

1 Ancestral process for infectious disease outbreaks with superspreading

2 Xavier Didelot^{1,2,*}, David Helekal³, Ian Roberts², ...

3 ¹ School of Life Sciences, University of Warwick, Coventry, United Kingdom

4

5 ² Department of Statistics, University of Warwick, Coventry, United Kingdom

6

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

9

10 * Corresponding author. Tel: 0044 (0)2476 572827. Email: xavier.didelot@gmail.com

11 Running title: Ancestry for outbreaks with superspreading

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

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 and infect each of the susceptible individuals, who will then become infected in their turn and spread the disease further. Most infectious disease modelling theory describes 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 (Anderson and May 1991; Keeling and Rohani 2008). Here however we focus on the early stages of an epidemic, where the number of infected individuals is small and the number of susceptibles relatively high and unchanging. In this situation it is useful to think about the number of infections that each newly infected individual is likely to cause, and the probabilistic distribution for this number is often called the offspring distribution (Grassly and Fraser 2008). 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 (Fraser et al. 2004; Ferguson et al. 2006).

If we consider that all individuals are infectious for the same duration and with the same infectiousness, the offspring distribution is Poisson distributed with mean R_0 , which means that 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, or being more infectious than others, or having more 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 (Lloyd-Smith et al. 2005; Grassly and Fraser 2008). 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 (Woolhouse et al. 1997; Stein 2011; Kucharski and Althaus 2015; Wang et al. 2021), and most recently for SARS-CoV-2 (Wang et al. 2020; Lemieux et al. 2021; Gómez-Carballa et al. 2021; Du et al. 2022).

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 sample 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 or 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 if the population size is large, our results agree with several previous studies (Volz 2012; Koelle and Rasmussen 2012; Fraser and Li 2017). Finally, we show how our results can be incorporated into a new lambda-coalescent model (Pitman 1999; Sagitov 1999; Donnelly and Kurtz 1999) and compare it with previously described models.

2 General case

Let time be measured in discrete units and denoted t . Each discrete value of t correspond to a unique non-overlapping generations of infected individuals, so that individuals infected at t will 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 . We have:

$$N_{t+1} = \sum_{i=1}^{N_t} s_{t,i} \quad (1)$$

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 $t + 1$ find a common ancestor in generation t , conditional on population sizes N_t and N_{t+1} .

Given full information about offspring counts from individuals in generation t , $\mathbf{s}_t = (s_{t,1}, \dots, s_{t,N_t})$, we have

$$\begin{aligned} p_{k,t}(\mathbf{s}_t, N_t) &= \sum_{i=1}^{N_t} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \\ &= \sum_{i=1}^{N_t} \frac{\Gamma(s_{t,i} + 1) \Gamma(N_{t+1} - k + 1)}{\Gamma(s_{t,i} - k + 1) \Gamma(N_{t+1})} \end{aligned} \quad (2)$$

Full information $\{s_{t,i}\}$ yields the population size N_{t+1} but is not feasible to observe in practice. We can instead express the inclusive coalescence probability conditioning on the next population size N_{t+1} by summing over possible offspring counts $\mathbf{s}_t = (s_{t,1}, \dots, s_{t,N_t})$ conditional on the total generation size. Let $S_t^{-(1)} = (S_{t,2}, \dots, S_{t,N_t})$.

$$\begin{aligned} p_{k,t}(N_t, N_{t+1}) &= \sum_{\mathbf{s}_t \in \mathbb{N}_0^{N_t}} \mathbb{P} \left[\mathbf{S}_t = \mathbf{s}_t \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] p_{k,t}(\mathbf{s}_t, N_t) \\ &= \sum_{\mathbf{s}_t \in \mathbb{N}_0^{N_t}} \mathbb{P} \left[\mathbf{S}_t = \mathbf{s}_t \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] \sum_{i=1}^{N_t} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \\ &= \sum_{i=1}^{N_t} \sum_{\mathbf{s}_t \in \mathbb{N}_0^{N_t}} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \mathbb{P} \left[S_{t,1} = s_{t,1}, \mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] \end{aligned}$$

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

75

76 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
77 obtained by differentiating the probability generating function of $S_{t,1} | (\sum_{i=1}^{N_t} S_{t,i} = N_{t+1})$.

78 2.2 Exclusive coalescence probability

79 Generally, we observe a sample of individuals from each generation rather than the entire population.
80 In this case, we are interested in the exclusive coalescence probability $p_{n,k,t}(N_t, N_{t+1})$ that exactly k
81 individuals from a sample of n arose from a common ancestor one generation in the past given knowledge
82 of the total population sizes N_t and N_{t+1} .

83 Given full information about offspring counts of the parents of sampled individuals at the present,
84 $\mathbf{x}_t = (x_{t,1}, \dots, x_{t,N_t})$, we have

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

85 Similarly to the exclusive coalescence probability, we can use this to evaluate the exclusive probability
86 given N_t and N_{t+1} by summing over possible parent offspring configurations (for $k \leq n$),

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

87 Note that $X_{t,i}$ does not follow the same offspring distribution as $S_{t,i}$. $(X_{t,1}, \dots, X_{t,N_t})$ consists of n
 88 individuals sampled from generation $t+1$ without replacement - there is no guarantee that all offspring
 89 from any given parent are included in the sample.

90 **2.3 Complementarity of exclusive coalescence probabilities**

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

$$\sum_{k=1}^n \binom{n-1}{k-1} p_{n,k,t} = 1 \tag{6}$$

94 We can show that it is indeed satisfied by the formula in Equation (5):

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

95 3 Poisson case

96 Here the offspring distribution is $\alpha_t = \text{Poisson}(R_t)$. In this case, we have

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

97 and conditional distribution

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

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

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

100 The k -th falling factorial moments of $X \sim \text{Binomial}(n, p)$ are (Potts 1953):

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

101 By injecting this formula into Equation (3) we obtain the inclusive probability of coalescence for k
102 lines:

$$\mathbb{E} \left[\binom{S_{t,1}}{k} \left| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right. \right] = \frac{1}{k!} \mathbb{E} \left[\frac{S_{t,1}!}{(S_{t,1}-k)!} \left| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right. \right] = \frac{1}{k!} \frac{N_{t+1}!}{(N_{t+1}-k)!} \left(\frac{1}{N_t} \right)^k \quad (12)$$

103 Consequently, the inclusive probability of coalescence for k lines is

$$p_{k,t} = \frac{1}{N_t^{k-1}} \quad (13)$$

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

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

106 It is interesting to note that neither the inclusive nor the exclusive coalescence probability depend on
107 the mean R_t of the Poisson offspring distribution or the size N_{t+1} of the population at time $t+1$. The
108 inclusive coalescent probability in Equation (13) can also be obtained conceptually by considering that
109 among the k lines, the first one has an ancestor with probability one, and the remaining $k-1$ need
110 to have the same ancestor among a set of N_t from which they choose uniformly at random so that
111 the probability of picking the same ancestor is $1/N_t$. The exclusive coalescent probability in Equation
112 (14) can be derived likewise by considering that in addition to the above, each of the $n-k$ other lines
113 need to choose a different ancestor, which happens with probability $(N_t - 1)/N_t$.

114 Figure 1 illustrates the inclusive and exclusive coalescence probabilities for the Poisson case for a set
115 of size $k = 1$ to $k = 10$ amongst a total of $n = 10$ observed lines, in a population of size $N_t = 10$,
116 $N_t = 20$ or $N_t = 30$.

117 4 Negative-Binomial case

118 Here the offspring distribution is $\alpha_t = \text{Negative-Binomial}(r, p)$ with parameters (r, p) set by moment-
119 matching mean R_t and variance V_t . The resulting parameters for this distribution are $r = R_t^2/(V_t - R_t)$

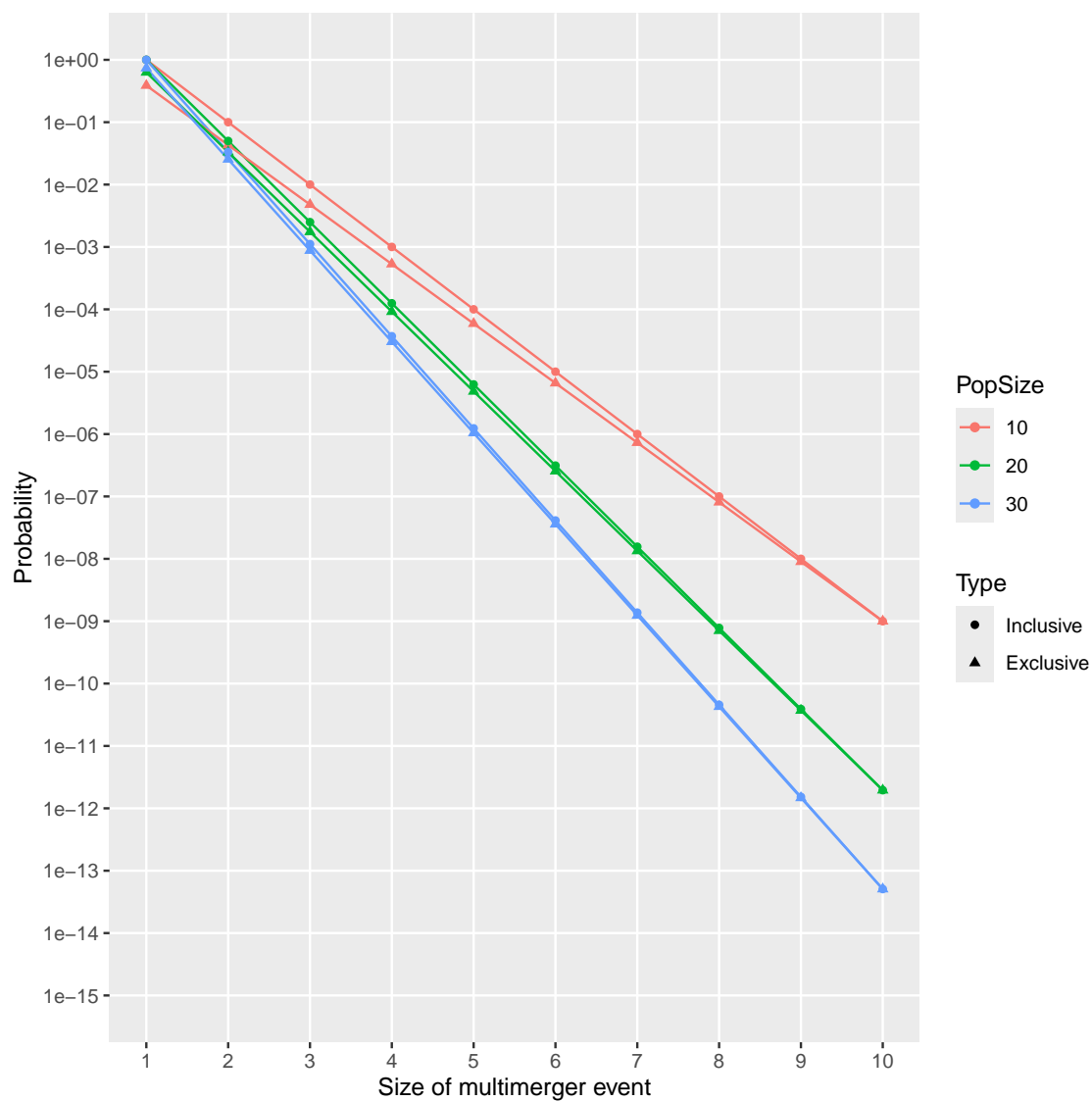


Figure 1: Inclusive and exclusive coalescence probabilities for the Poisson case.

120 and $p = R_t/V_t$. In this case, we have

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

121 and similarly to the Poisson(λ) offspring distribution identify the conditional distribution of
 122 $S_{t,1} | \sum_{i=1}^{N_t} S_{t,i}$ as follows,

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

123

124 where $B(x, y)$ denotes the Beta function defined as $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$. This is the probability
 125 mass function of Beta-Binomial and therefore we deduce that:

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

126 The k -th falling factorial moments of $X \sim \text{Beta-Binomial}(\alpha, \beta)$ are (Tripathi et al. 1994):

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

127 Injecting this formula into Equation (3), we deduce that the inclusive probability of coalescence for k
 128 lines is:

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

129 By injecting the probability mass function of a beta-binomial distribution in Equation (5) we deduce
 130 that the exclusive probability of coalescence for k lines is:

$$p_{n,k,t} = \frac{N_t B(k+r, n-k+N_t r-r)}{B(r, N_t r-r)} \quad (20)$$

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 .

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 of size $N_t = 12$. Several Negative-Binomial offspring distributions are compared, all of which have the same mean $R_t = 2$, and with the dispersion parameter equal to $r = 1$, $r = 2$, $r = 10$ and $r = 100$. When $r = 1$ the Negative-Binomial reduces to a Geometric distribution. When r is high (for example $r = 100$ as shown in Figure 2) the dispersion is low and the Negative-Binomial case behaves almost like the Poisson case. When r is lower the dispersion of the offspring distribution increases, so that both the inclusive and exclusive probabilities of larger multimerger events increase.

5 Limit when the population size is large

If we consider that the population size N_t is fixed and large, we can show the connections between our model and several previous studies.

Show that inclusive probabilities $p_{k,t}$ for $k > 2$ are small compared to $p_{2,t}$.

Show that exclusive probabilities $p_{n,k,t}$ for $k > 2$ are small compared to $p_{n,2,t}$, when $n \ll N_t$.

Show that inclusive and exclusive probabilities become equal, when $n \ll N_t$ in exclusive probabilities.

For Poisson offspring distribution we have:

$$p_{2,t} = p_{n,2,t} = \frac{1}{N_t} \quad (21)$$

For Negative-Binomial offspring distribution we have:

$$p_{2,t} = p_{n,2,t} = \frac{r+1}{N_t r+1} \approx \frac{r+1}{N_t r} \quad (22)$$

Fraser and Li (2017) calculated the effective population size $N_e(t)$ as a function of the actual population size $N(t)$ and the mean and variance of the offspring distribution R and σ^2 :

$$N_e(t) = \frac{N(t)}{\sigma^2/R + R - 1} \quad (23)$$

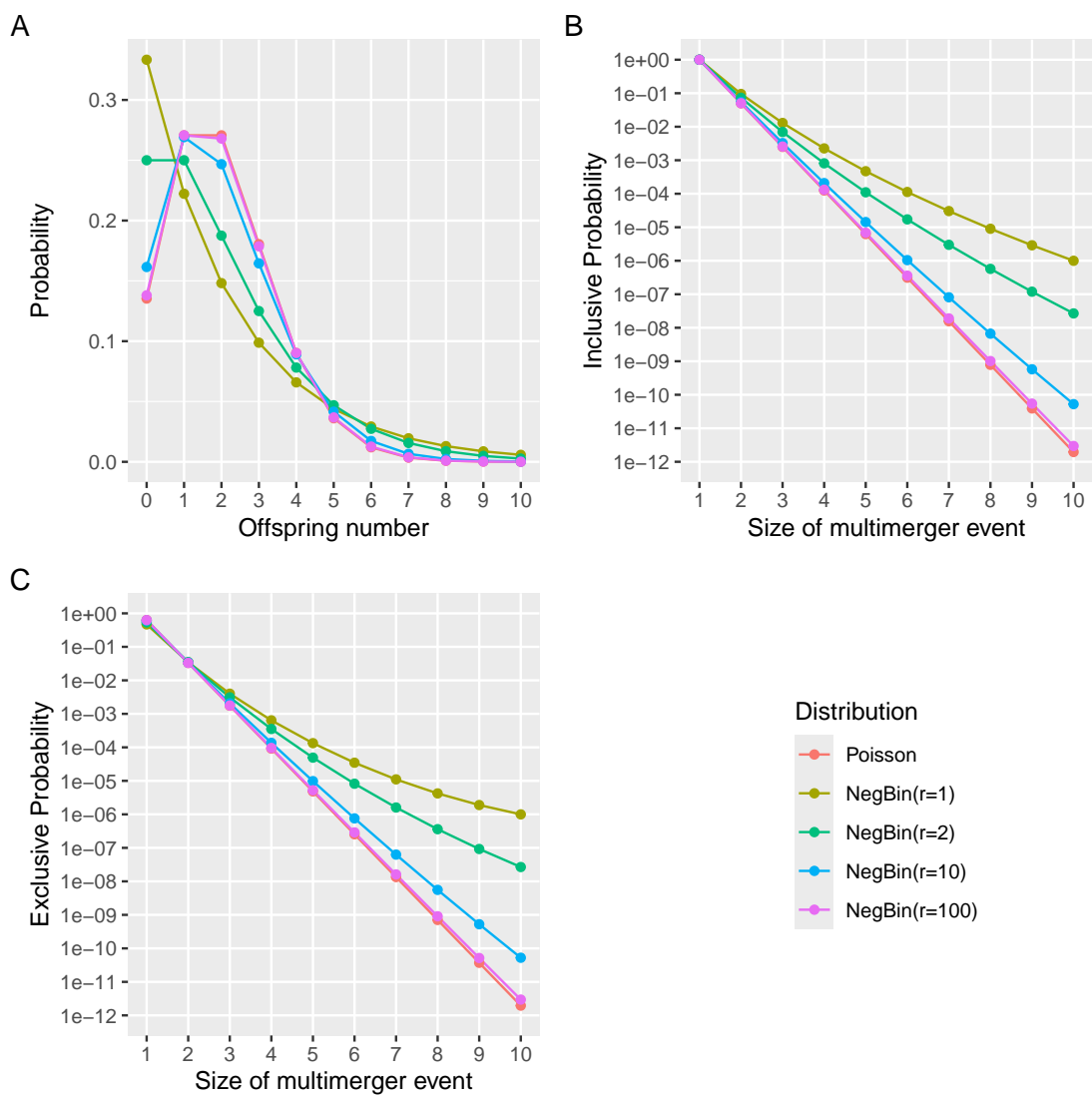


Figure 2: (A) Offspring distribution. (B) Inclusive probability of coalescence. (C) Exclusive probability of coalescence.

152 This formula was used to estimate the dispersion parameter from genetic data (Li et al. 2017). In our
 153 notation, this is equivalent to:

$$p_{2,t} = \frac{V_t/R_t + R_t - 1}{N_t R_t} \quad (24)$$

154 In the Poisson case we have $V_t = R_t$ so that Equation (24) simplifies to $p_{2,t} = 1/N_t$ which agrees with
 155 Equation (21). In the Negative-Binomial case we have $V_t/R_t = 1/p = (r + R_t)/r$ so that Equation
 156 (24) simplifies to $(r + 1)/(rN_t)$ which agrees with our Equation (22). Conversely, if we substitute
 157 $r = R_t^2/(V_t - R_t)$ in Equation (22) we obtain the formula Equation (24).

158 Koelle and Rasmussen (2012) derived the rates of coalescence of two lineages for several epidemiological
 159 models, assuming a large population at equilibrium. For each model they use the equation $N_e = N/\sigma^2$
 160 to relate the effective population size N_e to the actual population size N and the variance σ^2 in the
 161 number of offspring. This relationship was first established by Kingman (1982a) to apply the coalescent
 162 model to Cannings exchangeable models (Cannings 1974). From Equation (22) we can take $R_t = 1$ to
 163 achieve equilibrium of the population size and $r = R_t^2/(V_t - R_t) = 1/(V_t - 1)$ to deduce the equivalent
 164 $p_{2,t} = V_t/N_t$.

165 Volz (2012) showed that the rate of coalescence for two lineages under a continuous-time epidemic
 166 coalescent model is $2f(t)/I(t)^2$ where $f(t)$ is the incidence and $I(t)$ the prevalence. Setting in this
 167 formula the prevalence as $I(t) = N_{t+1} = N_t R_t$ and the incidence as $f(t) = R_t N_{t+1} = R_t^2 N_t$ we get
 168 a coalescent rate of $2/N_t$. To apply the Equation (22) we need to set $r = 1$ so that the offspring
 169 distribution is Geometric, which yields the same result.

170 6 Lambda-coalescent

171 The coalescent model (Kingman 1982a,b) describes the ancestry of a sample from a large population
 172 evolving according to many forward-in-time models such as the Wright-Fisher model (Wright 1931;
 173 Fisher 1930), the Moran model (Moran 1958) and the Cannings exchangeable model (Cannings 1974).
 174 Since the coalescent considers a large population in which each individual only has a number of
 175 offspring that is small compared to the population size, coalescent trees are always binary and do not
 176 feature multimergers, making them unsuitable to represent the ancestry of outbreaks considered in
 177 this study. However, the lambda-coalescent models are an extension of the coalescent model that do
 178 allow multimergers (Pitman 1999; Sagitov 1999; Donnelly and Kurtz 1999).

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

$$\lambda_{n,k} = \int_0^1 x^{k-2} (1-x)^{n-k} \Lambda(dx) \quad (25)$$

182 The beta-coalescent (Schweinsberg 2003) is a specific type of lambda-coalescent. Was used in (Hoscheit
 183 and Pybus 2019) and (Menardo et al. 2021). David's paper on inference of multiple mergers while

184 dating a pathogen phylogeny (Helekal et al. 2024). The Beta($2 - \alpha, \alpha$)-coalescent model has a single
 185 parameter $\alpha \in [0, 2]$ and is defined as:

$$\Lambda(dx) = \frac{x^{1-\alpha}(1-x)^{\alpha-1}}{B(2-\alpha, \alpha)}dx \quad (26)$$

186 By combining Equations (25) and (26) we can deduce that:

$$\lambda_{n,k} = \frac{B(k-\alpha, n-k+\alpha)}{B(2-\alpha, \alpha)} \quad (27)$$

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

190 We now define our own lambda-coalescent based on the Negative-Binomial case described previously.
 191 For ease of comparison with other coalescent models, we consider that time is continuous and that
 192 the population size remains constant equal to N_t . The exclusive coalescent probability $p_{n,k,t}$ in the
 193 Negative-Binomial case given by Equation (20) can be used to determine the corresponding rate of
 194 our lambda-coalescent, if we consider that the probability of each event in discrete time is the result
 195 of the event happening at a constant rate in continuous time:

$$\lambda_{n,k} = -\log(1 - p_{n,k,t}) \quad (28)$$

196 In order to compare our lambda-coalescent with other models, we consider the distribution of the size
 197 k of the next event among a set of n lineages. For any lambda-coalescent this can be computed as:

$$p(k|n) = \frac{\binom{n}{k}\lambda_{n,k}}{\sum_{i=2}^n \binom{n}{i}\lambda_{n,i}} \quad (29)$$

198 Figure 3 compares this distribution for $n = 10$ in the Beta-coalescent with parameter $\alpha \in \{0.5, 1, 1.5\}$
 199 and for our lambda-coalescent with parameters $N_t \in \{15, 25, 50\}$ and $r \in \{0.1, 0.5, 1\}$.

200 Figure 4 shows examples of trees simulated for a sample of size $n = 20$ and with constant population
 201 size $N_t = 40$.

202 Figure 5 shows summary statistics for 10,000 trees simulated in the same conditions as the individual
 203 trees shown in Figure 4. As the dispersion parameter increases from $r = 0.1$ to $r = 10$ multimerger
 204 events become less and less likely and large. Simultaneously, the time to the most recent common
 205 ancestor increases, as well as the stemminess of the tree (ie the proportion of branch lengths in non-
 206 terminal branches).

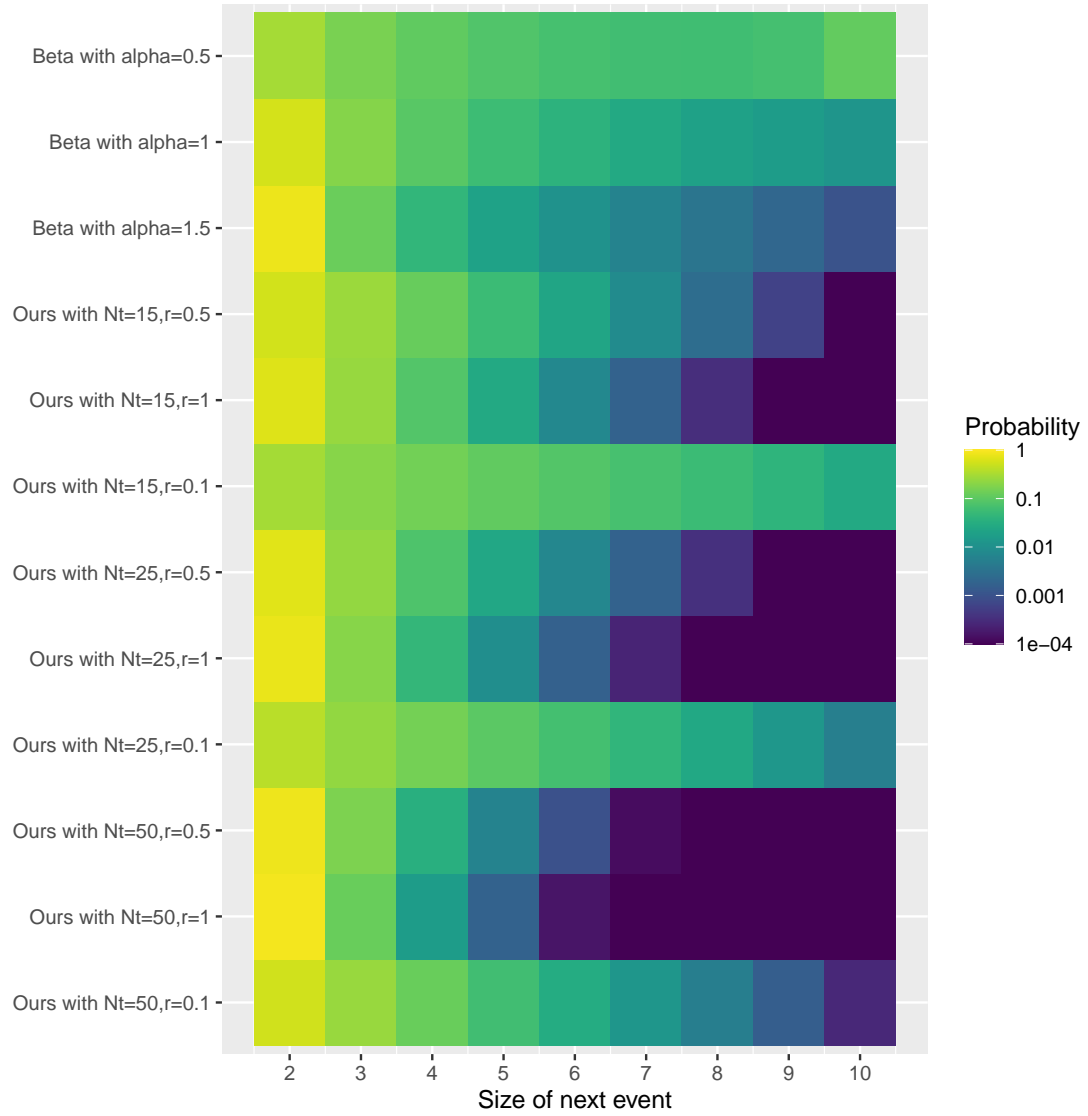


Figure 3: Distribution of the size of the next event among a set of $n = 10$ lineages, compared between the Beta-coalescent and our lambda-coalescent model with various parameters.

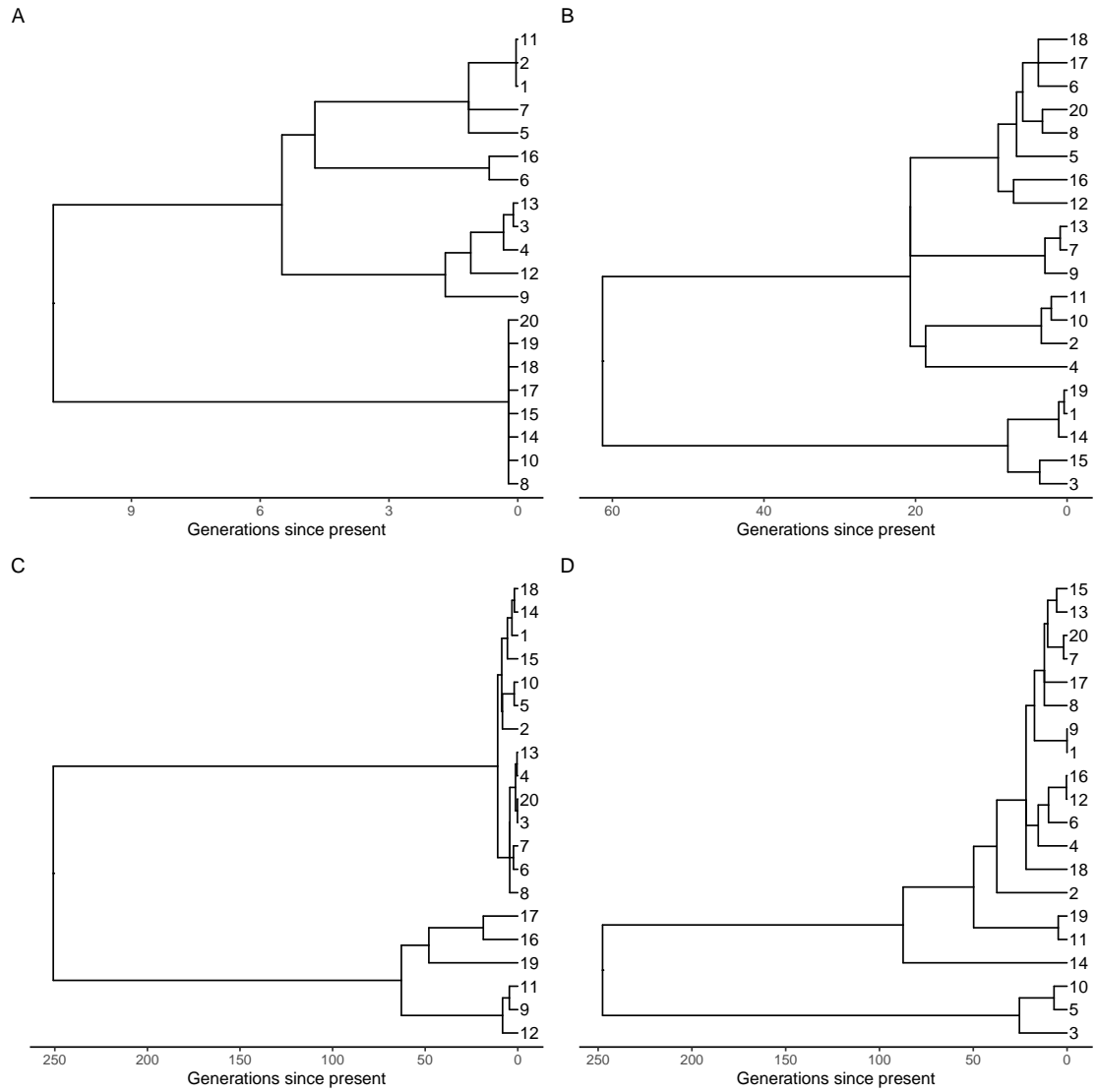


Figure 4: Example of trees simulated under our lambda-coalescent with $r = 0.1$ (A), $r = 1$ (B), $r = 5$ (C) and $r = 10$ (D).

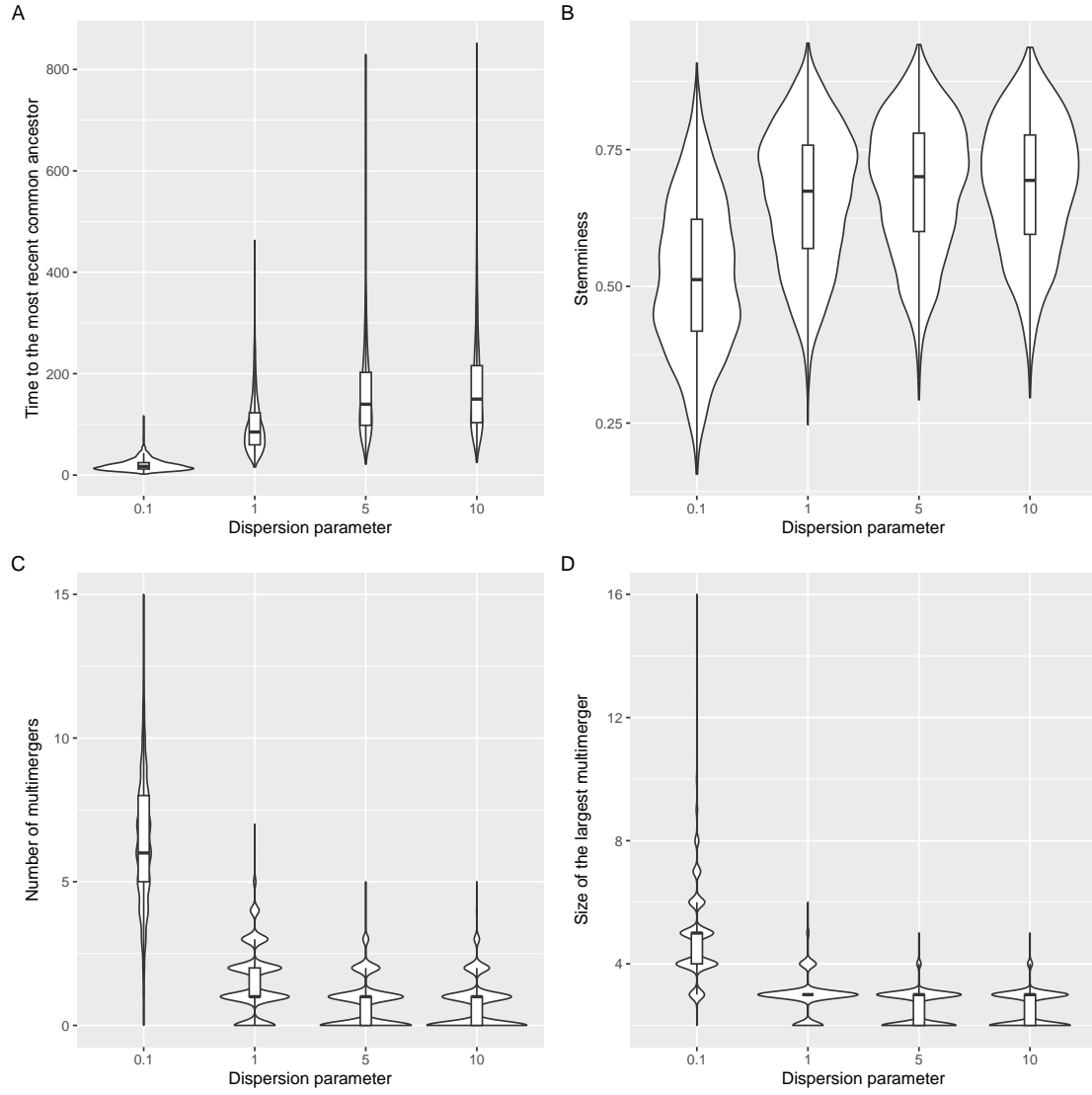


Figure 5: Summary statistics for trees simulated under our lambda-coalescent with $r = 0.1, r = 1, r = 5$ and $r = 10$, namely the time to the most recent common ancestor (A), stemminess (B), number of multimerers (C) and the size of the largest multimerger (D).

207 7 Implementation

208 We implemented the analytical methods described in this paper in a new R package entitled *EpiLambda*
209 which is available at <https://github.com/xavierdidelot/EpiLambda> for R version 3.5 or later. All
210 code and data needed to replicate the results are included in the “run” directory of the *EpiLambda*
211 repository.

212 8 Discussion

213 Our lambda-coalescent could be defined in a varying population size following the same approach as
214 previously described for the coalescent (Griffiths and Tavaré 1994) and the beta-coalescent (Hoscheit
215 and Pybus 2019). Could also extend to temporally offset leaves following work on the coalescent
216 (Drummond et al. 2003) and the beta-coalescent (Hoscheit and Pybus 2019).

217 The Xi-coalescent models admit multiple simultaneous mergers (Schweinsberg 2000).

218 Difference between transmission tree and phylogenetic tree (Jombart et al. 2011). Modelling within-
219 host evolution to bridge the gap (Didelot et al. 2014; Hall et al. 2015; Didelot et al. 2017).
220 Superspreading individuals vs superspreading events (Riley et al. 2003; Wallinga and Teunis 2004;
221 Ho et al. 2023).

222 Acknowledgements

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

References

- Anderson, R.M., May, R.M., 1991. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, USA.
- Cannings, C., 1974. The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Adv. Appl. Probab.* 6, 260–290. doi:10.2307/1426293.
- Didelot, X., Fraser, C., Gardy, J., Colijn, C., 2017. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution* 34, 997–1007. doi:10.1093/molbev/msw275.
- Didelot, X., Gardy, J., Colijn, C., 2014. Bayesian inference of infectious disease transmission from whole genome sequence data. *Molecular Biology and Evolution* 31, 1869–1879. doi:10.1093/molbev/msu121.
- Donnelly, P., Kurtz, T.G., 1999. Particle Representations for Measure-Valued Population Models. *The Annals of Probability* 27. doi:10.1214/aop/1022677258.
- Drummond, A.J., Pybus, O.G., Rambaut, A., Forsberg, R., Rodrigo, A.G., 2003. Measurably evolving populations. *Trends in Ecology and Evolution* 18, 481–488. doi:10.1016/S0169-5347(03)00216-7.
- Du, Z., Wang, C., Liu, C., Bai, Y., Pei, S., Adam, D.C., Wang, L., Wu, P., Lau, E.H.Y., Cowling, B.J., 2022. Systematic review and meta-analyses of superspreading of SARS-CoV-2 infections. *Transboundary and Emerging Diseases* 69. doi:10.1111/tbed.14655.
- Ferguson, N.M., Cummings, D.A.T., Fraser, C., Cajka, J.C., Cooley, P.C., Burke, D.S., 2006. Strategies for mitigating an influenza pandemic. *Nature* 442, 448–452. doi:10.1038/nature04795.
- Fisher, R.A., 1930. *The genetical theory of natural selection*. Clarendon Press. doi:10.5962/bh1.title.27468.
- Fraser, C., Li, L.M., 2017. Coalescent models for populations with time-varying population sizes and arbitrary offspring distributions. *bioRxiv* , 10.1101/131730doi:10.1101/131730.
- Fraser, C., Riley, S., Anderson, R.M., Ferguson, N.M., 2004. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences* 101, 6146–6151. doi:10.1073/pnas.0307506101.
- Gómez-Carballa, A., Pardo-Seco, J., Bello, X., Martínón-Torres, F., Salas, A., 2021. Superspreading in the emergence of COVID-19 variants. *Trends in Genetics* 37, 1069–1080. doi:10.1016/j.tig.2021.09.003.
- Grassly, N.C., Fraser, C., 2008. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology* 6, 477–87. doi:10.1038/nrmicro1845.
- Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society B* 344, 403–410.
- Hall, M., Woolhouse, M., Rambaut, A., 2015. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Computational Biology* 11, e1004613. doi:10.1371/journal.pcbi.1004613.
- Helekal, D., Koskela, J., Didelot, X., 2024. Inference of multiple mergers while dating a pathogen phylogeny. *bioRxiv* , 2023.09.12.557403doi:10.1101/2023.09.12.557403.

264 Ho, F., Parag, K.V., Adam, D.C., Lau, E.H.Y., Cowling, B.J., Tsang, T.K., 2023. Accounting for the
 265 Potential of Overdispersion in Estimation of the Time-varying Reproduction Number. *Epidemiology*
 266 34, 201–205. doi:10.1097/EDE.0000000000001563.

267 Hoscheit, P., Pybus, O.G., 2019. The multifurcating skyline plot. *Virus Evolution* 5, 1–10.
 268 doi:10.1093/ve/vez031.

269 Jombart, T., Eggo, R.M., Dodd, P.J., Balloux, F., 2011. Reconstructing disease outbreaks from genetic
 270 data: A graph approach. *Heredity* 106, 383–90. doi:10.1038/hdy.2010.78.

271 Keeling, M.J., Rohani, P., 2008. Modeling infectious diseases in humans and animals. Princeton
 272 university press.

273 Kingman, J., 1982a. The coalescent. *Stochastic Processes and their Applications* 13, 235–248.
 274 doi:10.1016/0304-4149(82)90011-4.

275 Kingman, J.F.C., 1982b. On the genealogy of large populations. *Journal of Applied Probability* 19,
 276 27–43. doi:10.2307/3213548.

277 Koelle, K., Rasmussen, D.A., 2012. Rates of coalescence for common epidemiological models at
 278 equilibrium. *Journal of The Royal Society Interface* 9, 997–1007. doi:10.1098/rsif.2011.0495.

279 Kucharski, A.J., Althaus, C.L., 2015. The role of superspreading in Middle East respiratory syndrome
 280 coronavirus (MERS-CoV) transmission. *Eurosurveillance* 20. doi:10.2807/1560-7917.ES2015.20.
 281 25.21167.

282 Lemieux, J.E., Siddle, K.J., Shaw, B.M., Loreth, C., Schaffner, S.F., Gladden-Young, A., Adams,
 283 G., Fink, T., Tomkins-Tinch, C.H., Krasilnikova, L.A., DeRuff, K.C., Rudy, M., Bauer, M.R.,
 284 Lagerborg, K.A., Normandin, E., Chapman, S.B., Reilly, S.K., Anahtar, M.N., Lin, A.E., Carter,
 285 A., Myhrvold, C., Kembell, M.E., Chaluvadi, S., Cusick, C., Flowers, K., Neumann, A., Cerrato,
 286 F., Farhat, M., Slater, D., Harris, J.B., Branda, J.A., Hooper, D., Gaeta, J.M., Baggett, T.P.,
 287 O’Connell, J., Gnirke, A., Lieberman, T.D., Philippakis, A., Burns, M., Brown, C.M., Luban, J.,
 288 Ryan, E.T., Turbett, S.E., LaRocque, R.C., Hanage, W.P., Gallagher, G.R., Madoff, L.C., Smole, S.,
 289 Pierce, V.M., Rosenberg, E., Sabeti, P.C., Park, D.J., MacInnis, B.L., 2021. Phylogenetic analysis
 290 of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* 371, eabe3261.
 291 doi:10.1126/science.abe3261.

292 Li, L.M., Grassly, N.C., Fraser, C., 2017. Quantifying Transmission Heterogeneity Using Both
 293 Pathogen Phylogenies and Incidence Time Series. *Molecular Biology and Evolution* 34, 2982–2995.
 294 doi:10.1093/molbev/msx195.

295 Lloyd-Smith, J., Schreiber, S., Kopp, P., Getz, W., 2005. Superspreading and the effect of individual
 296 variation on disease emergence. *Nature* 438, 355–9. doi:10.1038/nature04153.

297 Menardo, F., Gagneux, S., Freund, F., 2021. Multiple Merger Genealogies in Outbreaks of
 298 Mycobacterium tuberculosis. *Molecular Biology and Evolution* 38, 290–306. doi:10.1093/molbev/
 299 msaa179.

300 Moran, P., 1958. Random Processes in Genetics. *Mathematical Proceedings of the Cambridge*
 301 *Philosophical Society* 54, 60–71.

302 Pitman, J., 1999. Coalescents with multiple collisions. *The Annals of Probability* 27, 1870–1902.

303 Potts, R.B., 1953. Note on the Factorial Moments of Standard Distributions. *Australian Journal of*
 304 *Physics* 6, 498–499. URL: <https://www.publish.csiro.au/ph/ph530498>, doi:10.1071/ph530498.
 305 publisher: CSIRO PUBLISHING.

306 Riley, S., Fraser, C., a Donnelly, C., Ghani, A.C., Abu-Raddad, L.J., Hedley, A.J., Leung, G.M.,
307 Ho, L.M., Lam, T.H., Thach, T.Q., Chau, P., Chan, K.P., Lo, S.V., Leung, P.Y., Tsang, T., Ho,
308 W., Lee, K.H., Lau, E.M.C., Ferguson, N.M., Anderson, R.M., 2003. Transmission dynamics of the
309 etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* 300, 1961–6.
310 doi:10.1126/science.1086478.

311 Sagitov, S., 1999. The general coalescent with asynchronous mergers of ancestral lines. *Journal of*
312 *Applied Probability* 36, 1116–1125. doi:10.1239/jap/1032374759.

313 Schweinsberg, J., 2000. Coalescents with Simultaneous Multiple Collisions. *Electronic Journal of*
314 *Probability* 5. doi:10.1214/EJP.v5-68.

315 Schweinsberg, J., 2003. Coalescent processes obtained from supercritical Galton–Watson processes.
316 *Stochastic Processes and their Applications* 106, 107–139. doi:10.1016/S0304-4149(03)00028-0.

317 Stein, R.A., 2011. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*
318 15, e510–e513. doi:10.1016/j.ijid.2010.06.020.

319 Tripathi, R.C., Gupta, R.C., Gurland, J., 1994. Estimation of parameters in the beta binomial model.
320 *Annals of the Institute of Statistical Mathematics* 46, 317–331. URL: [https://doi.org/10.1007/](https://doi.org/10.1007/BF01720588)
321 [BF01720588](https://doi.org/10.1007/BF01720588), doi:10.1007/BF01720588.

322 Volz, E.M., 2012. Complex population dynamics and the coalescent under neutrality. *Genetics* 190,
323 187–201. doi:10.1534/genetics.111.134627.

324 Wallinga, J., Teunis, P., 2004. Different Epidemic Curves for Severe Acute Respiratory Syndrome
325 Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology* 160, 509–516.

326 Wang, J., Chen, X., Guo, Z., Zhao, S., Huang, Z., Zhuang, Z., Wong, E.L.y., Zee, B.C.Y., Chong,
327 M.K.C., Wang, M.H., Yeoh, E.K., 2021. Superspreading and heterogeneity in transmission of SARS,
328 MERS, and COVID-19: A systematic review. *Computational and Structural Biotechnology Journal*
329 19, 5039–5046. doi:10.1016/j.csbj.2021.08.045.

330 Wang, L., Didelot, X., Yang, J., Wong, G., Shi, Y., Liu, W., Gao, G.F., Bi, Y., 2020. Inference of
331 person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early
332 outbreak phase. *Nature Communications* 11, 5006. doi:10.1038/s41467-020-18836-4.

333 Woolhouse, M.E.J., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., Hagan, P., Hii,
334 J.L.K., Ndhlovu, P.D., Quinnell, R.J., Watts, C.H., Chandiwana, S.K., Anderson, R.M., 1997.
335 Heterogeneities in the transmission of infectious agents: Implications for the design of control
336 programs. *Proceedings of the National Academy of Sciences* 94, 338–342. doi:10.1073/pnas.94.1.
337 338.

338 Wright, S., 1931. Evolution in Mendelian populations. *Genetics* 16, 97–159. doi:10.1093/genetics/
339 16.2.97.