Ancestral process for infectious disease outbreaks with superspreading

Xavier Didelot[1,2,*], David Helekal[3], Ian Roberts[2,4]

[1] School of Life Sciences, University of Warwick, Coventry, United Kingdom

[2] Department of Statistics, University of Warwick, Coventry, United Kingdom

[3] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

[4] Pandemic Sciences Institute, University of Oxford, Oxford, United Kingdom

[*] Corresponding author. Tel: 0044 (0)2476 572827. Email: `xavier.didelot@warwick.ac.uk`

Running title: Ancestry for outbreaks with superspreading

Keywords: infectious disease epidemiology modelling; offspring distribution; superspreading; outbreaks; Lambda-coalescent model; multiple mergers

Abstract

When an infectious disease outbreak is of a relatively small size, describing the ancestry of a sample of infected individuals is difficult because most ancestral models assume large population sizes. Given a set of infected individuals, we show that it is possible to express exactly the probability that they have the same infector, either inclusively (so that other individuals may have the same infector too) or exclusively (so that they may not). To compute these probabilities requires knowledge of the offspring distribution, which determines how many infections each infected individual causes. We consider transmission both without and with superspreading, in the form of a Poisson and a Negative-Binomial offspring distribution, respectively. We show how our results can be incorporated into a new Lambda-coalescent model which allows multiple lineages to coalesce together. We call this new model the Omega-coalescent, we compare it with previously proposed alternatives, and advocate its use in future studies of infectious disease outbreaks.

1 Introduction

An outbreak of an infectious disease typically starts when a single or a small number of infected individuals appear within a susceptible population. Each infected individual may come in contact with and transmit the disease to each of the susceptible individuals, who will then become infected in their turn and spread the disease further. Most mathematical models of infectious diseases describe situations where the disease is at an equilibrium, when the number of infected individuals is high and/or with a significant part of the population already infected (**??**). Here however we focus on the early stages of an epidemic, where the number of infected individuals is small and the number of susceptibles comparatively high and constant. In this situation it is useful to consider the number of new infections that each infected individual is likely to cause, and the probabilistic distribution for this number is often called the offspring distribution (**?**). The mean of the offspring distribution is called the basic reproduction number R_0 and has been given much attention especially since it determines how likely the outbreak is to spread, and how much effort would be needed to bring it under control (**??**).

If we consider that all individuals are infectious for the same duration and with the same transmission rate, the offspring distribution is Poisson distributed with mean R_0, in which case the variance of the offspring distribution is also R_0. We would then say that there is no transmission heterogeneity. However, in practice there are many reasons why this may not be the case, with some individuals being infectious for longer than others, or being more infectious than others, or having more frequent contacts with susceptibles, or being less symptomatic and therefore less likely to reduce contact numbers, etc. All these factors cause the offspring distribution to be more dispersed than it would otherwise be, that is to have a variance greater than its mean R_0. A frequent choice to capture this overdispersion is to model the offspring distribution using a Negative-Binomial distribution with mean R_0 and dispersion parameter r (**??**). When r is close to zero the variance is high compared to the mean, whereas when r is high the variance becomes close to the mean. This transmission heterogeneity is often called superspreading, although this is perhaps misleading as it is the rule rather than the exception of how infectious diseases spread. Superspreading has indeed been described in many diseases (**????**), and most recently for SARS-CoV-2 (**????**).

As an outbreak unfolds forward-in-time, a transmission tree is generated representing who-infected-whom, in which each node is an infected individual and points towards a number of nodes distributed according to the offspring distribution. Here we consider the reverse problem of the transmission ancestry, going backward-in-time, from a sample of infected individuals, until reaching the last common transmission ancestor of the whole sample. Given a set of n sampled individuals, we show how to calculate the probability that a given subset of size k have the same infector, either inclusively (so that the remaining $n-k$ may also have the same infector or not) or exclusively (so that none of the remaining $n-k$ have the same infector). We start by considering the general case of an offspring distribution with arbitrary form, and then the specific cases of offspring distributions that follow a Poisson and a Negative-Binomial distribution. The main novelty of our approach is that we consider that the overall population size is small, but we show that in the limit where the population size is large, our results agree with several previous studies (**???**). Finally, we show how our results can be incorporated into a new Lambda-coalescent model (**???**) and compare it with previously proposed models.

2 General offspring distribution case

Let time be measured in discrete units and denoted t. Each discrete value of t corresponds to a unique non-overlapping generation of infected individuals, so that individuals infected at t have offspring at $t+1$, etc. Let N_t denote the number of infectious individuals at time t. Each of them creates a number $s_{t,i}$ of secondary infections at time $t+1$, following the offspring distribution $\alpha_t(s)$. The mean of this distribution is the basic reproduction number R_t and the variance is V_t. The total number of infected individuals at time $t+1$ is given by:

$$N_{t+1} = \sum_{i=1}^{N_t} s_{t,i} \tag{1}$$

2.1 Inclusive coalescence probability

We define the inclusive coalescence probability $p_{k,t}(N_t, N_{t+1})$ as the probability that a specific set of k individuals from generation $t+1$ have the same infector in generation t, conditional on population sizes N_t and N_{t+1}. Conditioning on offspring counts $\mathbf{S}_t = (S_{t,1}, S_{t,2}, \ldots, S_{t,N_t})$ where each $S_{t,j}$ is an independent draw from the offspring distribution α_t, we have:

$$p_{k,t}(N_t, N_{t+1} | \mathbf{S}_t = \mathbf{s}_t) = \sum_{i=1}^{N_t} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}}$$

$$= \sum_{i=1}^{N_t} \frac{s_{t,i}!}{(s_{t,i}-k)!} \frac{(N_{t+1}-k)!}{N_{t+1}!} \tag{2}$$

Full information about offspring counts $\mathbf{S}_t$ yields the population size N_{t+1} as shown in Equation 1, but this is not available in practice. We can instead express the inclusive coalescence probability conditioning only on the next population size N_{t+1} by summing over possible offspring counts $\mathbf{s}_t = (s_{t,1}, \dots s_{t,N_t})$ such that $\sum_{i=1}^{N_t} s_{t,i} = N_{t+1}$. Letting $S_t^{-(1)} = (S_{t,2}, \dots, S_{t,N_t})$, we have:

$$\begin{aligned}
p_{k,t}(N_t, N_{t+1}) &= \sum_{\mathbf{s}_t \in \mathbb{N}_0^{N_t}} \mathbb{P}\left[\mathbf{S}_t = \mathbf{s}_t \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] p_{k,t}(N_t, N_{t+1} | \mathbf{S}_t = \mathbf{s}_t) \\
&= \sum_{\mathbf{s}_t \in \mathbb{N}_0^{N_t}} \mathbb{P}\left[\mathbf{S}_t = \mathbf{s}_t \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \sum_{i=1}^{N_t} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \\
&= \sum_{i=1}^{N_t} \sum_{\mathbf{s}_t \in \mathbb{N}_0^{N_t}} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \mathbb{P}\left[S_{t,1} = s_{t,1}, \mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \\
&= \frac{N_t}{\binom{N_{t+1}}{k}} \sum_{\mathbf{s}_t \in \mathbb{N}_0^{N_t}} \binom{s_{t,1}}{k} \mathbb{P}\left[S_{t,1} = s_{t,1} \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \\
&\qquad \times \mathbb{P}\left[\mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \middle| S_{t,1} = s_{t,1}, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \\
&= \frac{N_t}{\binom{N_{t+1}}{k}} \sum_{s_{t,1}=0}^{N_{t+1}} \binom{s_{t,1}}{k} \mathbb{P}\left[S_{t,1} = s_{t,1} \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \\
&\qquad \times \underbrace{\sum_{\mathbf{s}_t^{-(1)} \in \mathbb{N}_0^{N_t - 1}} \mathbb{P}\left[\mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \middle| \sum_{i=2}^{N_t} S_{t,i} = N_{t+1} - s_{1,t}\right]}_{=1} \\
&= \frac{N_t}{\binom{N_{t+1}}{k}} \mathbb{E}\left[\binom{S_{t,1}}{k} \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \\
&= N_t \frac{(N_{t+1}-k)!}{N_{t+1}!} \mathbb{E}\left[\frac{S_{t,1}!}{(S_{t,1}-k)!} \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \qquad (3)
\end{aligned}$$

The k-th falling factorial moments $\mathbb{E}\left[\frac{S_{t,1}!}{(S_{t,1}-k)!} \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]$ in Equation 3 can be readily obtained by differentiating the probability generating function of $S_{t,1}|(\sum_{i=1}^{N_t} S_{t,i} = N_{t+1})$.

2.2 Exclusive coalescence probability

Generally, we observe a sample of individuals from each generation rather than the entire population. In this case, we are interested in the exclusive coalescence probability $p_{n,k,t}(N_t, N_{t+1})$ that a specific subset of k individuals amongst n sampled individuals arose from a common infector one generation in the past given knowledge of the total population sizes N_t and N_{t+1}. Let us begin by conditioning on the offspring counts of the individuals at time t amongst the sample at time $t+1$, namely $\mathbf{X}_t = (X_{t,1}, X_{t,2}, \ldots, X_{t,N_t})$ such that $X_{t,1} + \cdots + X_{t,N_t} = n$. Note that $X_{t,i}$ may not follow the same offspring distribution as $S_{t,i}$. We have:

$$\begin{aligned} p_{n,k,t}(N_t, N_{t+1} | \mathbf{X}_t = \mathbf{x}_t) &= \sum_{i=1}^{N_t} \frac{\binom{x_{t,i}}{k}}{\binom{n}{k}} \mathbb{I}\{x_{t,i} = k\} \\ &= \sum_{i=1}^{N_t} \frac{x_{t,i}!}{(x_{t,i} - k)!} \frac{(n-k)!}{n!} \mathbb{I}\{x_{t,i} = k\} \end{aligned} \tag{4}$$

Similarly to the inclusive coalescence probability in Equation 3, we can use this to evaluate the exclusive probability conditioning on N_t and N_{t+1} by summing over possible parent offspring configurations (for $k \leq n$):

$$\begin{aligned} p_{n,k,t}(N_t, N_{t+1}) &= \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^{n} X_{t,i} = n\right] p_{n,k,t}(N_t, N_{t+1} | \mathbf{X}_t = \mathbf{x}_t) \\ &= \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^{n} X_{t,i} = n\right] \sum_{i=1}^{N_t} \frac{\binom{x_{t,i}}{k}}{\binom{n}{k}} \mathbb{I}\{x_{t,i} = k\} \\ &= \frac{N_t}{\binom{n}{k}} \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \binom{x_{t,1}}{k} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \mathbb{I}\{x_{t,1} = k\} \\ &= \frac{N_t}{\binom{n}{k}} \sum_{\mathbf{x}_t^{-(1)} \in \mathbb{N}_0^{N_t - 1}} \binom{k}{k} \mathbb{P}\left[X_{t,1} = k, \mathbf{X}_t^{-(1)} = \mathbf{x}_t^{-(1)} \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \\ &= \frac{N_t}{\binom{n}{k}} \mathbb{P}[X_{t,1} = k \middle| \sum_{i=1}^{N_t} X_{t,i} = n\Bigg] \underbrace{\sum_{\mathbf{x}_t^{-(1)} \in \mathbb{N}_0^{N_t - 1}} \mathbb{P}\left[\mathbf{X}_t^{-(1)} = \mathbf{x}_t^{-(1)} \middle| \sum_{i=1}^{N_t} X_{t,i} = n, X_{t,1} = k\right]}_{=1} \\ &= \frac{N_t}{\binom{n}{k}} \mathbb{P}\left[X_{t,1} = k \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \end{aligned} \tag{5}$$

If we consider one of the lines observed amongst a set of n, it can either remain uncoalesced with probability $p_{n,1,t}(N_t, N_{t+1})$ or coalesce in an event of size k with probability $p_{n,k,t}(N_t, N_{t+1})$ with any set of $k-1$ lines among the $n-1$ other lines, leading to the following complementarity equation:

$$\sum_{k=1}^{n} \binom{n-1}{k-1} p_{n,k,t}(N_t, N_{t+1}) = 1 \tag{6}$$

We can show that it is indeed satisfied by the formula in Equation 5, thus providing a sanity check on this formula:

$$\begin{aligned}
\sum_{k=1}^{n} \binom{n-1}{k-1} p_{n,k,t}(N_t, N_{t+1}) &= \sum_{k=1}^{n} \binom{n-1}{k-1} \frac{N_t}{\binom{n}{k}} \mathbb{P}\left[X_1 = k \middle| \sum_{i=1}^{N_t} X_i = n \right] \\
&= \sum_{k=1}^{n} N_t \frac{k}{n} \mathbb{P}\left[X_1 = k \middle| \sum_{i=1}^{N_t} X_i = n \right] \\
&= \frac{N_t}{n} \sum_{k=0}^{n} k \mathbb{P}\left[X_1 = k \middle| \sum_{i=1}^{N_t} X_i = n \right] \\
&= \frac{N_t}{n} \mathbb{E}\left[X_1 \middle| \sum_{i=1}^{N_t} X_i = n \right] \\
&= \frac{1}{n} \sum_{i=1}^{N_t} \mathbb{E}\left[X_i \middle| \sum_{i=1}^{N_t} X_i = n \right] \\
&= \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{N_t} X_i \middle| \sum_{i=1}^{N_t} X_i = n \right] \\
&= 1
\end{aligned} \tag{7}$$

3 Poisson offspring distribution case

In this section we consider that the offspring distribution is $\alpha_t = \text{Poisson}(R_t)$. In this case, we have:

$$\sum_{i=1}^{N_t} S_{t,i} \sim \text{Poisson}(N_t R_t) \tag{8}$$

and the conditional distribution:

$$
\begin{aligned}
\mathbb{P}\left[S_{t,1} = s \,\middle|\, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] &= \frac{\mathbb{P}\left[S_{t,1} = s, \displaystyle\sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]}{\mathbb{P}\left[\displaystyle\sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]} \\
&= \frac{\alpha_t(s)\,\mathbb{P}\left[\displaystyle\sum_{i=2}^{N_t} S_{t,i} = N_{t+1} - s\right]}{\mathbb{P}\left[\displaystyle\sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]} \\
&= \frac{\dfrac{R_t^s e^{-R_t}}{s!} \cdot \dfrac{((N_t - 1)R_t)^{N_{t+1}-s}}{(N_{t+1} - s)!}}{\dfrac{(N_t R_t)^{N_{t+1}} e^{-N_t R_t}}{N_{t+1}!}} \\
&= \binom{N_{t+1}}{s}\left(\frac{1}{N_t}\right)^s \left(1 - \frac{1}{N_t}\right)^{N_{t+1}-s}
\end{aligned}
\tag{9}
$$

This is the probability mass function of a Binomial distribution and therefore we deduce that:

$$
S_{t,1} \,\middle|\, \left(\sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right) \sim \text{Binomial}\left(N_{t+1}, \frac{1}{N_t}\right) \tag{10}
$$

The k-th falling factorial moments of $X \sim \text{Binomial}(n, p)$ are (**?**):

$$
\mathbb{E}\left[\frac{X!}{(X-k)!}\right] = \binom{n}{k} p^k k! \tag{11}
$$

By applying this formula to the Binomial distribution in Equation 10 and injecting into Equation 3, we deduce that the inclusive probability of coalescence for k lines is:

$$
p_{k,t}(N_t, N_{t+1}) = \frac{1}{N_t^{k-1}} \tag{12}
$$

In addition, following a similar reasoning as for Equation 10 we can show that:

$$
X_{t,1} \,\middle|\, \left(\sum_{i=1}^{N_t} X_{t,i} = n\right) \sim \text{Binomial}\left(n, \frac{1}{N_t}\right) \tag{13}
$$

By injecting the probability mass function of this Binomial distribution into Equation 5 we deduce that the exclusive probability of coalescence for k lines from a sample of n ($n \geq k$) is:

$$p_{n,k,t}(N_t, N_{t+1}) = \frac{(N_t - 1)^{n-k}}{N_t^{n-1}} \tag{14}$$

The definitions of the inclusive and exclusive coalescence probabilities imply that the former is a special case of the latter, with specifically $p_{k,t}(N_t, N_{t+1}) = p_{k,k,t}(N_t, N_{t+1})$ and this is clearly verified in Equations 12 and 14. It is interesting to note that neither the inclusive nor the exclusive coalescence probability depend on the mean R_t of the Poisson offspring distribution or the size N_{t+1} of the population at time $t+1$. Both only depend on the population size N_t at time t. The intuition behind this result is the same as for the models of **?**: each individual at time $t+1$ has the same probability of having any ancestor at time t, irrespective of the population size at time $t+1$ and of the offspring distribution of the ancestors at time t, as long as they are exchangeable. The inclusive coalescent probability in Equation 12 can also be obtained conceptually by considering that among the k lines, the first one has an ancestor with probability one, and the remaining $k-1$ need to have the same ancestor among a set of N_t from which they choose uniformly at random so that the probability of picking the same ancestor is $1/N_t$. The exclusive coalescent probability in Equation 14 can be derived likewise by considering that in addition to the above, each of the $n-k$ other lines need to choose a different ancestor, which happens with probability $(N_t - 1)/N_t$. Figure 1 illustrates the inclusive and exclusive coalescence probabilities for a set of size $k = 1$ to $k = 10$ amongst a total of $n = 10$ observed individuals, in a population of size $N_t = 10$, $N_t = 20$ or $N_t = 30$.

4 Negative-Binomial offspring distribution case

In this section we consider that the offspring distribution is Negative-Binomial, a distribution often used to model superspreading individuals (**?**) and which can also be used to model superspreading events (**?**). Let $\alpha_t =$ Negative-Binomial(r, p) with parameters (r, p) set by moment-matching the mean R_t and variance V_t of the offspring distribution which are assumed constant over time. The resulting parameters for this distribution are $r = R_t^2/(V_t - R_t)$ and $p = R_t/V_t$. In this case, we have:

$$\sum_{i=1}^{N_t} S_{t,i} \sim \text{Negative-Binomial}(N_t r, p) \tag{15}$$

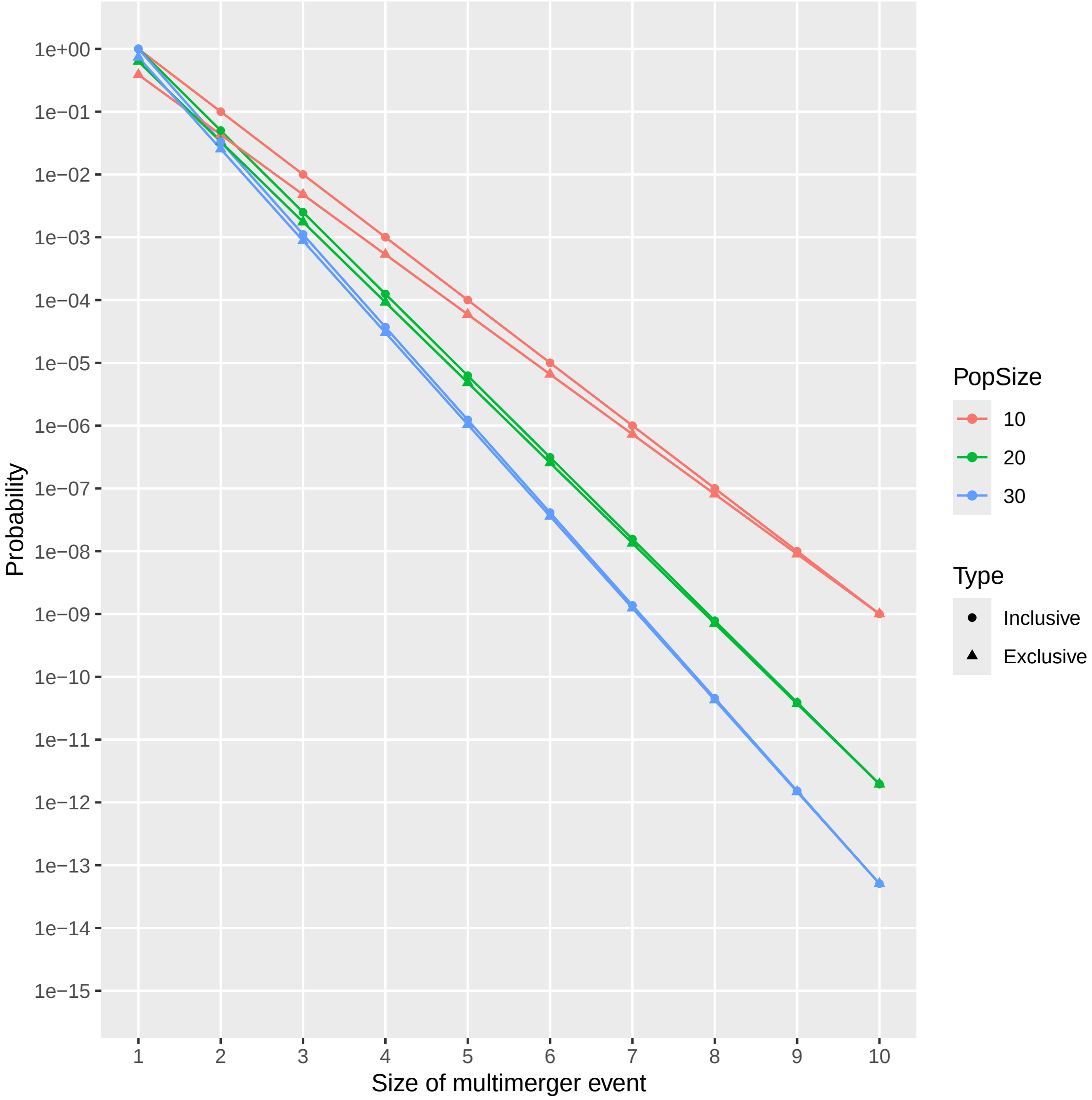

Figure 1: Inclusive and exclusive coalescence probabilities for the Poisson case.

and similarly to the Poisson offspring distribution case we identify that the conditional distribution of $S_{t,1}|\sum_{i=1}^{N_t} S_{t,i}$ is as follows:

$$\begin{aligned}
\mathbb{P}\left[S_{t,1}=s\,\middle|\,\sum_{i=1}^{N_t} S_{t,i}=N_{t+1}\right] &= \frac{\alpha_t(s)\cdot\mathbb{P}\left[\sum_{i=2}^{N_t} S_{t,i}=N_{t+1}-s\right]}{\mathbb{P}\left[\sum_{i=1}^{N_t} S_{t,i}=N_{t+1}\right]} \\
&= \frac{\dfrac{\Gamma(r+s)}{s!\Gamma(r)}(1-p)^s p^r\cdot\dfrac{\Gamma\big((N_t-1)r+(N_{t+1}-s)\big)}{(N_{t+1}-s)!\Gamma((N_t-1)r)}(1-p)^{N_{t+1}-s}p^{(N_t-1)r}}{\dfrac{\Gamma(N_t r+N_{t+1})}{N_{t+1}!\Gamma(N_t r)}(1-p)^{N_{t+1}}p^{N_t r}} \\
&= \frac{N_{t+1}!}{s!(N_{t+1}-s)!}\frac{\Gamma(r+s)\Gamma\big((N_t-1)r+(N_{t+1}-s)\big)}{\Gamma(N_t r+N_{t+1})}\frac{\Gamma(N_t r)}{\Gamma(r)\Gamma\big((N_t-1)r\big)} \\
&= \binom{N_{t+1}}{s}\frac{\mathrm{B}(s+r,N_{t+1}-s+(N_t-1)r)}{\mathrm{B}(r,(N_t-1)r)} \qquad (16)
\end{aligned}$$

where $\mathrm{B}(x,y)$ denotes the Beta function defined as $\mathrm{B}(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y)=\int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t$. This is the probability mass function of a Beta-Binomial distribution and therefore we deduce that:

$$S_{t,1}\,\middle|\left(\sum_{i=1}^{N_t} S_{t,i}=N_{t+1}\right)\sim\text{Beta-Binomial}(N_{t+1},r,(N_t-1)r) \qquad (17)$$

The k-th falling factorial moments of $X\sim$ Beta-Binomial(n,α,β) are (**?**):

$$\mathbb{E}\left[\frac{X!}{(X-k)!}\right]=\binom{n}{k}\frac{\mathrm{B}(\alpha+k,\beta)k!}{\mathrm{B}(\alpha,\beta)} \qquad (18)$$

By applying this formula to the Beta-Binomial distribution in Equation 17 and injecting into Equation 3, we deduce that the inclusive probability of coalescence for k lines is:

$$p_{k,t}(N_t,N_{t+1})=\frac{\mathrm{B}(N_t r+1,r+k)}{\mathrm{B}(r+1,N_t r+k)} \qquad (19)$$

In addition, following a similar reasoning as for Equation 17, we can show that:

$$X_{t,1}\bigg|\left(\sum_{i=1}^{N_t} X_{t,i} = n\right) \sim \text{Beta-Binomial}(n, r, (N_t-1)r) \tag{20}$$

By injecting the probability mass function of this Beta-Binomial distribution into Equation 5 we deduce that the exclusive probability of coalescence for k lines is:

$$p_{n,k,t}(N_t, N_{t+1}) = \frac{N_t \mathrm{B}(k+r, n-k+N_t r - r)}{\mathrm{B}(r, N_t r - r)} \tag{21}$$

As for the Poisson case we can check that $p_{k,t}(N_t, N_{t+1}) = p_{k,k,t}(N_t, N_{t+1})$ using the formulas in Equations 19 and 21 and the definition of the Beta function. It is interesting to note that as for the Poisson case, the inclusive and exclusive coalescence probabilities do not depend on the size N_{t+1} of the population at time $t+1$. They both depend on the Negative-Binomial offspring distribution only through the dispersion parameter r. If we consider that r is large in Equations 19 and 21, we can derive that the asymptotic behaviour is the same as in the Poisson case shown in Equations 12 and 14. For example this can be derived by rewriting the Beta functions using Gamma functions, and using the following form of Stirling's approximation:

$$\mathrm{lim}_{a\to\infty} \frac{\Gamma(a+b)}{\Gamma(a)} = a^b \mathrm{e}^{-b} \tag{22}$$

Figure 2 illustrates the inclusive and exclusive coalescence probabilities for the Negative-Binomial case for a set of size $k=1$ to $k=10$ amongst a total of $n=10$ observed lines, in a population with size $N_t = 20$. Several Negative-Binomial offspring distributions are compared, all of which have the same mean $R_t = 2$, and with the dispersion parameter equal to $r=0.1$, $r=1$, $r=10$ and $r=100$ (Figure 2, inset). When $r=1$ the Negative-Binomial reduces to a Geometric distribution. When r is high the dispersion is low and the Negative-Binomial case behaves almost like the Poisson case for both the inclusive and the exclusive coalescence probabilities. When r is lower the dispersion of the offspring distribution increases, so that both the inclusive and exclusive probabilities of larger multiple merger events are increased compared to the Poisson case. In particular, when $r=0.1$ we see that the exclusive probability can increase with the size of the event considered (Figure 2). This happens because the probability is not much lower for the common ancestor having say 10 rather than 9 offspring, while on the other hand if the event is of size 9 only then another individual in the generation of the ancestor needs to have had at least one sampled offspring.

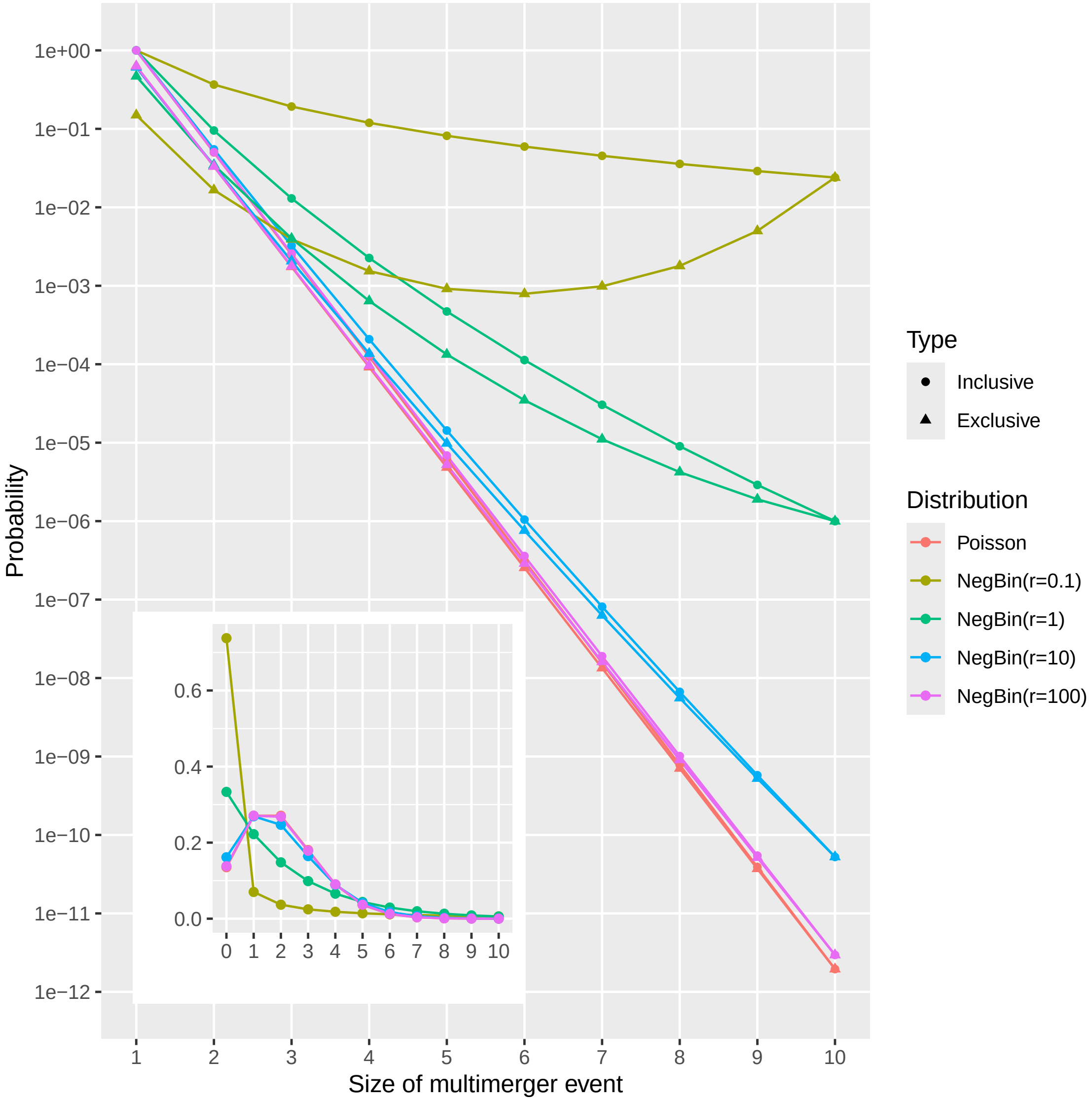

Figure 2: Inclusive and exclusive coalescence probabilities with $N_t = 20$ and $n = 10$ and for several offspring distributions. Inset: offspring distributions shown, all with mean $R_t = 2$.

5 Limit when the population size is large

Here we consider that the population size N_t is fixed and large, so that we can show the connections between our results and several previous studies on the ancestral process of infectious diseases. We only consider the population made of infected individuals, with no representation of uninfected individuals, so that the notation N_t is consistent with its previous use so far. For a population of large size to have multiple merger events, the number of individuals contributed by a single parent in the previous generation needs to not be negligibly small compared to the population size, so that it is possible for an individual to at least occasionally produce enough offspring to contribute a non-negligible fraction of the population (**??**). In the Poisson case, from Equations 12 and 14 we can see that both inclusive and exclusive probabilities are of order $\mathcal{O}(N_t^{1-k})$. When time is measured in units scaled by the population size as in the standard coalescent model (**?**), we can therefore ignore events with $k > 2$ and retain only the events with $k = 2$, which means that there are only binary coalescent events and no multiple merger events. The binary coalescent events occur with the same inclusive and exclusive probabilities:

$$p_{2,t}(N_t, N_{t+1}) = p_{n,2,t}(N_t, N_{t+1}) = \frac{1}{N_t} \tag{23}$$

For the Negative-Binomial case, from Equations 19 and 21 we can rewrite using Gamma functions and apply the form of Stirling's equation given in Equation 22 to show that once again both inclusive and exclusive probabilities are also of order $\mathcal{O}(N_t^{1-k})$. We can therefore once again ignore events with $k > 2$ and retain only the events with $k = 2$ which occur with the same inclusive and exclusive probabilities:

$$p_{2,t}(N_t, N_{t+1}) = p_{n,2,t}(N_t, N_{t+1}) = \frac{r+1}{N_t r + 1} \approx \frac{r+1}{N_t r} \tag{24}$$

? derived the rates of coalescence of two lineages for several epidemiological models, assuming a large population at equilibrium. For each model they use the equation $N_\mathrm{e} = N/\sigma^2$ to relate the effective population size N_e to the actual population size N and the variance σ^2 in the number of offspring. This relationship was first established by **?** to derive the backward-in-time coalescent model from the forward-in-time Cannings exchangeable models (**?**). This result implies that the rate of coalescence for two lineages is $1/N_\mathrm{e} = \sigma^2/N$. From Equation 24 we can take $R_t = 1$ to achieve equilibrium of the population size and the method of moments estimator $r = R_t^2/(V_t - R_t) = 1/(V_t - 1)$ to deduce the equivalent result $p_{2,t}(N_t, N_{t+1}) = V_t/N_t$.

? showed that the rate of coalescence for two lineages under a continuous-time epidemic coalescent model is $2f(t)/I(t)^2$ where $f(t)$ is the incidence of the disease and $I(t)$ its prevalence. Setting in this formula the prevalence as $I(t) = N_{t+1} = N_t R_t$ and the incidence as $f(t) = R_t I(t) = R_t^2 N_t$ we get a coalescent rate of $2/N_t$. To apply our methodology we need to consider that the offspring distribution is Geometric, since the epidemiological models considered have successes (transmission) happening until the first failure (removal). We therefore set $r = 1$ in Equation 24 to make the Negative-Binomial offspring distribution reduce to a Geometric distribution and the same result follows.

? calculated the effective population size $N_\mathrm{e}(t)$ as a function of the actual population size $N(t)$ and the mean and variance of the offspring distribution R and σ^2. This formula was used to estimate the dispersion parameter of a Negative-Binomial offspring distribution from genetic data (**?**). Using our notations, their formula is equivalent to the inclusive coalescence probability for two lineages:

$$p_{2,t}(N_t, N_{t+1}) = \frac{V_t/R_t + R_t - 1}{N_t R_t} \tag{25}$$

In the Poisson case we have $V_t = R_t$ so that Equation 25 simplifies to $1/N_t$ which agrees with Equation 23. In the Negative-Binomial case we have $V_t/R_t = 1/p = 1 + R_t/r$ so that Equation 25 simplifies to $(r+1)/(N_t r)$ which agrees with our Equation 24. Conversely, if we substitute the method of moments estimator $r = R_t^2/(V_t - R_t)$ in Equation 24 we obtain the Equation 25 originally from **?**.

6 Definition of a new Lambda-coalescent model

The coalescent model (**??**) describes the ancestry of a sample from a large population evolving according to many forward-in-time models such as the Wright-Fisher model (**??**), the Moran model (**?**) and the Cannings exchangeable model (**?**). Since the coalescent considers a large population in which each individual only has a number of offspring that is small compared to the population size, coalescent trees are always binary and do not feature multiple mergers, making them unsuitable to represent the ancestry of outbreaks considered in this study. However, if the population size is small in any of the aforementioned forward-in-time models, multiple mergers can occur, where more than two sampled individuals have the same ancestor. The Lambda-coalescent model is an extension of the coalescent model that allows for such multiple merger events (**???**).

A Lambda-coalescent model is defined by a Lambda measure $\Lambda(\mathrm{d}x)$ on the interval $[0, 1]$, from which we deduce the rate $\lambda_{n,k}$ at which any subset of k lineages within a set of n observed lineages coalesce:

$$\lambda_{n,k} = \int_0^1 x^{k-2}(1-x)^{n-k}\,\Lambda(\mathrm{d}x) \tag{26}$$

The Beta-coalescent (**?**) is a specific type of Lambda-coalescent that has been used recently in several studies analysing genetic data from infectious disease agents (**????**). The Beta-coalescent model has a single parameter $\alpha \in [0,2]$ and is defined as:

$$\Lambda(\mathrm{d}x) = \frac{x^{1-\alpha}(1-x)^{\alpha-1}}{\mathrm{B}(2-\alpha,\alpha)}\mathrm{d}x \tag{27}$$

By combining Equations 26 and 27 we deduce that:

$$\lambda_{n,k} = \frac{\mathrm{B}(k-\alpha, n-k+\alpha)}{\mathrm{B}(2-\alpha,\alpha)} \tag{28}$$

Special cases of the Beta-coalescent include $\alpha = 2$ corresponding to the Kingman coalescent, $\alpha = 1$ which is known as the Bolthausen-Sznitman coalescent and $\alpha = 0$ for which the phylogeny is always star-shaped.

We now define a new Lambda-coalescent based on the Negative-Binomial case described previously. We call this new Lambda-coalescent model the Omega-coalescent (where Omega stands for outbreak). For ease of comparison with other coalescent models, we consider that time is continuous and that the population size remains constant equal to $N_t = N$. The exclusive coalescent probability $p_{n,k,t}(N_t, N_{t+1})$ in the Negative-Binomial case given by Equation 21 can be used to determine the corresponding rate of the Omega-coalescent, if we consider that the probability of each event in discrete time is equal to the constant rate of this event happening in continuous time:

$$\lambda_{n,k} = p_{n,k,t}(N_t = N, N_{t+1} = N) = \frac{N\mathrm{B}(k+r, n-k+Nr-r)}{\mathrm{B}(r, Nr-r)} \tag{29}$$

Note that this equation implies that continuous time is measured approximately in number of transmission generations. For example to measure time in decimal days instead, the time scale would need to be multiplied by the mean of the generation time distribution measured in days (**?**). From

Equations 26 and 29 we can deduce the Lambda measure associated with the Omega-coalescent:

$$\Lambda(\mathrm{d}x) = \frac{Nx^{r+1}(1-x)^{Nr-r-1}}{\mathrm{B}(r, Nr-r)}\mathrm{d}x \tag{30}$$

For a Lambda-coalescent model to be consistent, when a multiple merger of size k amongst n lineages occurs, if an additional lineage is revealed it must either take part in the multiple merger or remain unaffected (**??**). This implies that the rates must satisfy:

$$\lambda_{n,k} = \lambda_{n+1,k} + \lambda_{n+1,k+1} \tag{31}$$

This consistency property is easily verified for the Beta-coalescent in Equation 28 and likewise for the Omega-coalescent in Equation 29, in both cases using recursive properties of the Beta functions used in the respective definitions.

The Omega-coalescent has two parameters: the constant population size N and the dispersion parameter r. In order to compare the Omega-coalescent defined in Equation 29 with other models such as the Beta-coalescent defined in Equation 28, we consider the distribution of the size k of the next event among a set of n lineages. For any Lambda-coalescent this can be computed as:

$$p(k|n) = \frac{\binom{n}{k}\lambda_{n,k}}{\sum_{i=2}^{n}\binom{n}{i}\lambda_{n,i}} \tag{32}$$

Figure 3 compares this distribution for $n = 10$ in the Beta-coalescent with parameter $\alpha \in \{0.5, 1, 1.5\}$ and for the Omega-coalescent with parameters $N \in \{10, 20, 30\}$ and $r \in \{0.1, 1, 10\}$. In the Beta-coalescent, the distribution shifts towards more larger multiple merger events as the parameter α decreases. In the Omega-coalescent a wider range of behaviours is obtained when varying the two parameters N and r. For a given value of N, decreasing the value of r results in more larger events. Conversely, for a given value of r we can see that increasing the value of N reduces the probability of larger events.

Genealogies can be simulated from the Omega-coalescent model defined in Equation 29 using the same algorithm as for other Lambda-coalescent models (**?**). Given n lineages, the next coalescent event happens after a time that is exponentially distributed with rate $\sum_{i=2}^{n}\binom{n}{i}\lambda_{n,i}$, the size k of this event is drawn according to Equation 32, and the k lineages that coalesce are chosen uniformly amongst the n lineages. This process is repeated iteratively until all lineages have coalesced. Figure 4 shows

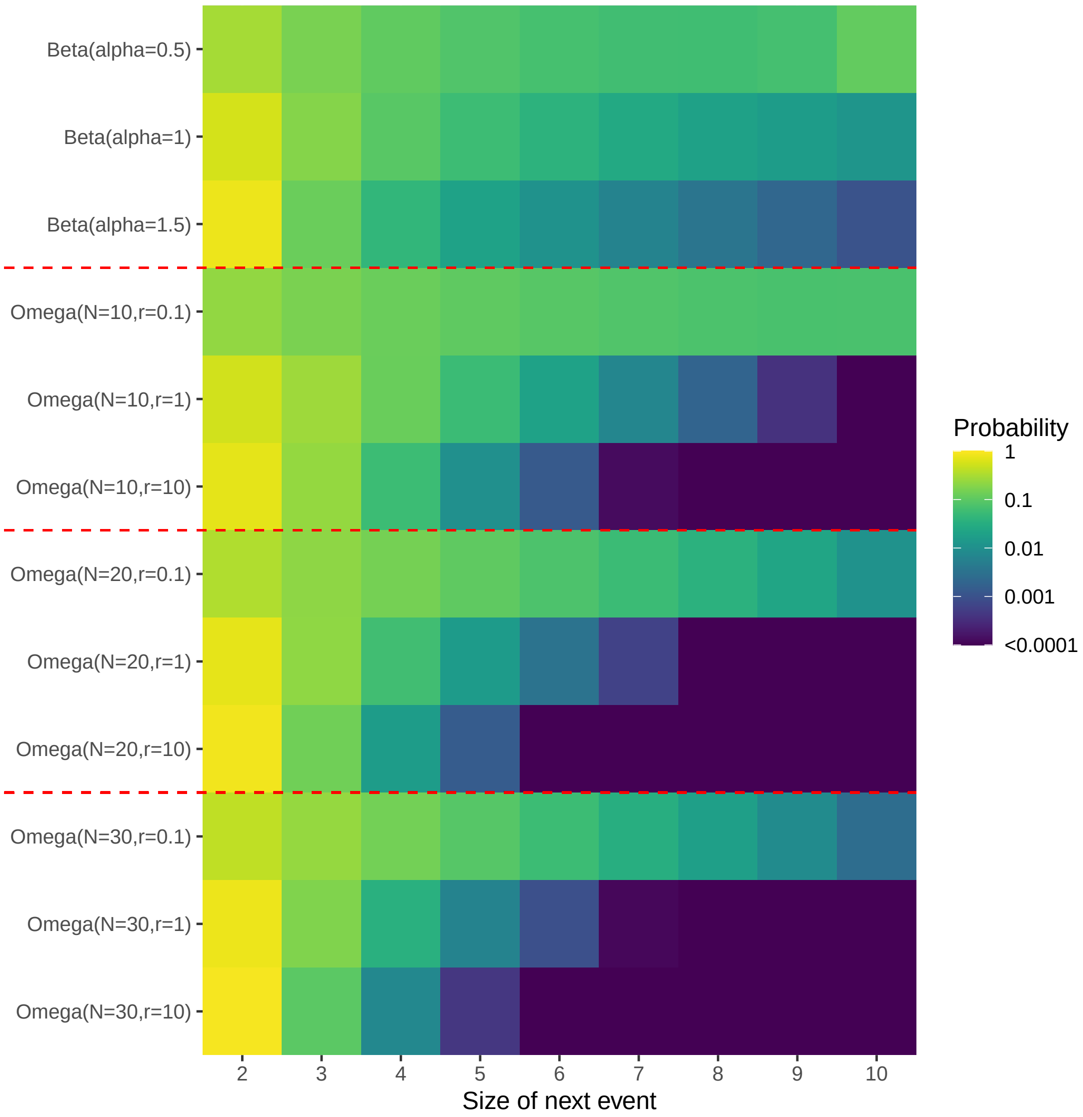

Figure 3: Distribution of the size of the next event among a set of $n = 10$ lineages, compared between the Beta-coalescent and the Omega-coalescent model with various parameters.

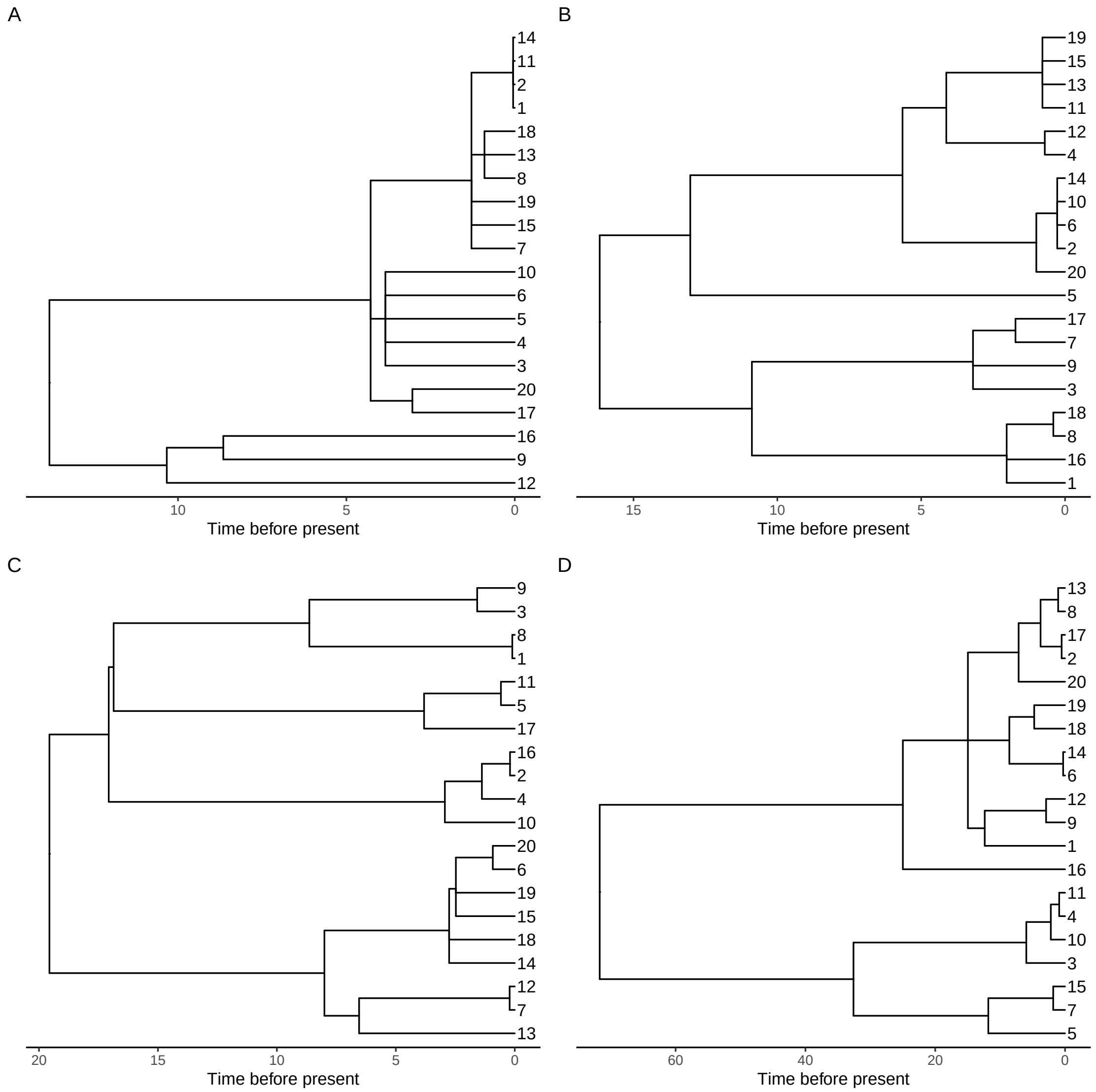

Figure 4: Example of trees simulated under the Omega-coalescent with $r = 0.1$ (A), $r = 1$ (B), $r = 10$ (C) and $r = 100$ (D).

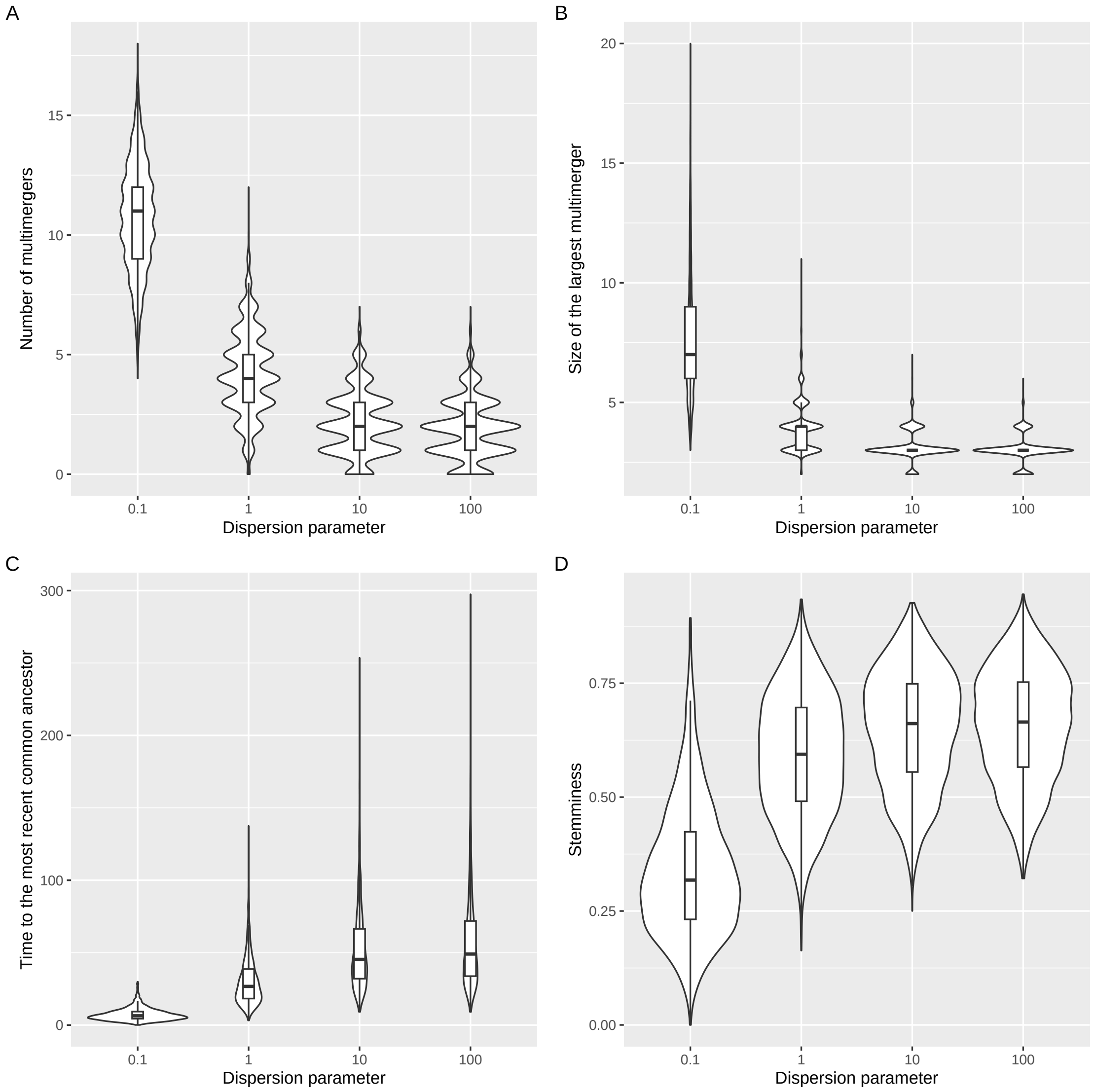

Figure 5: Summary statistics for trees simulated under the Omega-coalescent with $r = 0.1$, $r = 1$, $r = 10$ and $r = 100$, namely number of multiple mergers (A) the size of the largest multiple merger (B), the time to the most recent common ancestor (C) and the stemminess (D).

examples of trees simulated for a sample of size $n = 20$, constant population size $N = 30$ and dispersion parameter $r \in \{0.1, 1, 10, 100\}$. It is already clear from these single realisations that the lower values of r result in trees with more larger multiple merger events and lower time to the most recent common ancestor, but to quantify these properties we need to consider many trees. Figure 5 shows summary statistics for 10,000 trees simulated in the same conditions as the individual trees shown in Figure 4. As the dispersion parameter increases from $r = 0.1$ to $r = 100$ multiple merger events become less and less likely and less large (Figure 5A and B), and the time to the most recent common ancestor increases (Figure 5C). Furthermore, the stemminess of the tree increases, which is defined as the sum of lengths of internal branches divided by the total sum of branch lengths (Figure 5D). Stemminess is usually taken as a sign of population size dynamics (**??**), which would be misleading here since all simulations assumed a constant population size. The patterns of summary statistics shown in Figure 5 for $N = 30$ and $n = 20$ are consistent across different values of these parameters.

7 Parameter inference

Let us now consider a genealogy T with n leaves and c coalescent nodes, with $t_0 = 0$ the sampling time, $t_1, ..., t_c$ the times of the coalescent nodes in increasing order and k_i the number of lineages coalescing at time t_i. The number of lineages existing between time t_{i-1} and t_i is then $n_i = n - \sum_{j=1}^{i-1} k_j$. Under a Lambda-coalescent model, the genealogy T has likelihood:

$$p(T|\Lambda) = \prod_{i=1}^{c} \lambda_{n_i,k_i} \exp\left(- \sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i,j} (t_i - t_{i-1}) \right) \tag{33}$$

Note that in Equation 33 the term $\binom{n_i}{k_i}$ term from the coalescent rate cancels out with its reciprocal from the probability of sampling k_i specific lineages to coalesce within a set of n_i. Estimating the Lambda measure from Equation 26 in general is a difficult problem (**??**). Here however we focus on estimation under the Omega-coalescent model, where the $\lambda_{n,k}$ terms are given by Equation 29. There are therefore two parameters to estimate which have direct and important biological meaning: the effective population size N (which remains constant) and the dispersion parameter r of the Negative-Binomial offspring distribution. We perform estimation simply by maximising the likelihood in Equation 33, using the Brent algorithm (**?**) when estimating a single parameter and the L-BFGS-B algorithm (**?**) when estimating both parameters.

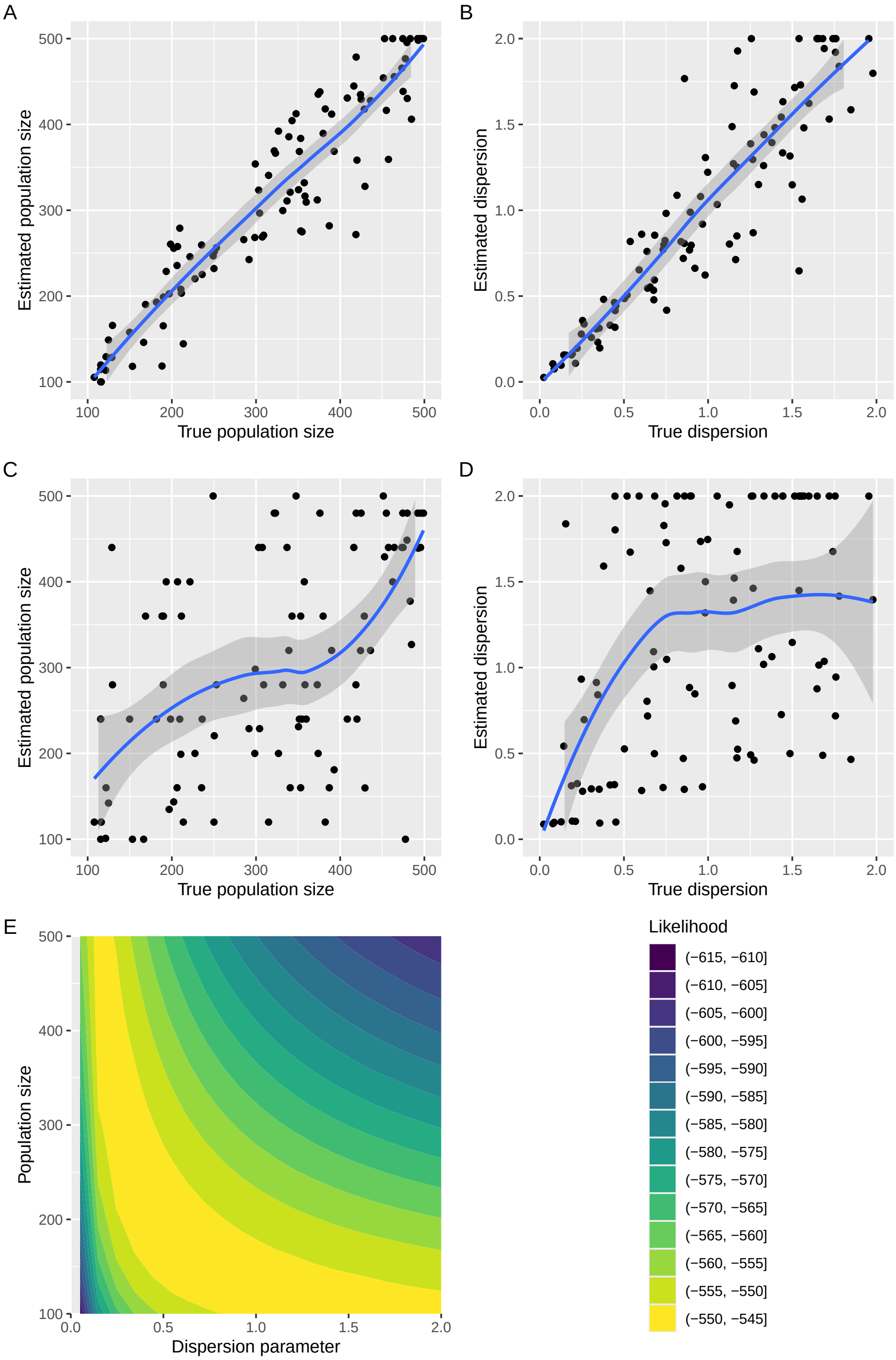

Figure 6: Maximum likelihood estimation of parameters. (A) Estimation of the population size given the dispersion parameter. (B) Estimation of the dispersion parameter given the population size. (C and D) Joint estimation of both the population size and dispersion parameters. (E) Example of likelihood surface as a function of both parameters.

We simulated 100 genealogies from the Omega-coalescent model each of which had $n = 100$ leaves, with parameter N drawn uniformly at random between 100 and 500 and parameter r drawn uniformly at random between 0.01 and 2. If we assume knowledge of the dispersion parameter, then estimating the population size works really well (Figure 6A). Conversely we obtain good result when estimating the dispersion parameter given a known population size (Figure 6B). However, attempting to estimate both parameters at the same time performed significantly less well (Figures 6C and D). To illustrate the cause of this, we consider a simulation for which the true parameters were $N = 200$ and $r = 0.5$, and we construct the likelihood surface (Figure 6E). This shows a strong inverse tradeoff between the two parameters, which is why it is harder to infer both parameters jointly. This poor identifiability is analogous to the situation of a large population following the Cannings models (?). In this case the coalescent process is fully determined by the effective population size $N_\mathrm{e} = N/\sigma^2$ as previously noted (?), where N is the population size and σ^2 is the variance in the number of offspring. Consequently there is a full tradeoff between N and σ^2, so that the ratio N_e can be estimated but not the parameters N and σ^2 separately.

8 Implementation

We implemented the analytical methods described in this paper in a new R package entitled *EpiLambda* which is available at `https://github.com/xavierdidelot/EpiLambda` for R version 3.5 or later. All code and data needed to replicate the results are included in the “run” directory of the *EpiLambda* repository. The R package `ape` was used to store, manipulate and visualise phylogenetic trees (?).

9 Discussion

We have described an ancestral process for infectious diseases which is relevant to the analysis of outbreaks of a relatively small size, and to diseases with transmission heterogeneity. We have shown how this process can be incorporated into a new Lambda-coalescent which we called the Omega-coalescent. We only considered the situation where all samples are taken at the same time, but the Omega-coalescent could be extended to allow temporally offset leaves following similar work on the coalescent (?) and the Beta-coalescent (?). We also made the simplifying assumption of a constant population size, but this choice was motivated by mathematical convenience and simplicity rather than biological realism. Considering the Omega-coalescent with variable population size in future work could be especially important, since this model aims to describe relatively small outbreaks, in which the number of infected individuals is likely to vary significantly. For example, in a stochastic version of the Susceptible-Infectious-Susceptible (SIS) or Susceptible-Infectious-Recovered (SIR) models with basic reproduction number greater than one, the early phase of a new outbreak is expected to show exponential growth, with much stochasticity (?). Furthermore, we showed that the probability of multiple merger events of various sizes depends explicitly on the population size in Equation 21. Changes in population size will therefore have an effect on the distribution of events observed, as can be seen for example in Figure 3. It should be possible to relax the constant size assumption following the same approach as previously described for integrating variable population size into the coalescent (???) and the Beta-coalescent (??). We showed that it is difficult to jointly infer a fixed population size with the dispersion parameter (Figure 6). Bayesian inference with an informative prior on at least one of these two parameters could help resolve this situation. Allowing the population size to vary could also help with the identifiability of these two parameters, since their tradeoff would not be the same for different values of the population size. Investigating the Omega-coalescent with variable population size therefore seems a promising avenue for future research.

We compared the Omega-coalescent only to the Beta-coalescent (?) in Figure 3 as it is the model that has been most frequently used for infectious diseases (???). Several other Lambda-coalescent models have been proposed previously, such as the Dirac coalescent (?), the Durrett-Schweinsberg coalescent (?) or the extended Beta-coalescent (?). However, none of these models is equivalent to the Omega-coalescent model. Indeed these previously described Lambda-coalescent models are mostly concerned with situations where an individual can be the father of a significant portion of a population in spite of the population being large, as opposed to the small populations with superspreading we considered here. The xi-coalescent models are extensions to the Lambda-coalescent models that admit multiple simultaneous mergers (?). This is clearly relevant to our basic discrete time model for small outbreaks, since in small populations it is quite likely that separate subsets of individuals have the same infector in the previous generation. However the exact timing of ancestry events is never available so that we must rely on ancestral dating estimation with no notion of event co-occurrence (????). We therefore introduced a continuous time approximation in Equation 29 so that ancestry events do not co-occur. The exact coalescent process for the discrete time Wright-Fisher process in a small population has been previously described (?). It would be difficult however to extend this approach to the more complex forward-in-time model we considered here, with variable population size and specific offspring distribution, which further justifies our continuous time approximation.

Finally, it should be noted that our model describes the transmission tree during an outbreak, which is different from a phylogeny (?). This difference is often ignored and in some settings it might be appropriate to do so, but not always. Consequently, some previous studies have used models of within-host evolution to bridge the gap between transmission and phylogenetic trees (???). However, these models assume that each transmission event happens independently from one infector to each of its infectees. This is not necessarily true especially when considering superspreading events in which many individuals can become infected simultaneously (????). In conclusion, we have described a new ancestral model for infectious disease outbreaks, which we hope will be useful especially in settings where the outbreaks are small or in the presence of high transmission heterogeneity.

Acknowledgements

We acknowledge funding from the National Institute for Health Research (NIHR) Health Protection Research Unit in Genomics and Enabling Data.