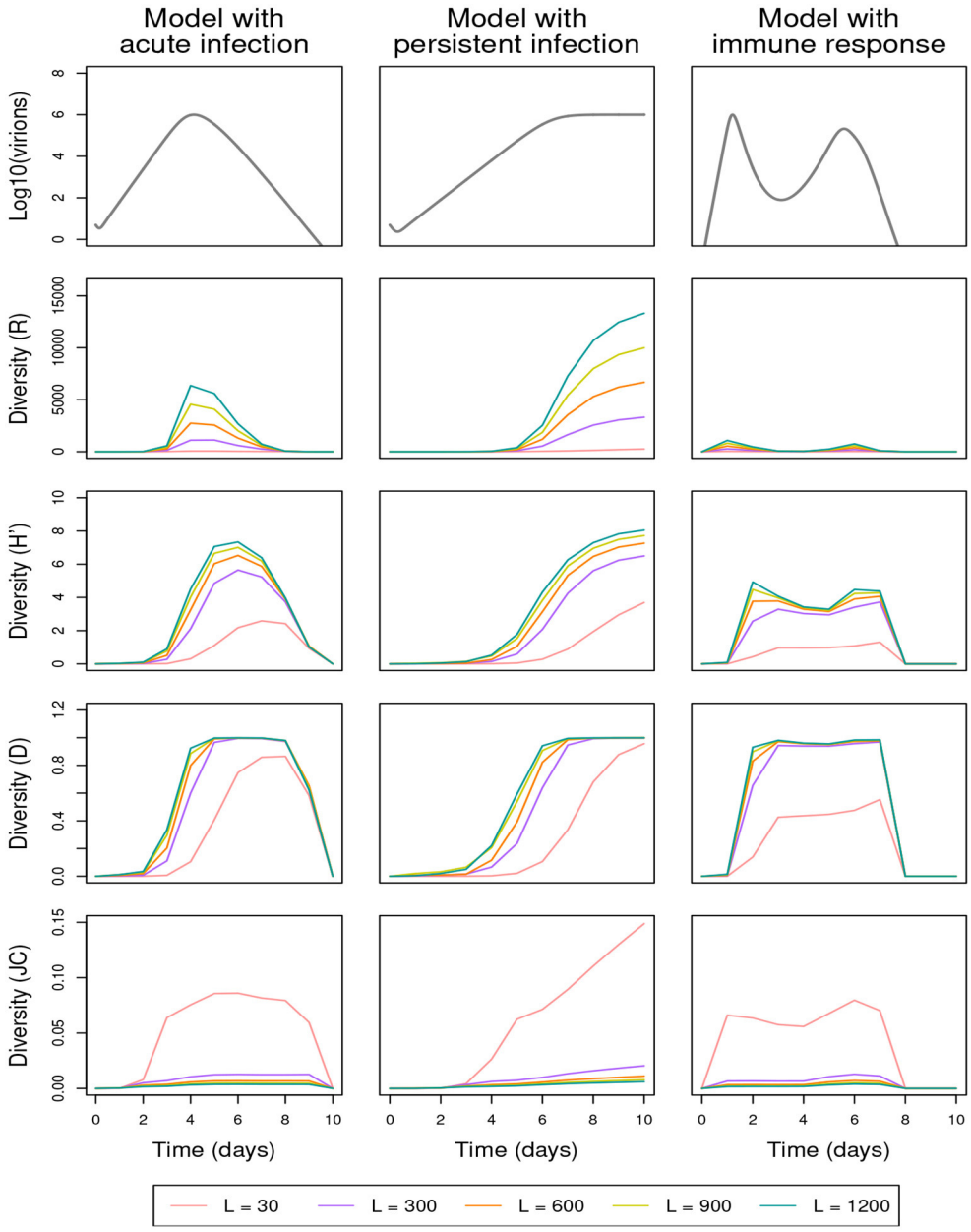


FIGURE 6
Effect of the genetic sequence size L on within-host genetic diversity in the presence of the shuffling process [$(\gamma_1, \gamma_2, \gamma_3) = (0.8, 0.4, 70)$] and the elimination of lethal genomes ($\alpha = 0.4$); See Table 3 for full parameter specification. Row 1: variation in within-host virion quantity under the models with acute infection (column 1), persistent infection (column 2), and immune response (column 3). Rows 2–5: variation in within-host genetic diversity measured by richness (R), Shannon (H'), Gini-Simpson (D), and Jukes-Cantor (JC) indices, respectively. See Table 3 for full parameter specification.

and the shuffling of genotype proportions generating over-dispersion with respect to multinomial draws (corresponding to genetic drift and positive selection). We observed a major impact of the shuffling process on within-host genetic diversity, both qualitatively and quantitatively, whatever the diversity index. There are two explanations to this observation: firstly, 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); secondly, 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. Thus, by coupling the model that we propose with a host-to-host transmission model, we could obtain a flexible basis to challenge, in very

diverse settings, the methods that reconstruct transmission trees between hosts on the basis of multiple virus sequences collected from each host (e.g., De Maio et al., 2018; Alamil et al., 2019). The host-to-host transmission model may specifically include an impact of viral load in the source host at transmission time (i) on transmission probability and/or (ii) on the initial viral load (and hence, on the subsequent dynamics) in the recipient host.

In contrast to the Wright-Fisher model considering that the total pathogen population size is constant (Fisher, 1923; Wright, 1931; Imhof and Nowak, 2006) and to the Worby and Read model (Worby and Read, 2015) assuming that the size of the pathogen population converges to an attraction function *via* the sum of binomial jumps, in our approach, virion-quantity changes during an infection are explicitly modeled (and hence controlled), and we can use many existing viral kinetic models found in the specialized literature. Predictions from the neutral theory, which highlight the importance of population size and genetic drift, provide a useful null model allowing to assess the occurrence and strength of selection on intra-host genetic diversity in rapidly evolving pathogens (Nelson and Hughes, 2015; Frost et al., 2018; Lauring, 2020). However, the succession of demographic processes (e.g., founder effect, expansion, bottleneck) and their consequences on genetic diversity during the course of an infection can be quite complex. Thus, by accounting for temporal variation in virus load (under different kinetic assumptions) and contrasting diversity indices, we can investigate the relative importance of viral load and genetic drift in shaping intra-host genetic diversity dynamics. With this in mind, we specifically observed a non-monotonic relationship between pathogen population size and genetic diversity. In Figure 3, this non-monotonic relationship is mostly exemplified by the comparison between the richness index (R) at the peak population size across the three kinetic models, or by the contrasting patterns of the different diversity indices for the model with immune response (and to a lesser extent of the JC index for the persistent infection model). Even for the acute infection model where all diversity curves seem to grossly correlate with viral load, there are delays between the peak population size and the peaks in diversity indices, and these delays often induce huge differences in diversity indices for similar viral loads, as further illustrated by Supplementary Figure 9). This may result from the complex interplay between diversity accumulation through time and changes in the size of the pathogen population. Analyzing and confronting the variations of different diversity indices in further analyses may provide some clues on the major processes shaping genetic diversity across time through main and interaction effects.

Interactions between genetic and demographic forces have been pointed out in numerous studies: pathogen population size can impact mutational robustness (Elena et al., 2007) as well as robustness to random genetic drift (Kuo et al., 2009; Didelot et al., 2016; LaBar and Adami, 2017) and selection intensity (Gutiérrez et al., 2012; Didelot et al., 2016; Frickel et al., 2018), which directly affects the composition of the viral population. Our study also illustrates such interactions (the word "interaction" being considered in its statistical meaning, i.e., the effect of a factor on a dependent variable, here a diversity index, changes according to the values of one or more other factors). Consider as an example the demographic force consisting of the immune response included in the kinetic model with immune response. The level of within-host genetic diversity and the mutation rate are known to be positively correlated (Xu et al., 2019; Castellano et al., 2020) 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). From a methodological point of view, a global sensitivity analysis and accompanying graphs (Saltelli et al., 2008; Wainwright et al., 2014) may be applied in a further study to deepen the multidimensional understanding of the covariation between parameters and diversity indices.

4.2. Perspectives in modeling

In the analysis presented in this paper, we compared diversity for three fixed viral kinetics that were standardized by setting the same value for the maximum number of virions reached over the 10-day study period (similarly, one may constrain the cumulative number of virions over 10 days to be the same). This choice allowed us to mitigate the effect of population size on diversity (which is major, as illustrated by Supplementary Figure 10) and, hence, to investigate the effect of the shape of the viral kinetics on diversity. However, beyond this objective, we aim in further studies to use our model to numerically test methods for reconstructing transmissions between hosts. In this perspective, one may simulate scenarios where the virus has intrinsic growth characteristics that are modulated from one host to another by different immunity components (Blanco-Rodriguez et al., 2021) and/or different values for parameters driving immunity strength (Hernandez-Vargas and Velasco-Hernandez, 2020). Supplementary Figure 11 somehow illustrates such a configuration at the within-host level: using the three models of Section 2.1 and setting the same values for the shared parameters, the original model components induce differences in the viral kinetics, especially in the maximum viral load, and subsequently on diversity dynamics. This is obviously a schematic setting, and host-to-host variations in the shared parameters may be included to match previous work (e.g., Baccam et al., 2006; Pawelek et al., 2012)

where parameters (even those not directly related to immunity) are separately estimated for different hosts. In addition, other kinetic models (with or without shared parameters) adapted to the disease and the population under study may deserve to be considered, as discussed below.

Our model can easily incorporate more advanced kinetic models of the number of virions and, hence, be used, e.g., to study within-host pathogen diversity in the presence of alternative immunity processes (Blanco-Rodriguez et al., 2021), antiviral treatment (Beauchemin et al., 2008; Smith and Perelson, 2011), chronic infection (Perelson and Nelson, 1999; Pinky and Dobrovolny, 2017), co-infection (Pinky and Dobrovolny, 2017, 2020), multiple target cells (Wang et al., 2020), variation of virion infectivity over time (Smith and Ribeiro, 2010; Vaidya et al., 2010; Beauchemin and Handel, 2011), co-receptor switch (Alizon and Boldin, 2010) and virion loss due to cell entry (Beauchemin et al., 2008; Gonçalves et al., 2020). We implemented a few of these models in the R code `MoWPP` accompanying this manuscript in addition to the three kinetic models presented in Section 2.1. Another perspective is to provide alternative choices in the genetic component of our hybrid model. The basic substitution model used here may notably be replaced by more realistic mutation processes (Kimura, 1980; Tavaré, 1986; Nishimaki and Sato, 2019), which would constrain the frequencies of different sequence modifications and might possibly modify the observed diversity patterns. While viral load was modeled as a continuous process (using ODEs) like most of the standard virus kinetics models found in the literature, sampling of virus genomes during replication and within-host infection was modeled as a discrete process for numerical tractability, with genotype growth and genome mutation occurring at a discrete time step (once per day). Modifying this time step while keeping the same virus kinetics impacts the resulting diversity dynamics, as illustrated by Supplementary Figure 12 where genotype growth and genome mutation happen twice a day, another biologically plausible generation time for viruses (Roizman, 1996). In this case, the level of genetic diversity increases or reaches its largest value more rapidly, because the processes that generate and increase diversity occur twice more often. Other perspectives would consist in including relative fitness depending on the genetic sequence or frequency-dependent selection (Sanjuán et al., 2004; Alizon and Boldin, 2010); note that our model, where genetics is modeled conditionally on demographics, effectively copes with variations in the relative fitness of variants, but would need to be adapted to handle variations in absolute fitness that can impact population size (i.e., the numbers of virions).

Beyond considering the way the model components are defined, models can be improved by using realistic parameter values and implementing goodness-of-fit procedures of the fitted model(s) to validate their components. The statistical estimation of the model parameters from host-level kinetic data and within-host genetic data is likely to be a challenge that first requires to identify the appropriate inference approach and level of data accuracy using models such as the one presented here. This point is further discussed in Supplementary Text 2.

Data availability statement

No original raw data were generated for this study. The publicly available datasets analyzed in this study can be found at: https://doi.org/10.5281/zenodo.6783246.

Author contributions

MA, GT, KB, and SS conceived the methodology and MA implemented it. MA, GT, KB, and SS analyzed model output, contributed to interpretation of results, and gave final approval for publication. MA prepared the initial draft of the manuscript. All authors participated to the writting of the subsequent versions. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by an ANR grant (SMITID project; ANR-16-CE35-0006).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.983938/full#supplementary-material

References

Abel, S., zur Wiesch, P. A., Davis, B. M., and Waldor, M. K. (2015). Analysis of bottlenecks in experimental models of infection. *PLoS Pathogens* 11:e1004823. doi: 10.1371/journal.ppat.1004823

Abuin, P., Anderson, A., Ferramosca, A., Hernandez-Vargas, E. A., and Gonzalez, A. H. (2020). Characterization of SARS-CoV-2 dynamics in the host. *Annu. Rev. Control* 50, 457–468. doi: 10.1016/j.arcontrol.2020.09.008

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

Alizon, S., and Boldin, B. (2010). Within-host viral evolution in a heterogeneous environment: insights into the HIV co-receptor switch. *J. Evol. Biol.* 23, 2625–2635. doi: 10.1111/j.1420-9101.2010.02139.x

Alizon, S., Luciani, F., and Regoes, R. R. (2011). Epidemiological and clinical consequences of within-host evolution. *Trends Microbiol.* 19, 24–32. doi: 10.1016/j.tim.2010.09.005

Baccam, P., Beauchemin, C., Macken, C. A., Hayden, F. G., and Perelson, A. S. (2006). Kinetics of influenza A virus infection in humans. *J. Virol.* 80, 7590–7599. doi: 10.1128/JVI.01623-05

Beauchemin, C. A., and Handel, A. (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:S7. doi: 10.1186/1471-2458-11-S1-S7

Beauchemin, C. A., McSharry, J. J., Drusano, G. L., Nguyen, J. T., Went, G. T., Ribeiro, R. M., et al. (2008). Modeling amantadine treatment of influenza A virus *in vitro*. *J. Theor. Biol.* 254, 439–451. doi: 10.1016/j.jtbi.2008.05.031

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., and Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.* 30, 306–313. doi: 10.1016/j.tree.2015.03.009

Blanco-Rodriguez, R., Du, X., and Hernandez-Vargas, E. (2021). Computational simulations to dissect the cell immune response dynamics for severe and critical cases of SARS-CoV-2 infection. *Comput. Methods Prog. Biomed.* 211:106412. doi: 10.1016/j.cmpb.2021.106412

Boldogh, I., Albrecht, T., and Porter, D. D. (1996). "Chapter 46: Persistent viral infections," in *Medical Microbiology, 4th Edn*, ed S. Baron (Galveston, TX: University of Texas; Medical Branch at Galveston).p. 585–596.

Brunker, K., Hampson, K., Horton, D., and Biek, R. (2012). Integrating the landscape epidemiology and genetics of RNA viruses: rabies in domestic dogs as a model. *Parasitology* 139, 1899–1913. doi: 10.1017/S003118201200090X

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13:581. doi: 10.1038/nmeth.3869

Canini, L., and Perelson, A. S. (2014). Viral kinetic modeling: state of the art. *J. Pharmacokinet. Pharmacodyn.* 41, 431–443. doi: 10.1007/s10928-014-9363-3

Castellano, D., Eyre-Walker, A., and Munch, K. (2020). Impact of mutation rate and selection at linked sites on DNA variation across the genomes of humans and other homininae. *Genome Biol. Evol.* 12, 3550–3561. doi: 10.1093/gbe/evz215

Chaisson, E. J. (2004). Complexity: an energetics agenda-Energy as the motor of evolution. *Complexity* 9, 14–21. doi: 10.1002/cplx.20009

Cottam, E. M., Thébaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., et al. (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. Lond. B Biol. Sci.* 275, 887–895. doi: 10.1098/rspb.2007.1442

Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J., and Sanjuán, R. (2015). Extremely high mutation rate of HIV-1 *in vivo*. *PLoS Biol.* 13:e1002251. doi: 10.1371/journal.pbio.1002251

De Maio, N., Worby, C. J., Wilson, D. J., and Stoesser, N. (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Comput. Biol.* 14:e1006117. doi: 10.1371/journal.pcbi.1006117

Didelot, X., Gardy, J., and 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

Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W., and Wilson, D. J. (2016). Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* 14:150. doi: 10.1038/nrmicro.2015.13

Domingo, E., and Holland, J. (1997). RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* 51, 151–178. doi: 10.1146/annurev.micro.51.1.151

Du, S. Q., and Yuan, W. (2020). Mathematical modeling of interaction between innate and adaptive immune responses in COVID-19 and implications for viral pathogenesis. *J. Med. Virol.* 92, 1615–1628. doi: 10.1002/jmv.25866

Elena, S. F., Wilke, C. O., Ofria, C., and Lenski, R. E. (2007). Effects of population size and mutation rate on the evolution of mutational robustness. *Evolution* 61, 666–674. doi: 10.1111/j.1558-5646.2007.00064.x

Fisher, R. A. (1923). XXI. On the dominance ratio. *Proc. R. Soc. Edinburgh* 42, 321–341. doi: 10.1017/S0370164600023993

Frickel, J., Feulner, P. G., Karakoc, E., and Becks, L. (2018). Population size changes and selection drive patterns of parallel evolution in a host-virus system. *Nat. Commun.* 9, 1–10. doi: 10.1038/s41467-018-03990-7

Frost, S. D. W., Magalis, B. R., and Kosakovsky Pond, S. L. (2018). Neutral theory and rapidly evolving viral pathogens. *Mol. Biol. Evol.* 35, 1348–1354. doi: 10.1093/molbev/msy088

Galan, M., Guivier, E., Caraux, G., Charbonnel, N., and Cosson, J.-F. (2010). A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics* 11:296. doi: 10.1186/1471-2164-11-296

Galan, M., Pages, M., and Cosson, J.-F. (2012). Next-generation sequencing for rodent barcoding: species identification from fresh, degraded and environmental samples. *PLoS ONE* 7:e48374. doi: 10.1371/journal.pone.0048374

Ghosh, I. (2021). Within host dynamics of SARS-CoV-2 in humans: modeling immune responses and antiviral treatments. *SN Comput. Sci.* 2, 1–12. doi: 10.1007/s42979-021-00919-8

Gonçalves, A., Bertrand, J., Ke, R., Comets, E., De Lamballerie, X., Malvy, D., et al. (2020). Timing of antiviral treatment initiation is critical to reduce SARS-CoV-2 viral load. *Pharmacometr. Syst. Pharmacol.* 9, 509–514. doi: 10.1002/psp4.12543

Goyal, A., Cardozo-Ojeda, E. F., and Schiffer, J. T. (2020). Potency and timing of antiviral therapy as determinants of duration of SARS-CoV-2 shedding and intensity of inflammatory response. *Sci. Adv.* 6:eabc7112. doi: 10.1126/sciadv.abc7112

Gutiérrez, S., Michalakis, Y., and Blanc, S. (2012). Virus population bottlenecks during within-host progression and host-to-host transmission. *Curr. Opin. Virol.* 2, 546–555. doi: 10.1016/j.coviro.2012.08.001

Hall, M., Woolhouse, M., and 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

Handel, A., Longini, I. M., and Antia, R. (2010). Towards a quantitative understanding of the within-host dynamics of influenza A infections. *J. R. Soc. Interface* 7, 35–47. doi: 10.1098/rsif.2009.0067

Hernandez-Vargas, E. A. (2019). *Modeling and Control of Infectious Diseases in the Host: With MATLAB and R*. Frankfurt: Academic Press.

Hernandez-Vargas, E. A. and Velasco-Hernandez, J. X. (2020). In-host mathematical modelling of COVID-19 in humans. *Annu. Rev. Control* 50, 448–456. doi: 10.1016/j.arcontrol.2020.09.006

Holmes, E. C. (2009). *The Evolution and Emergence of RNA Viruses*. Oxford: Oxford University Press.

Hughes, J., Allen, R. C., Baguelin, M., Hampson, K., Baillie, G. J., Elton, D., et al. (2012). Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS Pathogens* 8:e1003081. doi: 10.1371/journal.ppat.1003081

Imhof, L. A., and Nowak, M. A. (2006). Evolutionary game dynamics in a Wright-Fisher process. *J. Math. Biol.* 52, 667–681. doi: 10.1007/s00285-005-0369-8

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., and 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

Jukes, T. H., and Cantor, C. R. (1969). "Evolution of protein molecules," in *Mammalian Protein Metabolism*, ed H. N. Munro (New York, NY: Academic Press), 21–132. doi: 10.1016/B978-1-4832-3211-9.50009-7

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/BF01731581

Kulkarni, P., and Frommolt, P. (2017). Challenges in the setup of large-scale next-generation sequencing analysis workflows. *Comput. Struct. Biotechnol. J.* 15, 471–477. doi: 10.1016/j.csbj.2017.10.001

Kuo, C.-H., Moran, N. A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res*. 19, 1450–1454. doi: 10.1101/gr.091785.109

LaBar, T., and Adami, C. (2017). Evolution of drift robustness in small populations. *Nat. Commun*. 8, 1–12. doi: 10.1038/s41467-017-01003-7

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30, 314–334. doi: 10.1111/j.1558-5646.1976.tb00911.x

Lau, M. S., Marion, G., Streftaris, G., and Gibson, G. (2015). A systematic Bayesian integration of epidemiological and genetic data. *PLoS Comput. Biol*. 11:e1004633. doi: 10.1371/journal.pcbi.1004633

Lauring, A. S. (2020). Within-host viral diversity: a window into viral evolution. *Annu. Rev. Virol*. 7, 63–81. doi: 10.1146/annurev-virology-010320-061642

Leitner, T., and Romero-Severson, E. (2018). Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. *Nat. Microbiol*. 3, 983–988. doi: 10.1038/s41564-018-0204-9

Mollentze, N., Nel, L. H., Townsend, S., Le Roux, K., Hampson, K., Haydon, D. T., et al. (2014). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. Lond. B Biol. Sci*. 281:20133251. doi: 10.1098/rspb.2013.3251

Morelli, M. J., Thébaud, G., Chadœuf, J., King, D. P., Haydon, D. T., and Soubeyrand, S. (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol*. 8:e1002768. doi: 10.1371/journal.pcbi.1002768

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., et al. (2014). Choosing and using diversity indices: insights for ecological applications from the German biodiversity exploratories. *Ecol. Evol*. 4, 3514–3524. doi: 10.1002/ece3.1155

Murcia, P. R., Hughes, J., Battista, P., Lloyd, L., Baillie, G. J., Ramirez-Gonzalez, R. H., et al. (2012). Evolution of an Eurasian avian-like influenza virus in naive and vaccinated pigs. *PLoS Pathogens* 8:e1002730. doi: 10.1371/journal.ppat.1002730

Nelson, C. W., and Hughes, A. L. (2015). Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. *Infect. Genet. Evol*. 30, 1–7. doi: 10.1016/j.meegid.2014.11.026

Nishimaki, T., and Sato, K. (2019). An extension of the Kimura two-parameter model to the natural evolutionary process. *J. Mol. Evol*. 87, 60–67. doi: 10.1007/s00239-018-9885-1

Nowak, M., and May, R. M. (2000). *Virus Dynamics: Mathematical Principles of Immunology and Virology: Mathematical Principles of Immunology and Virology*. Oxford: Oxford University Press.

Pawelek, K. A., Huynh, G. T., Quinlivan, M., Cullinane, A., Rong, L., and Perelson, A. S. (2012). Modeling within-host dynamics of influenza virus infection including immune responses. *PLoS Comput. Biol*. 8:e1002588. doi: 10.1371/journal.pcbi.1002588

Perelson, A. S., and Nelson, P. W. (1999). Mathematical analysis of HIV-1 dynamics *in vivo*. *SIAM Rev*. 41, 3–44. doi: 10.1137/S0036144598335107

Picard, C., Dallot, S., Brunker, K., Berthier, K., Roumagnac, P., Soubeyrand, S., et al. (2017). Exploiting genetic information to trace plant virus dispersal in landscapes. *Annu. Rev. Phytopathol*. 55, 139–160. doi: 10.1146/annurev-phyto-080516-035616

Pinky, L., and Dobrovolny, H. M. (2017). The impact of cell regeneration on the dynamics of viral coinfection. *Chaos* 27:063109. doi: 10.1063/1.4985276

Pinky, L., and Dobrovolny, H. M. (2020). SARS-CoV-2 coinfections: Could influenza and the common cold be beneficial? *J. Med. Virol*. 92, 2623–2630. doi: 10.1002/jmv.26098

Piry, S., Wipf-Scheibel, C., Martin, J.-F., Galan, M., and Berthier, K. (2017). High throughput amplicon sequencing to assess within- and between-host genetic diversity in plant viruses. *bioRxiv* 2017:168773. doi: 10.1101/168773

Poirier, E. Z., and Vignuzzi, M. (2017). Virus population dynamics during infection. *Curr. Opin. Virol*. 23, 82–87. doi: 10.1016/j.coviro.2017.03.013

Pybus, O. G., and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet*. 10, 540–550. doi: 10.1038/nrg2583

Roizman, B. (1996). "Chapter 42: Multiplication," in *Medical Microbiology, 4th Edn*, ed S. Baron (Galveston, TX: University of Texas, Medical Branch at Galveston). p. 543–554.

Saenz, R. A., Quinlivan, M., Elton, D., MacRae, S., Blunden, A. S., Mumford, J. A., et al. (2010). Dynamics of influenza virus infection and pathology. *J. Virol*. 84, 3974–3983. doi: 10.1128/JVI.02078-09

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Global Sensitivity Analysis: The Primer*. Chichester: John Wiley & Sons. doi: 10.1002/9780470725184

Sanjuán, R. (2010). Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. *Philos. Trans. R. Soc. B Biol. Sci*. 365, 1975–1982. doi: 10.1098/rstb.2010.0063

Sanjuán, R., Moya, A., and Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U.S.A*. 101, 8396–8401. doi: 10.1073/pnas.0400146101

Simmons, H., Dunham, J., Stack, J., Dickins, B., Pagan, I., Holmes, E., et al. (2012). Deep sequencing reveals persistence of intra-and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. *J. Gen. Virol*. 93, 1831–1840. doi: 10.1099/vir.0.042622-0

Smith, A. M., and Perelson, A. S. (2011). Influenza A virus infection kinetics: quantitative data and models. *Wiley Interdiscip. Rev. Syst. Biol. Med*. 3, 429–445. doi: 10.1002/wsbm.129

Smith, A. M., and Ribeiro, R. M. (2010). Modeling the viral dynamics of influenza A virus infection. *Crit. Rev. Immunol*. 30, 291–298. doi: 10.1615/CritRevImmunol.v30.i3.60

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci*. 17, 57–86.

Vaidya, N. K., Ribeiro, R. M., Miller, C. J., and Perelson, A. S. (2010). Viral dynamics during primary simian immunodeficiency virus infection: effect of time-dependent virus infectivity. *J. Virol*. 84, 4302–4310. doi: 10.1128/JVI.02284-09

Valdazo-González, B., Kim, J. T., Soubeyrand, S., Wadsworth, J., Knowles, N. J., Haydon, D. T., et al. (2015). The impact of within-herd genetic variation upon inferred transmission trees for foot-and-mouth disease virus. *Infect. Genet. Evol*. 32, 440–448. doi: 10.1016/j.meegid.2015.03.032

Wainwright, H. M., Finsterle, S., Jung, Y., Zhou, Q., and Birkholzer, J. T. (2014). Making sense of global sensitivity analyses. *Comput. Geosci*. 65, 84–94. doi: 10.1016/j.cageo.2013.06.006

Walker, T. M., Ip, C. L., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., 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

Wang, S., Pan, Y., Wang, Q., Miao, H., Brown, A. N., and Rong, L. (2020). Modeling the viral dynamics of SARS-CoV-2 infection. *Math. Biosci*. 328:108438. doi: 10.1016/j.mbs.2020.108438

Worby, C. J., Lipsitch, M., and Hanage, W. P. (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput. Biol*. 10:e1003549. doi: 10.1371/journal.pcbi.1003549

Worby, C. J., and Read, T. D. (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

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* 16, 97–159. doi: 10.1093/genetics/16.2.97

Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., et al. (2018). PHYLOSCANNER: inferring transmission from within- and between-host pathogen genetic diversity. *Mol. Biol. Evol*. 35, 719–733. doi: 10.1093/molbev/msx304

Xu, S., Stapley, J., Gablenz, S., Boyer, J., Appenroth, K. J., Sree, K. S., et al. (2019). Low genetic variation is associated with low mutation rate in the giant duckweed. *Nat. Commun*. 10, 1–6. doi: 10.1038/s41467-019-09235-5

Ypma, R. J., Bataille, A., Stegeman, A., Koch, G., Wallinga, J., and Van Ballegooijen, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. Lond. B Biol. Sci*. 279, 444–450. doi: 10.1098/rspb.2011.0913

Ypma, R. J., van Ballegooijen, W. M., and Wallinga, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195, 1055–1062. doi: 10.1534/genetics.113.154856

Yuan, Y., and Allen, L. J. (2011). Stochastic models for virus and immune system dynamics. *Math. Biosci*. 234, 84–94. doi: 10.1016/j.mbs.2011.08.007