Rates of coalescence for common epidemiological models at equilibrium

Katia Koelle1,2,* and David A. Rasmussen1

1Department of Biology, Duke University, Box 90338, Durham, NC 27708, USA
2Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

Coalescent theory provides a mathematical framework for quantitatively interpreting gene genealogies. With the increased availability of molecular sequence data, disease ecologists now regularly apply this body of theory to viral phylogenies, most commonly in attempts to reconstruct demographic histories of infected individuals and to estimate parameters such as the basic reproduction number. However, with few exceptions, the mathematical expressions at the core of coalescent theory have not been explicitly linked to the structure of epidemiological models, which are commonly used to mathematically describe the transmission dynamics of a pathogen. Here, we aim to make progress towards establishing this link by presenting a general approach for deriving a model’s rate of coalescence under the assumption that the disease dynamics are at their endemic equilibrium. We apply this approach to four common families of epidemiological models: standard susceptible-infected-susceptible/susceptible-infected-recovered/susceptible-infected-recovered-susceptible models, models with individual heterogeneity in infectivity, models with an exposed but not yet infectious class and models with variable distributions of the infectious period. These results improve our understanding of how epidemiological processes shape viral gene genealogies, as well as how these processes affect levels of viral diversity and rates of genetic drift. Finally, we discuss how a subset of these coalescent rate expressions can be used for phylodynamic inference in non-equilibrium settings. For the ones that are limited to equilibrium conditions, we also discuss why this is the case. These results, therefore, point towards necessary future work while providing intuition on how epidemiological characteristics of the infection process impact gene genealogies.

Keywords: phylodynamics; viral evolution; disease dynamics; epidemiological models; coalescent theory

1. INTRODUCTION

It is well known that an infectious disease’s population dynamics are affected by the epidemiological characteristics of the infection process. For example, it has been shown that infectious period distributions closer to normal than exponential lead to a destabilization of disease dynamics [1], that longer incubation periods lead to slower epidemic invasion rates [2], and that transmission rate heterogeneity can impact disease emergence [3]. Here, we instead consider how epidemiological characteristics such as these impact the structure of viral gene genealogies, patterns of genetic diversity and rates of genetic drift. We specifically focus our analysis on the case of neutrally evolving RNA viruses and address these questions by explicitly linking epidemiological models with coalescent theory.

Previous work applying coalescent theory to viral sequence data has commonly taken one of two approaches. The first approach uses sequence data to reconstruct a composite quantity that population geneticists generally call $\theta$. If the sequence data were isolated at a single point in time so that the mutation rate cannot be directly estimated, $\theta$ is twice the product of the effective viral population size $N_e$ and the viral mutation rate $\mu$ ($\theta = 2N_e\mu$). If instead the sequence data were isolated serially in time, $\theta$ is taken to be twice the product of $N_e$ and the generation time $\tau$ ($\theta = 2N_e\tau$) and the mutation rate $\mu$ can be estimated directly [4,5]. Estimation of $\theta$ from viral sequences has been popularized by the Bayesian skyline plot [6] implemented in the Bayesian phylogenetics program BEAST [7]. The second approach uses viral sequences to estimate population parameters, such as the basic reproduction number $R_0$. This approach has relied on simple demographic models that have analytical time-dependent solutions, most notably the constant, exponential and logistic growth models. With an analytical solution to the viral dynamics, one of several approaches can be used to determine the probability of a gene genealogy under the assumed population dynamics (e.g. [8], for the most general case), thereby enabling parameter estimation. Applications of these approaches have included the estimation of HIV’s generation time [9], the estimation of H1N1pdm’s basic reproduction number during its exponential growth.

*Author for correspondence (katia.koelle@duke.edu).
phase [10] and the estimation of the hepatitis C virus’s growth rate during the virus’s establishment [11]. Two more recent approaches have now also enabled parameter estimation for nonlinear susceptible-infected-recovered (SIR) models that are not analytically solvable [12,13]. The first approach, developed by Volz et al. [12], uses an equation for the rate of lineage coalescence in an SIR model to reconstruct how the number of infected individuals changes deterministically over time. This work also elegantly pointed out that the rate of lineage coalescence should depend on both disease incidence and disease prevalence, rather than on prevalence levels alone [12,14]. The second approach extends a Bayesian inference method based on particle filtering to accommodate gene genealogies as observed data [13]. Using Volz and co-workers’ rate of coalescence expression, this method allows the estimation of epidemiological parameters and reconstruction of infection dynamics from genealogies in the presence of both process and observation noise, which have been shown to play an important role in understanding patterns of disease incidence [15,16].

While all of these approaches have relied on gene genealogies to gain information on viral epidemiology, they do not directly address how epidemiological characteristics of the infection process affect gene genealogies and patterns of viral diversity. Here, instead of focusing on further developing statistical approaches, we consider this topic mathematically by explicitly linking epidemiological model structures to the familiar expressions that form the basis of coalescent theory. At the core of our analysis lies a general approach that can be used to derive effective viral population sizes from the variance of individuals’ basic reproduction numbers under the assumption that disease dynamics are at their endemic equilibrium. We combine these effective population sizes with generation times that are appropriate for the epidemiological models to obtain coalescent rate expressions. These expressions determine the expected level of circulating viral diversity for a given epidemiological model and its rate of genetic drift under neutrality. We apply our general approach to four sets of common epidemiological models: basic susceptible-infected-susceptible/susceptible-infected-recovered/susceptible-infected-recovered (SIS/SIR/SIRS) models, models that include disease ‘supershedding’, models with an exposed but not yet infectious class and models with variable infectious period distributions. For each of these applications, we demonstrate the accuracy of our derived rates of coalescence with forward simulations. These results, therefore, provide a simple framework for understanding how and why alternative epidemiological characteristics of the infection process impact viral gene genealogies.

2. A BRIEF REVIEW OF COALESCENT THEORY

We start with a brief review of the parts of coalescent theory that are necessary for the derivation of our general approach. The key analytical result of the Kingman coalescent is that the rate of lineage coalescence is given by

\[ \lambda_i = \binom{i}{2} \]  

in the limit of infinite population size, where \( i \) is the number of sampled lineages [17]. This yields the probability density function (PDF) for the time to first coalescence

\[ f(t) = \lambda_i e^{-\lambda_i t}, \]  

where \( t \) is measured backwards in time. The cumulative distribution function (CDF), providing the probability that the first coalescence has occurred by time \( t \), is given by

\[ F(t) = 1 - e^{-\lambda_i t}. \]  

To make use of these expressions, time \( t \) needs to be scaled appropriately. For the standard Wright–Fisher (WF) model, time is measured in units of \( N \) generations [18]. Assuming that a WF generation takes \( \tau \) units of calendar time, the rate of coalescence becomes

\[ \lambda_i = \binom{i}{2} \frac{1}{N\tau} = \frac{i(i-1)}{2N\tau}. \]  

For the standard Moran model, time is instead appropriately measured in units of \( N/2 \) Moran generations [18]. Assuming that a Moran generation takes \( \tau \) units of calendar time, this model’s rate of coalescence becomes

\[ \lambda_i = \binom{i}{2} \frac{2}{N\tau} = \frac{i(i-1)}{N\tau}. \]  

The difference in the factor of two between these models’ rates of coalescence is well known to arise from the difference in these models’ offspring distributions, rather than the overlapping versus discrete nature of their generations. Specifically, the offspring distribution has a greater variance under the Moran model than the WF model, leading to more rapid lineage coalescence.

Under the standard WF model, all \( N \) individuals are equally likely to leave offspring in the next generation. Unequal offspring distributions, however, are not difficult to consider as they simply change the scaling of time in the coalescent [18,19]. The coalescent effective population size \( N_e \) is defined as the population size that yields the appropriate measurement of time for this distribution [20]:

\[ \lambda_i = \frac{i(i-1)}{2N_e\tau}. \]  

\( N_e \) can be computed from the population size \( N \) and the variance \( \sigma^2 \) in the number of offspring that individuals in the population contribute [18,19]

\[ N_e = \frac{N}{\sigma^2}. \]  

The effective population size \( N_e \) therefore decreases as the reproductive output of individuals becomes increasingly disproportional.
3. A GENERAL APPROACH FOR DERIVING COALESCENT RATES OF COALESCENCE FROM EPIDEMIOLOGICAL MODELS AT THEIR ENDEMIC EQUILIBRIUM

To derive coalescent rates appropriate for epidemiological models, we first note that the relevant population size $N$ is not the size of the host population, but rather the number of infected individuals, with the relevant offspring being new infections. We can write the variance in the number of new infections contributed by already infected individuals as

$$\sigma^2 = \text{var} (\Omega) = E[\Omega^2] - (E[\Omega])^2. \quad (3.1)$$

At endemic equilibrium, the expected number of new infections contributed by an infected individual, $E[\Omega]$, is simply 1.

To evaluate the second moment of the offspring distribution, $E[\Omega^2]$, we can imagine that each infected individual has his own net reproductive rate, $r$. Given that an individual with a net reproductive rate of $r$ will have a Poisson number of offspring (with mean and variance of $r$), the second moment of $\Omega$ can be written as

$$E[\Omega^2] = \int_{r=0}^{\infty} (r^2 + r)f(r)dr, \quad (3.2)$$

where $f(r)$ is the distribution of individuals' net reproductive rates, which we denote below as $R$. Simplifying equation (3.2) yields $E[\Omega^2] = E[R^2] + E[R] = (\text{var}(R) + (E[R])^2) + E[R]$. With $E[R]$ equal to 1 at endemic equilibrium, substitution into equation (3.1) further yields

$$\sigma^2 = \text{var}(R) + 1. \quad (3.3)$$

Assuming mean-field mixing, the variance in the net reproductive rate distribution $\text{var}(R)$ can alternatively be written as $\text{var}(R_0(S/N_{N_H}))$, where $R_0$ is the basic reproduction number distribution and $S/N_{N_H}$ is the fraction of the host population that is susceptible to infection. This simplifies to $(S/N_{N_H})^2 \text{var}(R_0)$. At endemic equilibrium, $S/N_{N_H}$ is equal to $1/E[R_0]$, such that equation (3.3) becomes

$$\sigma^2 = \frac{\text{var}(R_0)}{(E[R_0])^2} + 1. \quad (3.4)$$

The rate of coalescence $\lambda_i$ for a given epidemiological model is therefore given by equation (2.5), where $\sigma^2$ needed to compute $N_i$ is given by equation (3.4). The four quantities we therefore need to know to write $\lambda_i$ for a specific disease model at endemic equilibrium with $i$ sampled lineages are the relevant population size $N$, the mean of the basic reproduction number distribution $E[R_0]$, the variance of this distribution $\text{var}(R_0)$ and the relevant generation time $\tau$. Below, we derive these four quantities for each of four common epidemiological models, and therewith their rates of coalescence at equilibrium. We further demonstrate the accuracy of our results with forward model simulations. Finally, using these results, we compute expected levels of genetic diversity and rates of genetic drift for each of these models to facilitate comparisons of these population genetic quantities across models.

4. APPLICATION TO FOUR COMMON FAMILIES OF EPIDEMIOLOGICAL MODELS

4.1. Standard SIS/SIR/SIRS models

To derive the rate of coalescence for standard SIS/SIR/SIRS models, we can first write the dynamics of the infected population that are common to all of them

$$\frac{dI}{dt} = \beta \left( \frac{S}{N_{N_H}} \right) I - (\nu + \delta)I, \quad (4.1)$$

where $\beta$ is the transmission rate, $\nu$ is the recovery rate, $\delta$ is the background mortality rate, $S$ is the number of susceptible hosts and $N_{N_H}$ is the total size of the host population. (We could equivalently write equation (4.1) using the density-dependent transmission term $\beta SI$, because we are considering models at endemic equilibrium, with $N_{N_H}$ constant, the density-dependent transmission rate would simply be written as $\beta/N_{N_H}$, where $\beta$ is the frequency-dependent transmission rate.)

Because transmission of viral lineages can only occur from infected individuals, the focal population $N$ is not the host population size $N_{N_H}$, but rather the size of the infected population, $I$. From equation (4.1), it is clear that each of the infected individuals in a population has the same transmission rate parameter $\beta$ and the same rate of infection loss $(\nu + \delta)$. Because the rate of infection loss is constant, the duration of time that individuals remain infected and infectious is exponentially distributed with mean $1/(\nu + \delta)$. As a result, individuals' basic reproduction numbers are exponentially distributed with mean $E[R_0] = \beta/(\nu + \delta)$ and variance $\text{var}(R_0) = (\beta/(\nu + \delta))^2$.

The generation time $\tau$ is given by the average amount of time between an individual becoming infected and his transmission of the disease. From equation (4.1), we see that $\tau$ is given by $1/(\beta(S/N_{N_H}))$. This is equivalent to the average duration of infection $1/(\nu + \delta)$ when $dI/dt = 0$, which is the case at endemic equilibrium.

By substituting these expressions for $N_i E[R_0], \text{var}(R_0)$ and $\tau$ into equations (3.4) and (2.5), we now have a rate of coalescence for standard SIS/SIR/SIRS models at equilibrium:

$$\lambda_i = \frac{i(i-1)(\nu + \delta)}{I}. \quad (4.2)$$

Figure 1a shows a genealogy that has been simulated using this rate of coalescence. A lineages-through-time plot derived from this genealogy is shown in figure 1c along with the expected lineages-through-time plot for this model.

To support the analytical result shown in equation (4.2), we forward-simulated an SIR model using Gillespie’s exact algorithm [21], starting with initial conditions at endemic equilibrium. In this simulation, we further tracked transmission chains of infected individuals. At the end of the simulation, we sampled $i = 40$ currently
infected individuals and, using the transmission chains, determined the time at which the first two of these i lineages coalesced. We repeated this sampling procedure 1000 times, yielding 1000 times to first coalescence. A histogram showing this distribution of first coalescent times is shown in figure 2a. Alongside this histogram, we show the PDF for the time to first coalescence for an SIR model under the same parameterization. A

Figure 1. Simulated genealogies for the four classes of epidemiological models considered, each starting with 250 homochronous tips. Each genealogy was simulated using the derived rate of coalescence for the epidemiological model specified. For a given model, the time to the first coalescence $t_{250}$ was drawn from an exponential distribution with mean $1/\lambda_{250}$. The time between the first coalescence and the second coalescence $t_{249}$ was drawn from an exponential distribution with mean $1/\lambda_{249}$, and so forth until all the coalescent time intervals had been drawn. Coalescing lineages were randomly chosen from the extant lineages at each coalescence event. (a) A genealogy for an SIS/SIR/SIRS model. (b) A genealogy for an epidemiological model with supershedding. (c) A genealogy for an epidemiological model with an exposed class. (d) A genealogy for an epidemiological model with a variable distribution of the infectious period. Epidemiological parameters used in (a–d) are: a constant number of infected individuals of $I = 10000$, a recovery rate of $\nu = 1/10$ d$^{-1}$, a mortality rate of $\mu = 1/70$ yr$^{-1}$. In (b), the transmissibility shape parameter $k_b = 3$. In (c), $E = 20000$ exposed individuals and the rate of leaving the exposed class $\sigma = 1/20$ d$^{-1}$. In (d), $n = 10$ infectious classes. Scale bars, measuring 6000 days, are the same across all shown genealogies. (e) Lineages-through-time plots for the four classes of epidemiological models considered, colour-coded according to the genealogies shown in (a–d). Dots show lineages-through-time plots derived from the simulated genealogies shown in (a–d). Lines show expected lineages-through-time plots for these models, using the means $1/\lambda_{250}$, $1/\lambda_{249}$ and so forth.

A histogram showing this distribution of first coalescent times is shown in figure 2a. Alongside this histogram, we show the PDF for the time to first coalescence for an SIR model under the same parameterization. A
comparison between this histogram and the PDF indicates that equation (4.2) is consistent with the simulations. In figure 2b, we plot the CDF for the time to first coalescence, given the rate in equation (4.2), along with the proportion of the 1000 sample sets that had at least one coalescent event by the time indicated on the x-axis. This provides a more effective way to graphically assess the accuracy of our analytical results using our simulation results.

Equation (4.2) should come as no surprise because continuous-time SIS/SIR/SIRS models at endemic equilibrium are all in essence Moran models: an infected individual gives ‘birth’ at a constant rate ($\beta(S/N_H)$) by transmitting an infection to a susceptible host and ‘dies’ at an equivalent constant rate ($\nu + \delta$) by recovering from infection or through natural mortality. It could be argued that one difference between SIS/SIR/SIRS models and the Moran model is that a single event consists of a birth coupled with a death in the Moran model. In contrast, stochastic simulations of SIS/SIR/SIRS models do not have new infections explicitly coupled with recoveries. Nevertheless, in the limit of infinite population sizes, one transmission event occurs for every recovery event when the infected population size is at equilibrium, such that the rate of coalescence expression is consistent with that of the Moran model.

The mean level of genetic diversity, defined as the expected number of pairwise nucleotide differences $p$, for any population model is given by $2Nc\mu$, where $\mu$ is the mutation rate per generation [18]. In the case of SIS/SIR/SIRS models at equilibrium, $\pi$ is therefore $\mu I$, or equivalently $\mu c\Gamma = \mu c I / (\nu + \delta)$ when $\mu c$ is defined in terms of mutations per unit of calendar time. The rate of genetic drift, in units of calendar time, is given by $1/(Nc\pi)$ [18]. For SIS/SIR/SIRS models, it is therefore $2(\nu + \delta)/I$.

We now return to our derived rate of coalescence given by equation (4.2). We can compare this expression to the expression derived in Volz et al. [12] and Frost et al. [14] that governs the dynamics of the number of ancestral lineages over time

$$-\frac{dA}{dt} = -f_{SI} \frac{A(A-1)}{I(I-1)},$$

where $f_{SI}$ is the rate at which transmission events occur ($\beta(S/N_H)I$ under our model), $A$ is the number of ancestral lineages and $I$ is the number of infected individuals (prevalence). At endemic equilibrium, $f_{SI}$ will be equal to $(\nu + \delta)I$ and $I(I-1)$ can be approximated by $I^2$ when $I$ is large. With these substitutions, reversing the notation of time in equation (4.3) yields equation (4.2) once the $I$ from the numerator and one of the $I$s from the denominator have been cancelled. Our results are therefore consistent with this previous work (which does not assume dynamics at equilibrium), although our derivations differ.

4.2. SIS/SIR/SIRS models with supershedding

We now consider epidemiological models in which there is heterogeneity in the infectivity of hosts. A common approach to consider this kind of host heterogeneity classifies hosts into groups that are unique in their transmission rates [2], with the population dynamics of infected individuals belonging to group $m$ given by

$$\frac{dI_m}{dt} = \left(\sum_{j} \beta_j I \right) \frac{S_m}{N_H} - (\nu + \delta)I_m,$$

where $\beta_j$ denotes the transmission rate of individuals belonging to group $j$. Now instead of imagining distinct
groups of infected hosts, we can alternatively imagine that each individual in a population of infected individuals has his own transmission rate drawn from a gamma distribution with shape parameter $k_\beta$ and scale parameter $\lambda_\beta$ (figure 3a). The mean transmission rate in this case is given by $E[\beta] = k_\beta \lambda_\beta$ and the variance of the individuals’ transmission rates is $\text{var}(\beta) = k_\beta \lambda_\beta^2$. As in the standard SIS/SIR/SIRS models, the duration of infection is exponentially distributed with mean $1/(\nu + \delta)$ and variance $1/(\nu + \delta)^2$. The mean basic reproduction number is therefore $E[R_0] = k_\beta \lambda_\beta / \nu + \delta$. To calculate the variance in the $R_0$ distribution, we can use Goodman’s equation for the variance of products [22]

$$\text{var}(xy) = E[x]^2 \text{var}(y) + E[y]^2 \text{var}(x) + \text{var}(x) \text{var}(y),$$

(4.5)

where $x$ and $y$ are independent random variables, $E[x]$ and $E[y]$ are the expected values of $x$ and $y$, respectively, and $\text{var}(x)$ and $\text{var}(y)$ are the variances of $x$ and $y$, respectively. Substituting $k_\beta \lambda_\beta$ for $E[x]$, $k_\beta \lambda_\beta^2$ for $\text{var}(x)$, $1/(\nu + \delta)$ for $E[y]$ and $1/(\nu + \delta)^2$ for $\text{var}(y)$, the variance of $R_0$ becomes

$$\text{var}(R_0) = \left(\frac{k_\beta \lambda_\beta}{\nu + \delta}\right)^2 \left(1 + \frac{2}{k_\beta}\right).$$

(4.6)

With $\tau = 1/(\nu + \delta)$ at endemic equilibrium, the rate of coalescence therefore becomes

$$\lambda_t = \frac{i(i - 1)(1 + 1/k_\beta)(\nu + \delta)}{I}.$$  

(4.7)

This equation indicates that the rate of coalescence is higher the more heterogeneous the transmission rates of infected individuals, an effect that is apparent in genealogies that have been simulated with this rate (figure 1b,c). As $k_\beta \to \infty$, infected individuals all approach the same transmission rate and equation (4.2) is recovered, as anticipated.

To support the analytical result shown in equation (4.7), we again used forward simulations, starting with initial conditions at endemic equilibrium and parametrized with the transmission heterogeneity distributions shown in figure 3a. For each of these three parametrizations, the proportion of sample sets that had at least one coalescent event by the time given on the $x$-axis is shown in figure 3b alongside the CDF for the time to first coalescence, calculated from the rate in equation (4.7). These comparisons indicate that equation (4.7) is accurate.

Finally, $\pi$ for this model is

$$\frac{\mu I}{(1 + 1/k_\beta)}$$

and the rate of genetic drift is

$$\frac{2(1 + 1/k_\beta)(\nu + \delta)}{I},$$

which are lower and higher, respectively, than for SIS/SIR/SIRS models with no transmission heterogeneity.
4.3. SEIS/SEIR/SEIRS models

Susceptible-exposed-infected-susceptible/susceptible-exposed-infected-recovered/susceptible-exposed-infected-recovered-susceptible (SEIS/SEIR/SEIRS) models are similar to SIS/SIR/SIRS models, but have an additional class of individuals who have become infected but are not yet infectious. These individuals are classified as exposed. The dynamics of exposed and infectious individuals in these models are given by:

\[
\frac{dE}{dt} = \beta \left( \frac{S}{N_H} \right) I - (\sigma + \delta)E \quad \text{and} \quad \frac{dI}{dt} = \sigma E - (\nu + \delta)I
\]

where \( \sigma \) is the rate at which exposed individuals transition to becoming infectious.

The mean basic reproduction number for this class of models is given by \([2]\)

\[
E[R_0] = \left( \frac{\beta}{\nu + \delta} \right) \left( \frac{\sigma}{\sigma + \delta} \right).
\]

This is because the rate at which an infected individual infects susceptible hosts is given by \( \beta \), but only a fraction of them, \( (\sigma/\sigma + \delta) \), will become infectious. The remaining fraction will die prior to becoming infectious. The variance in the basic reproduction number is given by

\[
\text{var}(R_0) = \left( \left( \frac{\beta}{\nu + \delta} \right) \left( \frac{\sigma}{\sigma + \delta} \right) \right)^2
\]

because the rate of infection loss is again constant, leading to an exponential distribution for the duration of time individuals remain infectious.

The two remaining quantities that require identification are \( \tau \) and \( N \). From equations (4.8), we might naively expect that the average amount of time it takes an individual to transmit the disease is

\[
\tau = \frac{1}{\sigma} + \frac{1}{\beta(S/N_H)}
\]

which is equivalent to

\[
\tau = \frac{1}{\sigma} + \frac{1}{\nu + \delta}
\]

at endemic equilibrium. Although this is the case for \( \delta \ll \sigma \), when this assumption is not met, \( 1/\sigma \) overestimates the average amount of time an individual who becomes infectious spends in the exposed class. To accurately calculate this time, we define \( 1/\sigma_{ad} \) as the average duration of time an individual spends in the exposed class, provided that he will transition to the infectious class. This mortality-adjusted expected time is given by \([23]\): 

\[
\frac{1}{\sigma_{ad}} = \int_0^\infty \frac{1 - e^{-\delta \tau}}{\delta} \sigma e^{-\sigma \tau} d\tau = \frac{1}{\sigma + \delta}.
\]

The generation time \( \tau \) is therefore

\[
\tau = \frac{1}{\sigma + \delta} + \frac{1}{\nu + \delta}.
\]

Finally, the relevant population size \( N \) is appropriately defined as the sum of the number of exposed individuals who will become infectious and the number of infectious individuals:

\[
N = \left( \frac{\sigma}{\sigma + \delta} \right) E + I.
\]

This is because, over a single generation, viral lineages are transmitted by individuals who are either already infectious at the beginning of the generation or by individuals who are exposed at the beginning of the generation, but who will become infectious during the viral generation, as defined above. With these quantities, the rate of coalescence for SEIS/SEIR/SEIRS models becomes

\[
\lambda_i = \frac{i(i-1)}{((\sigma/\sigma + \delta))E + I) (1/(\sigma + \delta) + 1/(\nu + \delta)).
\]

Figure 1c shows a genealogy that has been simulated using equation (4.13). A lineages-through-time plot for this genealogy is shown in figure 1c. This plot indicates that lineages coalesce more slowly when an exposed period is present than when absent. Equation (4.13) indicates that this is because exposed individuals increase the effective population size and because the generation time is longer than the period of infectiousness.

We again used forward simulations to check the accuracy of our analytical results. As before, figure 4 plots the proportion of sample sets that had a first coalescence by the time indicated on the x-axis against our analytical expectation, indicating that equation (4.13) is accurate. This figure also shows that that the rate of coalescence is significantly overestimated if \( N \) is incorrectly specified as only the number of infectious individuals, and that the rate of coalescence can be significantly underestimated if the generation time \( \tau \) is not adjusted for mortality from the exposed class.

The average number of pairwise nucleotide differences \( \pi \) for this class of models is given by

\[
\mu \left( \left( \frac{\sigma}{\sigma + \delta} \right) E + I \right),
\]

which is higher than that for SIS/SIR/SIRS models. The rate of genetic drift is given by

\[
2 \left( \frac{E(\sigma/\sigma + \delta)) + I)(1/(\sigma + \delta) + 1/(\nu + \delta)) \right),
\]

which is lower than that for SIS/SIR/SIRS models.

4.4. SIS/SIR/SIRS models with variable distributions of the infectious period

SIS/SIR/SIRS models are frequently extended to allow for variable distributions for the amount of time individuals remain infectious. This is commonly done by using the method of stages \([24]\): simulating \( n > 1 \) compartments of infected individuals, with individuals moving from one compartment to the next...
compartment at a constant rate \(vn\) [1]:

\[
\frac{dI}{dt} = \beta \frac{S}{N} \sum_{i=1}^{n} I_i - (vn + \delta)I_i
\]

and

\[
\frac{dI_m}{dt} = (vn)I_{m-1} - (vn + \delta)I_m \quad \text{for} \quad m = 2, \ldots, n.
\]

(4.14)

Under this formulation, the average amount of time individuals remain infectious is \(1/\nu\), provided that the mortality rate is negligible compared with the rate at which individuals transition between infected classes, \(\delta \ll vn\). As \(n \to \infty\), all individuals remain infectious for \(1/\nu\) units of time. As \(n \to 1\), the duration of time individuals remain infectious becomes exponentially distributed, with the classic class of SIS/SIR/SIRS models recovered when \(n = 1\).

To derive the rate of coalescence for this class of epidemiological models, we again need the relevant population size \(N\), \(E[R_0]\), \(\text{var}(R_0)\) and \(\tau\). We derive these quantities here under the assumption that \(\delta \ll vn\) in order to highlight the parts of this class of models that makes them unique. A similar approach to the one described in the previous class of models, however, can be used to mortality-adjust transition rates if this assumption is violated. In the equations that follow, we therefore use the notation \(\approx\) rather than \(\approx\) to remind the reader that these quantities are approximations.

Instead of deriving \(N\), \(E[R_0]\) and \(\text{var}(R_0)\) independently, it is easier to derive \(N_e\) by writing

\[N_e = \sum_{k=1}^{n} N_{e,k}\]

where \(N_{e,k}\) is the effective population size of individuals in infectious compartment \(k\). The average reproduction number of individuals in compartment \(k\) is given by \(E[R_{0,k}] = \beta/(vn)\), and the variance given by \(\text{var}(R_{0,k}) \approx (\beta/(vn))^2\) because the average duration of time individuals remain in compartment \(k\) is exponentially distributed with mean \(1/(vn)\). Substituting into equations (3.4) and (2.6) yields \(N_{e,k} = I_k/2\), such that the total effective population size becomes

\[N_e = \sum_{k=1}^{n} I_k/2.
\]

Finally, we can turn to the generation time \(\tau\). At endemic equilibrium, the average amount of time it takes for individuals in the first infectious compartment to transmit is given by \(1/(vn)\). The average amount of time, since infection, it takes for individuals in the second infectious compartment to transmit is given by \(2/(vn)\). This is because it takes on average \(1/(vn)\) amount of time for an individual in this second class to pass through the first class, and another \(1/(vn)\) units of time for the individual to transmit while in this second class. Analogously, the average amount of time for an individual in the \(k\)th infectious class to transmit is given by \(k/(vn)\). With our assumption of \(\delta \ll vn\), the number of individuals in each infectious class is the same, so that we can write the overall average time to the next transmission as

\[\tau \approx \frac{\sum_{k=1}^{n} (k/(vn))}{n} = \frac{1}{2v} \left(1 + 1/n\right).
\]

(4.15)
Putting these quantities together yields a rate of coalescence of

\[ \lambda_i \approx \frac{i(i-1)(2\nu)}{N(1+1/n)} \]  

(4.16)

where \( I = \sum_{k=1}^{n} I_k \). When the number of infectious compartments \( n \) is 1, this reduces to the family of models considered in our first application, and indeed the rate of coalescence becomes equation (4.2), once the mortality rate \( \delta \) has been assumed negligible. As \( n \) approaches infinity, the average time to infection is cut in half, and the rate of coalescence is therefore twice as fast. Figure 1d shows a genealogy that has been simulated using equation (4.16), with \( n = 10 \) infectious classes, and figure 1e the corresponding lineages-through-time plot. A comparison between the number of lineages through time under this model versus the SIS/SIR/SIRS model indicates that lineages coalesce more rapidly when the duration of infection is less variable among infected individuals, as expected by the higher rate of coalescence brought about by the shortening of the generation time.

To check the accuracy of our analytical results, we simulated three models of this class, with \( n = 1, 3 \) and 10 infectious compartments, keeping the average infectious period \( 1/\nu \) constant across the simulations. Figure 5 plots the proportion of sample sets from each of these three simulations that had a first coalescence by the time indicated on the x-axis. Comparisons against the CDFs calculated using the coalescent rate provided in equation (4.16) indicates that our analytical results are accurate.

As before, the average number of pairwise nucleotide differences \( \pi \) and rates of genetic drift can be computed using the effective population size and the generation time. The mean level of genetic diversity is, as before, given by \( 2\mu I \), where \( \mu \) is the per generation mutation rate. Although this level of genetic diversity initially appears to be the same as the one for SIS/SIR/SIRS models, if instead we consider that the mutation rate is more likely to be constant in terms of calendar time than generation time, we should instead write \( \pi \) as

\[ \frac{\mu_c I(1 + 1/n)}{2\nu} \]

with \( \mu_c \) as the mutation rate per unit of calendar time. In this case, because of the shorter generation time when \( n > 1 \), the amount of viral diversity is expected to be smaller than for SIS/SIR/SIRS models. The rate of genetic drift for these models is given by

\[ \frac{4\nu}{I(1+1/n)} \]

which is higher than that for SIS/SIR/SIRS models.

5. DISCUSSION

Here, we have presented a general approach to derive rates of coalescence for a wide range of epidemiological models at their endemic equilibrium and have applied this approach to four common families of epidemiological models. The different rates of coalescence between these models indicate that the structures of epidemiological models leave imprints on the shapes of viral genealogies. While it is well known that population size dynamics affect the shapes of viral genealogies [25], the observation that viral genealogies are also affected by factors such as the extent of disease supersheding, the presence of exposed periods and the distribution of infectious periods is novel.

The derivation of the rates of coalescence for the four families of epidemiological models we considered is not meant to be exhaustive. Many of the epidemiological models used today combine several components of the models we considered. For example, a disease may have a latent period prior to an infected person becoming infectious and also be highly variable in transmissibility among individuals. This combines the model of the third application with the model of the second application. From the general approach we outlined for deriving rates of coalescence, these more complex epidemiological models can nevertheless be easily considered by correctly specifying the relevant population size \( N \) and the relevant generation \( \tau \), and by deriving the mean and variance of individuals’ basic reproduction number distribution.

The ability to derive coalescent rates and effective population sizes specific to an epidemiological model is useful for several reasons. First, sequence data are now routinely used to reconstruct the demographic histories of viral populations, commonly visualized in the Bayesian skyline plot [6]. When sequences are serially sampled, these histories are given as the product of the effective population size and the generation time. Knowledge of how \( N_e \) and \( \tau \) are affected by epidemiological factors such as exposed periods, the distribution of infectious periods and the extent of supersheding, will therefore help us be able to better interpret the magnitude of \( N_e \tau \) and may help us understand why, for
example, a disease with low prevalence and a short duration of infection may nevertheless have higher \( \theta \) values than a persistent disease with high prevalence. (As shown here, a lower variability in transmission rates or a long exposed period in the former case may result in an overall higher \( N_t \tau \), despite a lower prevalence level and a shorter duration of infection.)

Second, statistical inference methods that seek to fit complex epidemiological models to viral genetic data will require model-specific coalescent rate expressions. This is because these rates provide the quantitative link between the structures and parameters of epidemiological models and the machinery of population genetics used to evaluate the probability of a genealogy given these models. Several statistical approaches already exist, but they have all been based on the simplest of epidemiological models (e.g. SI, SIS or SIR models) [11–13]; the consideration of more complex epidemiological models, such as the ones presented in applications 4.1 through 4.4, will therefore require an \textit{a priori} appropriate specification of coalescent rates. Our general approach thereby facilitates the extension of existing methods to consider more realistic epidemiological processes.

We caution the reader, however, in directly applying the general approach outlined here to derive model-specific coalescent rates for statistical inference methods. This is for several reasons. First, we have generally written the generation time \( \tau \) in terms of the average duration of infection, rather than in terms of the average time it takes for an infected individual to transmit the disease (which has been termed the \textit{generation interval} [26]). At endemic equilibrium, these two quantities are equal. However, during the rise of an epidemic, the generation interval is shorter than the average duration of infection [26]. Similarly, during the fall of an epidemic, the generation interval is longer than the average duration of infection. Care therefore needs to be taken to appropriately define \( \tau \) when dynamics are not at equilibrium.

Second, the formulation presented here is only valid if the transmission term is linear in the number of infecteds (i.e. if the transmission term is of the form \( \beta S^r T^p \), where \( \alpha = 1 \), E. Volz 2011, personal communication). This can be seen by returning to the rate of coalescence derived in Volz \textit{et al.} [12], shown in equation (4.3), which does not assume equilibrium conditions. When the transmission term is linear in the number of infected individuals, the \( I \) in the numerator and denominator cancel and the resulting rate of coalescence can be seen as only depending inversely on disease prevalence \( I \) (and directly on the number of susceptible hosts). However, when \( \alpha \neq 1 \), the \( I_s \) do not simply cancel, and the rate of coalescence can therefore never be described in terms of \( N_t \tau \) formalism. However, most epidemiological models that are commonly used are linear in \( I \), so we do not consider this a major limitation.

Finally, the general approach detailed here is limited in that it cannot be accurately applied to non-equilibrium situations when the population of infected individuals is structured in some way. For example, SEIR models are structured: infected individuals are classified as exposed (\( E \)) or infectious (\( I \)). Models with variable distributions of the infectious period are similarly structured, because distinct infectious classes exist (\( I_1, I_2, \ldots, I_n \)). The approach outlined here is not suitable for these situations because the rate of coalescence should depend on the number of individuals in each compartment, not just their sum. For example, if all individuals infected in an SEIR model belonged to the \( I \) class and none to the \( E \) class at time \( t \), then the instantaneous rate of coalescence at time \( t \) should be zero. This is because lineage coalescence can only occur when an infected individual has just infected a susceptible individual; backwards in time, this means that for there to be a positive rate of coalescence, both exposed and infected individuals need to be present in the population. The number of individuals in each class of infected individuals therefore matters, not only their sum.

Despite these limitations, some of the rates derived here can be directly integrated into current statistical inference methods (e.g. [13]). For example, the rate of coalescence derived for SIS/SIR/SIRS models, given by equation (4.2), can be applied to non-equilibrium dynamics as long as \( \tau \) is redefined as \( 1/([\beta S/N_S]) \). This is similarly so for the rate of coalescence derived for SIS/SIR/SIRS models with supershedding, given by equation (4.7). Furthermore, the work presented here can also be used to check the accuracy of coalescent rate expressions that have been more generally derived for structured populations with non-equilibrium dynamics: when considered at equilibrium, the rate expressions should be identical.

The work presented here therefore highlights the role that epidemiological model structures can play in shaping viral genealogies. Most importantly, our general approach for deriving rates of coalescence for models at endemic equilibrium shows how characteristics of the infection process alter effective population sizes and generation times. As such, it provides an intuitive framework for better understanding observed patterns of viral diversity. However, as elaborated on above, our approach has only limited applicability to non-equilibrium systems. It therefore points towards the need for an even more general approach that would enable a more effective use of viral genetic data for gaining insight into the factors driving infectious disease dynamics.

We thank Chris Castorena and Allen Rodrigo for helpful comments and useful discussions. This work was supported by grant NSF-EF-08-27416 to K.K., by the RAPIDD programme of the Science and Technology Directorate, Department of Homeland Security, and the Fogarty International Centre, National Institutes of Health and by an NSF graduate research fellowship to D.A.R.

REFERENCES


