

1 Ancestral process for infectious disease outbreaks with superspreading

2 Xavier Didelot<sup>1,2,\*</sup>, David Helekal<sup>3</sup>, Ian Roberts<sup>2</sup>

3 <sup>1</sup> School of Life Sciences, University of Warwick, Coventry, United Kingdom

4 <sup>2</sup> Department of Statistics, University of Warwick, Coventry, United Kingdom

5 <sup>3</sup> Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health,  
6 Boston, Massachusetts, USA

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

8 Running title: Ancestry for outbreaks with superspreading

9 Keywords: infectious disease epidemiology modelling; offspring distribution; superspreading;  
10 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 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 (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 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 (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 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  $r$  (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

55 according to the offspring distribution. Here we consider the reverse problem of the transmission  
 56 ancestry, going backward-in-time, from a sample of infected individuals, until reaching the last common  
 57 transmission ancestor of the whole sample. Given a set of  $n$  sampled individuals, we show how to  
 58 calculate the probability that a given subset of size  $k$  have the same infector, either inclusively (so that  
 59 the remaining  $n - k$  may also have the same infector or not) or exclusively (so that none of the remaining  
 60  $n - k$  have the same infector). We start by considering the general case of an offspring distribution  
 61 with arbitrary form, and then the specific cases of offspring distributions that follow a Poisson and  
 62 a Negative-Binomial distribution. The main novelty of our approach is that we consider that the  
 63 overall population size is small, but we show that in the limit where the population size is large, our  
 64 results agree with several previous studies (Volz 2012; Koelle and Rasmussen 2012; Fraser and Li 2017).  
 65 Finally, we show how our results can be incorporated into a new lambda-coalescent model (Pitman  
 66 1999; Sagitov 1999; Donnelly and Kurtz 1999) and compare it with previously proposed models.

## 67 **2 General offspring distribution case**

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

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

### 74 **2.1 Inclusive coalescence probability**

75 We define the inclusive coalescence probability  $p_{k,t}(N_t, N_{t+1})$  as the probability that a specific set of  
 76  $k$  individuals from generation  $t + 1$  have the same infector in generation  $t$ , conditional on population  
 77 sizes  $N_t$  and  $N_{t+1}$ . Given full information about offspring counts from individuals in generation  $t$ ,  
 78  $\mathbf{s}_t = (s_{t,1}, \dots, s_{t,N_t})$ , we have:

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

79

80 Full information  $\{s_{t,i}\}$  yields the population size  $N_{t+1}$  as shown in Equation 1, but this is not available  
81 in practice. We can instead express the inclusive coalescence probability conditioning on the next  
82 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  
83 total generation size. Let  $S_t^{-(1)} = (S_{t,2}, \dots, S_{t,N_t})$ :

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

84

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

## 87 2.2 Exclusive coalescence probability

88 Generally, we observe a sample of individuals from each generation rather than the entire population.  
 89 In this case, we are interested in the exclusive coalescence probability  $p_{n,k,t}(N_t, N_{t+1})$  that a specific  
 90 subset of  $k$  individuals amongst  $n$  sampled individuals arose from a common infector one generation  
 91 in the past given knowledge of the total population sizes  $N_t$  and  $N_{t+1}$ . Let us first assume full  
 92 knowledge about offspring counts of the individuals at time  $N_t$  amongst the sample at time  $N_{t+1}$ ,  
 93 namely  $\mathbf{x}_t = (x_{t,1}, \dots, x_{t,N_t})$  such that  $x_{t,1} + \dots + x_{t,N_t} = n$ . Note that  $X_{t,i}$  does not follow the same  
 94 offspring distribution as  $S_{t,i}$ . We have:

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

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

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

$$= \frac{N_t}{\binom{n}{k}} \mathbb{P} \left[ X_{t,1} = k \mid \sum_{i=1}^{N_t} X_{t,i} = n \right] \quad (5)$$

97

## 98 2.3 Complementarity of exclusive coalescence probabilities

99 If we consider one of the lines observed amongst a set of  $n$ , it can either remain uncoalesced with  
 100 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  
 101 set of  $k - 1$  lines among the  $n - 1$  other lines, leading to the following complementarity equation:

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

102 We can show that it is indeed satisfied by the formula in Equation 5:

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

### 103 3 Poisson offspring distribution case

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

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

105 and the conditional distribution:

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

106

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

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

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

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

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



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

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

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

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

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

114 It is interesting to note that neither the inclusive nor the exclusive coalescence probability depend on  
 115 the mean  $R_t$  of the Poisson offspring distribution or the size  $N_{t+1}$  of the population at time  $t + 1$ . Both  
 116 only depend on the population size  $N_t$  at time  $t$ . The inclusive coalescent probability in Equation 12  
 117 can also be obtained conceptually by considering that among the  $k$  lines, the first one has an ancestor  
 118 with probability one, and the remaining  $k - 1$  need to have the same ancestor among a set of  $N_t$  from  
 119 which they choose uniformly at random so that the probability of picking the same ancestor is  $1/N_t$ .  
 120 The exclusive coalescent probability in Equation 14 can be derived likewise by considering that in  
 121 addition to the above, each of the  $n - k$  other lines need to choose a different ancestor, which happens  
 122 with probability  $(N_t - 1)/N_t$ . Figure 1 illustrates the inclusive and exclusive coalescence probabilities  
 123 for a set of size  $k = 1$  to  $k = 10$  amongst a total of  $n = 10$  observed individuals, in a population of size  
 124  $N_t = 10$ ,  $N_t = 20$  or  $N_t = 30$ .

## 125 4 Negative-Binomial offspring distribution case

126 In this section we consider that the offspring distribution is  $\alpha_t = \text{Negative-Binomial}(r, p)$  with  
 127 parameters  $(r, p)$  set by moment-matching the mean  $R_t$  and variance  $V_t$  of the offspring distribution  
 128 which are assumed constant over time. The resulting parameters for this distribution are  $r =$

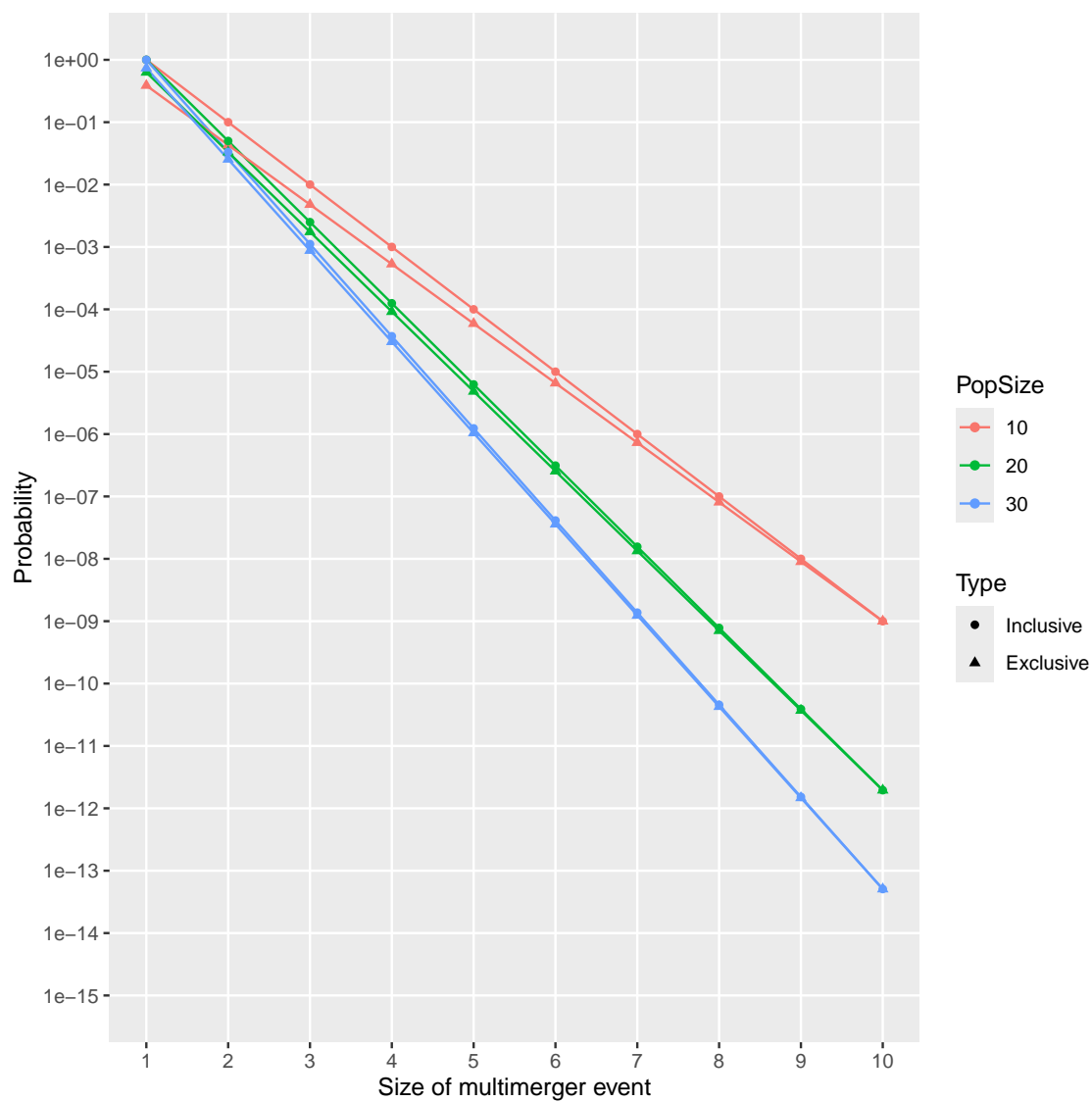


Figure 1: Inclusive and exclusive coalescence probabilities for the Poisson case.

129  $R_t^2/(V_t - R_t)$  and  $p = R_t/V_t$ . In this case, we have:

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

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

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

132

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

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

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

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

136 By applying this formula to the Beta-Binomial distribution in Equation 17 and injecting into Equation  
 137 3, we deduce that the inclusive probability of coalescence for  $k$  lines is:

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

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

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

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

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

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

$$\lim_{a \rightarrow \infty} \frac{\Gamma(a + b)}{\Gamma(a)} = a^b e^{-b} \quad (22)$$

147 Figure 2 illustrates the inclusive and exclusive coalescence probabilities for the Negative-Binomial case  
148 for a set of size  $k = 1$  to  $k = 10$  amongst a total of  $n = 10$  observed lines, in a population with  
149 size  $N_t = 20$ . Several Negative-Binomial offspring distributions are compared, all of which have the  
150 same mean  $R_t = 2$ , and with the dispersion parameter equal to  $r = 0.1$ ,  $r = 1$ ,  $r = 10$  and  $r = 100$   
151 (Figure 2A). When  $r = 1$  the Negative-Binomial reduces to a Geometric distribution. When  $r$  is high  
152 the dispersion is low and the Negative-Binomial case behaves almost like the Poisson case for both  
153 the inclusive (Figure 2B) and the exclusive coalescence probabilities (Figure 2C). When  $r$  is lower the  
154 dispersion of the offspring distribution increases, so that both the inclusive and exclusive probabilities  
155 of larger multimerger events are increased compared to the Poisson case. In particular, when  $r = 0.1$   
156 we see that the exclusive probability can increase with the size of the event considered (Figure 2C).

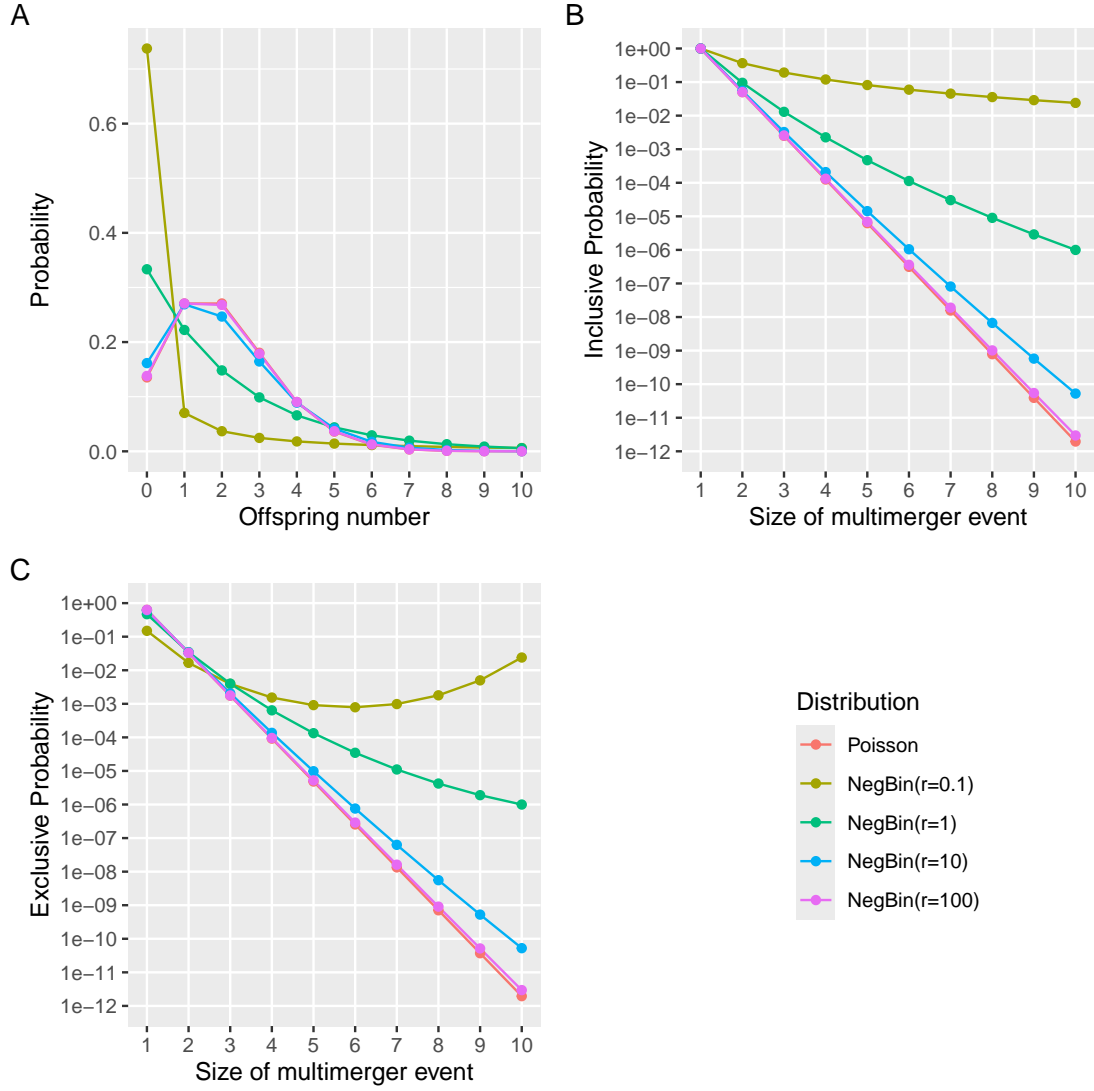


Figure 2: (A) Offspring distributions with mean  $R_t = 2$ . (B) Inclusive probability of coalescence for  $N_t = 20$  and  $n = 10$ . (C) Exclusive probability of coalescence for  $N_t = 20$  and  $n = 10$ .

157 This happens because the probability is not much lower for the common ancestor having say 10 rather  
 158 than 9 offspring, while on the other hand if the event is of size 9 only then another individual in the  
 159 generation of the ancestor needs to have had at least one sampled offspring.

## 160 5 Limit when the population size is large

161 If we consider that the population size  $N_t$  is fixed and large, we can show the connections between  
 162 our results and several previous studies on the ancestral process of infectious diseases. In the Poisson  
 163 case, from Equations 12 and 14 we can see that both inclusive and exclusive probabilities are of order  
 164  $\mathcal{O}(N_t^{1-k})$ . We can therefore ignore events with  $k > 2$  and retain only the events with  $k = 2$  which  
 165 occur with the same inclusive and exclusive probabilities:

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

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

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

170 Koelle and Rasmussen (2012) derived the rates of coalescence of two lineages for several epidemiological  
 171 models, assuming a large population at equilibrium. For each model they use the equation  $N_e = N/\sigma^2$   
 172 to relate the effective population size  $N_e$  to the actual population size  $N$  and the variance  $\sigma^2$  in the  
 173 number of offspring. This relationship was first established by Kingman (1982a) to derive the backward-  
 174 in-time coalescent model from the forward-in-time Cannings exchangeable models (Cannings 1974).  
 175 This result implies that the rate of coalescence for two lineages is  $1/N_e = \sigma^2/N$ . From Equation 24 we  
 176 can take  $R_t = 1$  to achieve equilibrium of the population size and the method of moments estimator  
 177  $r = R_t^2/(V_t - R_t) = 1/(V_t - 1)$  to deduce the equivalent result  $p_{2,t}(N_t, N_{t+1}) = V_t/N_t$ .

178 Volz (2012) showed that the rate of coalescence for two lineages under a continuous-time epidemic  
 179 coalescent model is  $2f(t)/I(t)^2$  where  $f(t)$  is the incidence of the disease and  $I(t)$  its prevalence. Setting

180 in this formula the prevalence as  $I(t) = N_{t+1} = N_t R_t$  and the incidence as  $f(t) = R_t I(t) = R_t^2 N_t$   
181 we get a coalescent rate of  $2/N_t$ . To apply our methodology we need to consider that the offspring  
182 distribution is Geometric, since the epidemiological models considered have successes (transmission)  
183 happening until the first failure (removal). We therefore set  $r = 1$  in Equation 24 to make the Negative-  
184 Binomial offspring distribution reduce to a Geometric distribution and the same result follows.

185 Fraser and Li (2017) calculated the effective population size  $N_e(t)$  as a function of the actual population  
186 size  $N(t)$  and the mean and variance of the offspring distribution  $R$  and  $\sigma^2$ . This formula was used to  
187 estimate the dispersion parameter of a Negative-Binomial offspring distribution from genetic data (Li  
188 et al. 2017). Using our notations, their formula is equivalent to the inclusive coalescence probability  
189 for two lineages:

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

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

## 194 6 Definition of a new lambda-coalescent model

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

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

$$\lambda_{n,k} = \int_0^1 x^{k-2} (1-x)^{n-k} \Lambda(dx) \quad (26)$$

205 The beta-coalescent (Schweinsberg 2003) is a specific type of lambda-coalescent that has been used  
 206 recently in several studies analysing genetic data from infectious disease agents (Hoscheit and Pybus  
 207 2019; Menardo et al. 2021; Helekal et al. 2025; Zhang and Palacios 2024). The beta-coalescent model  
 208 has a single parameter  $\alpha \in [0, 2]$  and is defined as:

$$\Lambda(dx) = \frac{x^{1-\alpha} (1-x)^{\alpha-1}}{B(2-\alpha, \alpha)} dx \quad (27)$$

209 By combining Equations 26 and 27 we deduce that:

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

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

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

$$\lambda_{n,k} = -\log(1 - p_{n,k,t}(N_t, N_{t+1})) \quad (29)$$

220 Note that this equation implies that time is measured continuously in number of transmission  
 221 generations. For example to measure time in decimal days instead, the time scale would need to  
 222 be multiplied by the mean of the generation time distribution measured in days (Svensson 2007). The  
 223 omega-coalescent has two parameters: the constant population size  $N_t$  and the dispersion parameter



224  $r$ . In order to compare the omega-coalescent defined in Equation 29 with other models such as the  
 225 beta-coalescent defined in Equation 28, we consider the distribution of the size  $k$  of the next event  
 226 among a set of  $n$  lineages. For any lambda-coalescent this can be computed as:

$$p(k|n) = \frac{\binom{n}{k} \lambda_{n,k}}{\sum_{i=2}^n \binom{n}{i} \lambda_{n,i}} \quad (30)$$

227 Figure 3 compares this distribution for  $n = 10$  in the beta-coalescent with parameter  $\alpha \in \{0.5, 1, 1.5\}$   
 228 and for the omega-coalescent with parameters  $N_t \in \{10, 20, 30\}$  and  $r \in \{0.1, 1, 10\}$ . In the beta-  
 229 coalescent, the distribution shifts towards more larger multimerger events as the parameter  $\alpha$  decreases.  
 230 In the omega-coalescent a wider range of behaviours is obtained when varying the two parameters  $N_t$   
 231 and  $r$ . For a given value of  $N_t$ , decreasing the value of  $r$  results in more larger events. Conversely, for  
 232 a given value of  $r$  we can see that increasing the value of  $N_t$  reduces the probability of larger events.

233 Genealogies can be simulation from the omega-coalescent model defined in Equation 29 using the same  
 234 algorithm as for other lambda-coalescent models (Pitman 1999). Given  $n$  lineages, the next coalescent  
 235 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  
 236 this event is drawn according to Equation 30, and the  $k$  lineages that coalesce are chosen uniformly  
 237 amongst the  $n$  lineages. This process is repeated iteratively until all lineages have coalesced. Figure 4  
 238 shows examples of trees simulated for a sample of size  $n = 20$ , constant population size  $N_t = 30$  and  
 239 dispersion parameter  $r \in \{0.1, 1, 10, 100\}$ . It is already clear from these single realisations that the  
 240 lower values of  $r$  result in trees with more larger multimerger events and lower time to the most recent  
 241 common ancestor, but to quantify these properties we need to consider many trees. Figure 5 shows  
 242 summary statistics for 10,000 trees simulated in the same conditions as the individual trees shown in  
 243 Figure 4. As the dispersion parameter increases from  $r = 0.1$  to  $r = 100$  multimerger events become  
 244 less and less likely and less large (Figure 5A and B), and the time to the most recent common ancestor  
 245 increases (Figure 5C). Furthermore, the stemminess of the tree increases, which is defined as the sum  
 246 of lengths of internal branches divided by the total sum of branch lengths (Figure 5D). Stemminess is  
 247 usually taken as a sign of population size dynamics (Fiala and Sokal 1985; Didelot et al. 2009), which  
 248 would be misleading here since all simulations assumed a constant population size.

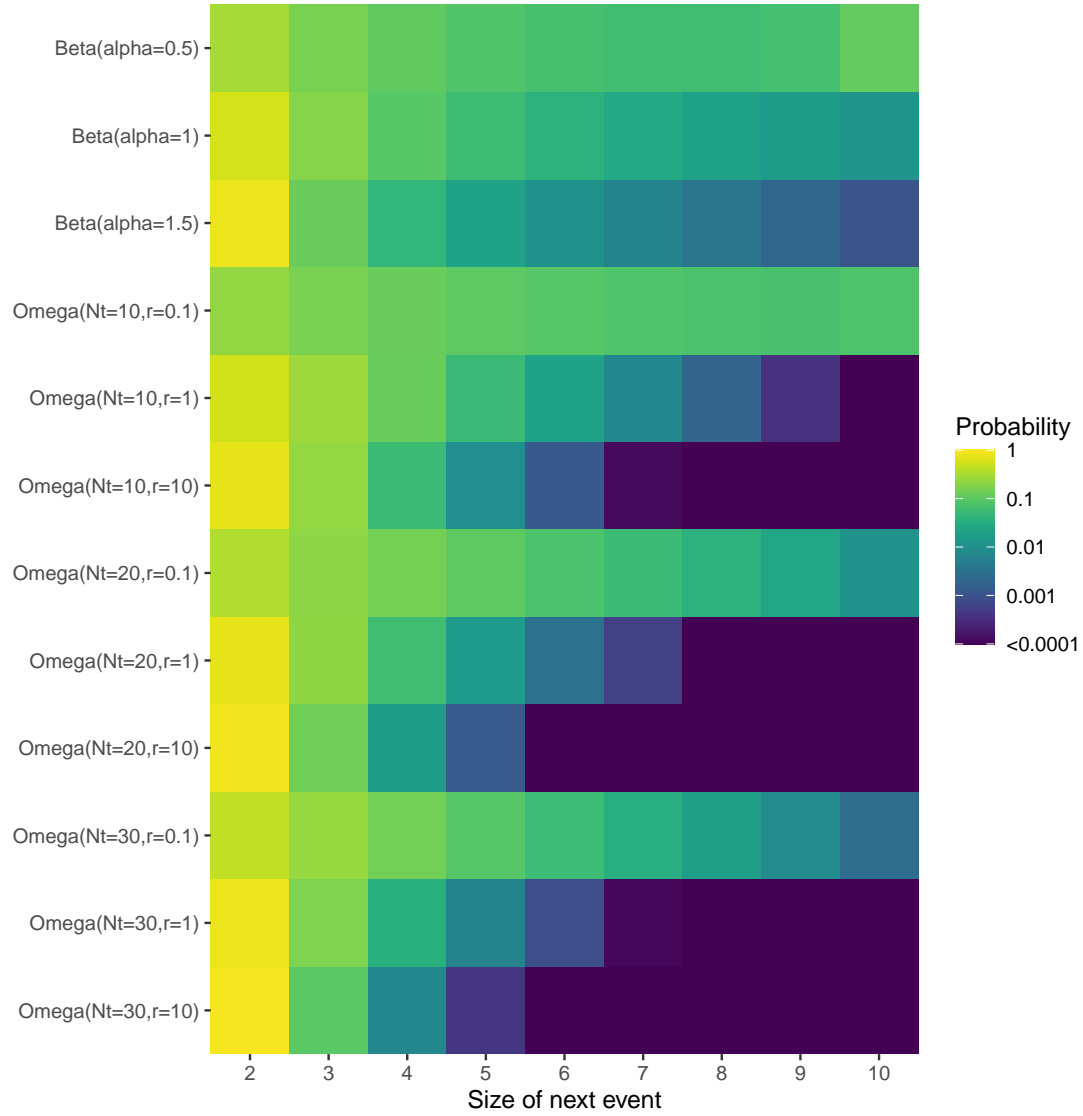


Figure 3: Distribution of the size of the next event among a set of  $n = 10$  lineages, compared between the beta-coalescent and the omega-coalescent model with various parameters.

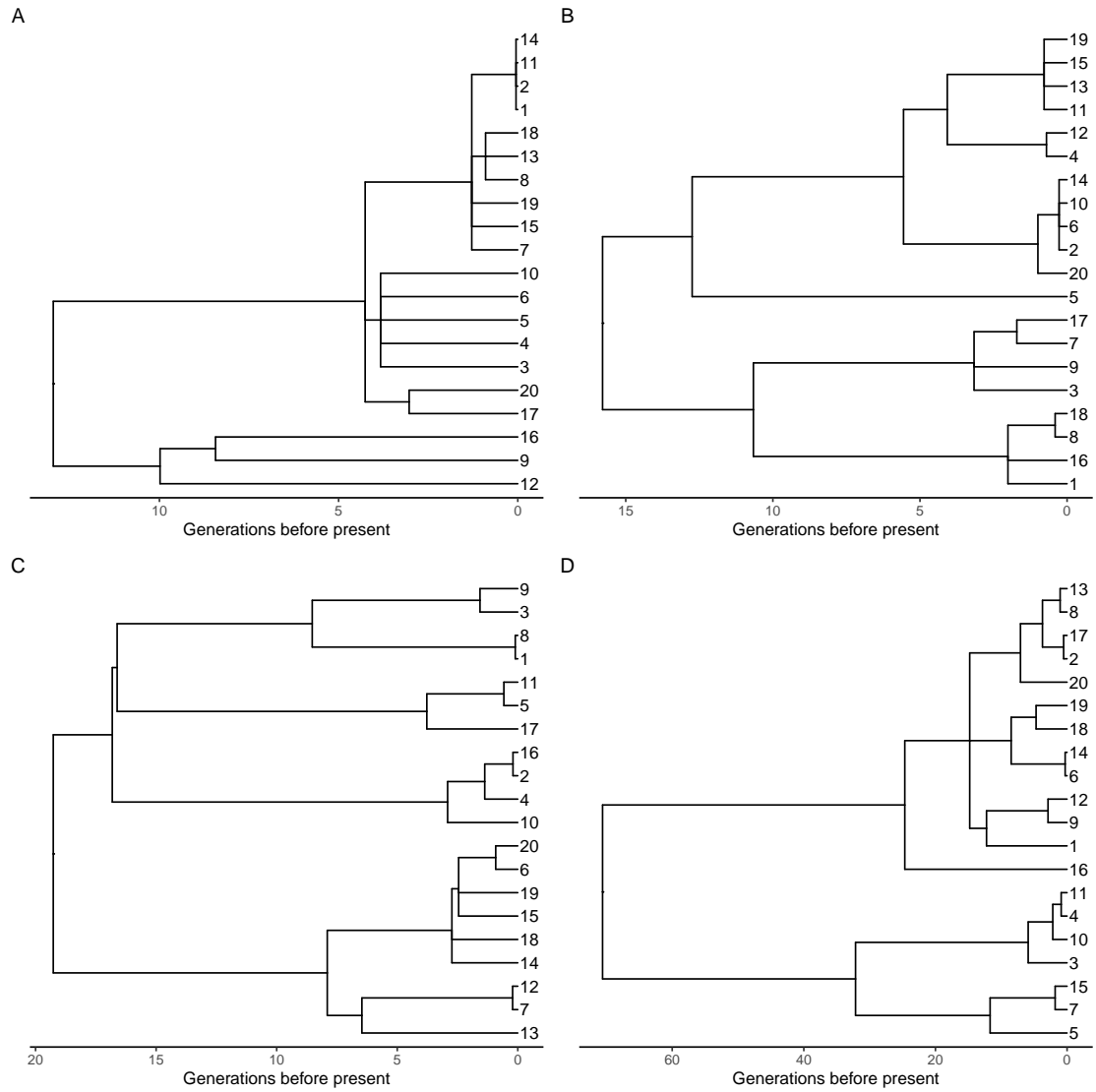


Figure 4: Example of trees simulated under the omega-coalescent with  $r = 0.1$  (A),  $r = 1$  (B),  $r = 10$  (C) and  $r = 100$  (D).

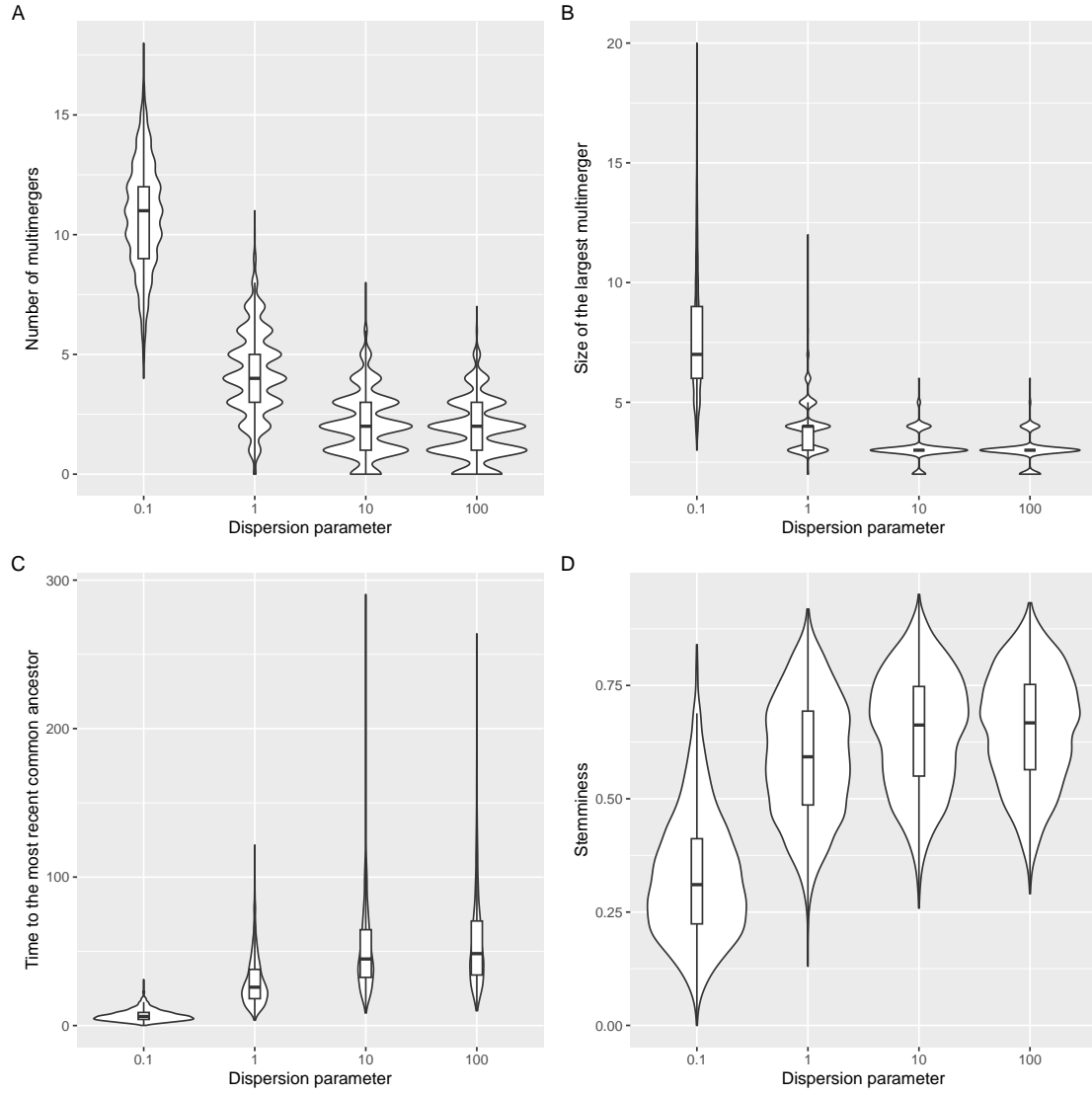


Figure 5: Summary statistics for trees simulated under the omega-coalescent with  $r = 0.1$ ,  $r = 1$ ,  $r = 10$  and  $r = 100$ , namely number of multimerers (A) the size of the largest multimer (B), the time to the most recent common ancestor (C) and the stemminess (D).

## 7 Parameter inference

Let us now consider a genealogy  $T$  with  $n$  leaves and  $c$  coalescent nodes, with  $t_0 = 0$  the sampling time,  $t_1, \dots, t_c$  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:

$$p(T|\Lambda) = \prod_{i=1}^c \lambda_{n_i, k_i} \exp \left( - \sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i, j} (t_i - t_{i-1}) \right) \quad (31)$$

Note that in Equation 31 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 (Koskela 2018; Miró Pina et al. 2023). 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  $N_t$  (which remains constant) and the dispersion parameter  $r$  of the Negative-Binomial offspring distribution. We perform estimation simply by maximising the likelihood in Equation 31, using the Brent algorithm (Brent 1971) when estimating a single parameter and the L-BFGS-B algorithm (Byrd et al. 1995) when estimating both parameters.

We simulated 100 genealogies from the omega-coalescent model each of which had  $n = 100$  leaves, with parameter  $N_t$  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  $N_t = 200$  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.

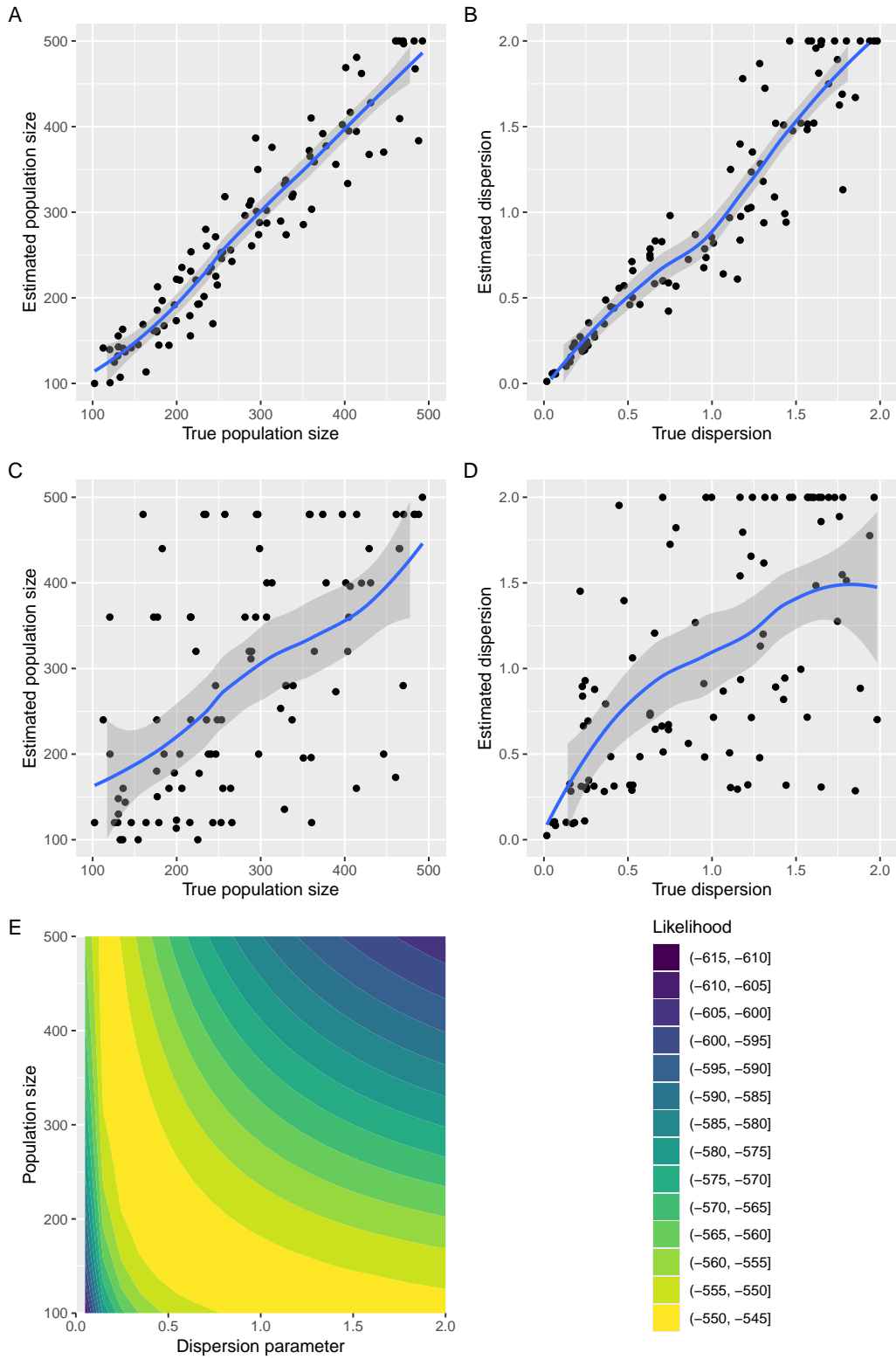


Figure 6: Maximum likelihood estimation of parameters. (A) Estimation of the population size given the dispersion parameter. (B) Estimation of the dispersion parameter given the population size. (C and D) Joint estimation of both the population size and dispersion parameters. (E) Example of likelihood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

## 272 8 Implementation

273 We implemented the analytical methods described in this paper in a new R package entitled *EpiLambda*  
274 which is available at <https://github.com/xavierdidelot/EpiLambda> for R version 3.5 or later. All  
275 code and data needed to replicate the results are included in the “run” directory of the *EpiLambda*  
276 repository. The R package `ape` was used to store, manipulate and visualise phylogenetic trees (Paradis  
277 and Schliep 2019).

## 278 9 Discussion

279 We have described an ancestral process for infectious diseases which is relevant to the analysis of  
280 outbreaks of a relatively small size, and to diseases with transmission heterogeneity. We have shown  
281 how this process can be incorporated into a new lambda-coalescent which we called the omega-  
282 coalescent. We only considered the situation where all samples are taken at the same time, but  
283 the omega-coalescent could be extended to allow temporally offset leaves following similar work on the  
284 coalescent (Drummond et al. 2003) and the beta-coalescent (Hoscheit and Pybus 2019). We also made  
285 the simplifying assumption of a constant population size, but this could be relaxed following the same  
286 approach as previously described for integrating variable population size into the coalescent (Griffiths  
287 and Tavaré 1994; Pybus et al. 2000; Ho and Shapiro 2011) and the beta-coalescent (Hoscheit and Pybus  
288 2019; Zhang and Palacios 2024). Allowing the population size to vary could be especially useful for the  
289 omega-coalescent for several reasons. Firstly, since it is aimed at relatively small outbreaks, it is likely  
290 that their sizes varies significantly. Secondly, the probability of multimerger events of various sizes  
291 depends explicitly on the population size in Equation 21. Changes in population size will therefore  
292 have an effect on the distribution of events observed, as can be seen for example in Figure 3. Thirdly,  
293 joint inference of a varying population size could help break the otherwise difficult joint inference of a  
294 fixed population size with the dispersion parameter (Figure 6).

295 We compared the omega-coalescent only to the beta-coalescent (Schweinsberg 2003) in Figure 3  
296 as it is the model that has been most frequently used for infectious diseases (Hoscheit and Pybus  
297 2019; Menardo et al. 2021; Helekal et al. 2025). Several other lambda-coalescent models have been  
298 proposed previously, such as the Dirac coalescent (Eldon and Wakeley 2006), the Durrett-Schweinsberg  
299 coalescent (Durrett and Schweinsberg 2005) or the extended Beta-coalescent (Helekal et al. 2025).  
300 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 (Schweinsberg 2000). 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 (Volz and Frost 2017; Didelot et al. 2018; Bouckaert et al. 2019; Helekal et al. 2025). We therefore introduced a continuous time approximation in Equation 29 so that ancestry events do not co-occur.

Finally, it should be noted that our model describes the transmission tree during an outbreak, which is different from a phylogeny (Jombart et al. 2011). 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 (Didelot et al. 2014; Hall et al. 2015; Didelot et al. 2017). 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 (Riley et al. 2003; Wallinga and Teunis 2004; Ho et al. 2023). 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.



## References

- Anderson, R.M., May, R.M., 1991. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, USA.
- Bouckaert, R., Vaughan, T.G., Fourment, M., Gavryushkina, A., Heled, J., Denise, K., Maio, N.D., Matschiner, M., Ogilvie, H., Plessis, L., Popinga, A., 2019. BEAST 2.5 : An Advanced Software Platform for Bayesian Evolutionary Analysis. *PLoS computational biology* 15, e1006650.
- Brent, R.P., 1971. An algorithm with guaranteed convergence for finding a zero of a function. *The computer journal* 14, 422–425.
- Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing* 16, 1190–1208.
- 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.
- Didelot, X., Croucher, N.J., Bentley, S.D., Harris, S.R., Wilson, D.J., 2018. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Research* 46, e134–e134.
- 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.
- 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.
- Didelot, X., Urwin, R., Maiden, M.C.J., Falush, D., 2009. Genealogical typing of *Neisseria meningitidis*. *Microbiology* 155, 3176–86.
- Donnelly, P., Kurtz, T.G., 1999. Particle Representations for Measure-Valued Population Models. *The Annals of Probability* 27.
- 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.
- 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.
- Durrett, R., Schweinsberg, J., 2005. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications* 115, 1628–1657.

353 Eldon, B., Wakeley, J., 2006. Coalescent Processes When the Distribution of Offspring Number Among  
354 Individuals Is Highly Skewed. *Genetics* 172, 2621–2633.

355 Ferguson, N.M., Cummings, D.A.T., Fraser, C., Cajka, J.C., Cooley, P.C., Burke, D.S., 2006. Strategies  
356 for mitigating an influenza pandemic. *Nature* 442, 448–452.

357 Fiala, K.L., Sokal, R.R., 1985. Factors determining the accuracy of cladogram estimation: Evolution  
358 using computer simulation. *Evolution* 39, 609–622.

359 Fisher, R.A., 1930. The genetical theory of natural selection. Clarendon Press.

360 Fraser, C., Li, L.M., 2017. Coalescent models for populations with time-varying population sizes and  
361 arbitrary offspring distributions. *bioRxiv* , 10.1101/131730.

362 Fraser, C., Riley, S., Anderson, R.M., Ferguson, N.M., 2004. Factors that make an infectious disease  
363 outbreak controllable. *Proceedings of the National Academy of Sciences* 101, 6146–6151.

364 Gómez-Carballa, A., Pardo-Seco, J., Bello, X., Martínón-Torres, F., Salas, A., 2021. Superspreading  
365 in the emergence of COVID-19 variants. *Trends in Genetics* 37, 1069–1080.

366 Grassly, N.C., Fraser, C., 2008. Mathematical models of infectious disease transmission. *Nature*  
367 *Reviews Microbiology* 6, 477–87.

368 Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment.  
369 *Philosophical Transactions of the Royal Society B* 344, 403–410.

370 Hall, M., Woolhouse, M., Rambaut, A., 2015. Epidemic Reconstruction in a Phylogenetics Framework:  
371 Transmission Trees as Partitions of the Node Set. *PLOS Computational Biology* 11, e1004613.

372 Helekal, D., Koskela, J., Didelot, X., 2025. Inference of multiple mergers while dating a pathogen  
373 phylogeny. *Systematic Biology* , in press.

374 Ho, F., Parag, K.V., Adam, D.C., Lau, E.H.Y., Cowling, B.J., Tsang, T.K., 2023. Accounting for the  
375 Potential of Overdispersion in Estimation of the Time-varying Reproduction Number. *Epidemiology*  
376 34, 201–205.

377 Ho, S.Y.W., Shapiro, B., 2011. Skyline-plot methods for estimating demographic history from  
378 nucleotide sequences. *Molecular Ecology Resources* 11, 423–434.

379 Hoscheit, P., Pybus, O.G., 2019. The multifurcating skyline plot. *Virus Evolution* 5, 1–10.

380 Jombart, T., Eggo, R.M., Dodd, P.J., Balloux, F., 2011. Reconstructing disease outbreaks from genetic  
381 data: A graph approach. *Heredity* 106, 383–90.

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

384 Kingman, J., 1982a. The coalescent. *Stochastic Processes and their Applications* 13, 235–248.

385 Kingman, J.F.C., 1982b. On the genealogy of large populations. *Journal of Applied Probability* 19,  
386 27–43.

387 Koelle, K., Rasmussen, D.A., 2012. Rates of coalescence for common epidemiological models at  
388 equilibrium. *Journal of The Royal Society Interface* 9, 997–1007.

389 Koskela, J., 2018. Multi-locus data distinguishes between population growth and multiple merger  
390 coalescents. *Statistical Applications in Genetics and Molecular Biology* 17, 1–24.

391 Kucharski, A.J., Althaus, C.L., 2015. The role of superspreading in Middle East respiratory syndrome  
392 coronavirus (MERS-CoV) transmission. *Eurosurveillance* 20.

393 Lemieux, J.E., Siddle, K.J., Shaw, B.M., Loreth, C., Schaffner, S.F., Gladden-Young, A., Adams,  
394 G., Fink, T., Tomkins-Tinch, C.H., Krasilnikova, L.A., DeRuff, K.C., Rudy, M., Bauer, M.R.,  
395 Lagerborg, K.A., Normandin, E., Chapman, S.B., Reilly, S.K., Anahtar, M.N., Lin, A.E., Carter,  
396 A., Myhrvold, C., Kembball, M.E., Chaluvadi, S., Cusick, C., Flowers, K., Neumann, A., Cerrato,  
397 F., Farhat, M., Slater, D., Harris, J.B., Branda, J.A., Hooper, D., Gaeta, J.M., Baggett, T.P.,  
398 O’Connell, J., Gnirke, A., Lieberman, T.D., Philippakis, A., Burns, M., Brown, C.M., Luban, J.,  
399 Ryan, E.T., Turbett, S.E., LaRocque, R.C., Hanage, W.P., Gallagher, G.R., Madoff, L.C., Smole, S.,  
400 Pierce, V.M., Rosenberg, E., Sabeti, P.C., Park, D.J., MacInnis, B.L., 2021. Phylogenetic analysis  
401 of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* 371, eabe3261.

402 Li, L.M., Grassly, N.C., Fraser, C., 2017. Quantifying Transmission Heterogeneity Using Both  
403 Pathogen Phylogenies and Incidence Time Series. *Molecular Biology and Evolution* 34, 2982–2995.

404 Lloyd-Smith, J., Schreiber, S., Kopp, P., Getz, W., 2005. Superspreading and the effect of individual  
405 variation on disease emergence. *Nature* 438, 355–9.

406 Menardo, F., Gagneux, S., Freund, F., 2021. Multiple Merger Genealogies in Outbreaks of  
407 *Mycobacterium tuberculosis*. *Molecular Biology and Evolution* 38, 290–306.

408 Miró Pina, V., Joly, É., Siri-Jégousse, A., 2023. Estimating the Lambda measure in multiple-merger  
409 coalescents. *Theoretical Population Biology* 154, 94–101.

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

412 Paradis, E., Schliep, K., 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary  
413 analyses in R. *Bioinformatics* 35, 526–528.

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

415 Potts, R.B., 1953. Note on the Factorial Moments of Standard Distributions. *Australian Journal of*  
416 *Physics* 6, 498–499. Publisher: CSIRO PUBLISHING.

417 Pybus, O.G., Rambaut, A., Harvey, P.H., 2000. An integrated framework for the inference of viral  
418 population history from reconstructed genealogies. *Genetics* 155, 1429–1437.

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

423 Sagitov, S., 1999. The general coalescent with asynchronous mergers of ancestral lines. *Journal of*  
424 *Applied Probability* 36, 1116–1125.

425 Schweinsberg, J., 2000. Coalescents with Simultaneous Multiple Collisions. *Electronic Journal of*  
426 *Probability* 5.

427 Schweinsberg, J., 2003. Coalescent processes obtained from supercritical Galton–Watson processes.  
428 *Stochastic Processes and their Applications* 106, 107–139.

429 Stein, R.A., 2011. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*  
430 15, e510–e513.

431 Svensson, A., 2007. A note on generation times in epidemic models. *Mathematical Biosciences* 208(1),  
432 300–311.

433 Tripathi, R.C., Gupta, R.C., Gurland, J., 1994. Estimation of parameters in the beta binomial model.  
434 *Annals of the Institute of Statistical Mathematics* 46, 317–331.

435 Volz, E.M., 2012. Complex population dynamics and the coalescent under neutrality. *Genetics* 190,  
436 187–201.

437 Volz, E.M., Frost, S.D.W., 2017. Scalable relaxed clock phylogenetic dating. *Virus Evolution* 3, vex025.

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

440 Wang, J., Chen, X., Guo, Z., Zhao, S., Huang, Z., Zhuang, Z., Wong, E.L.y., Zee, B.C.Y., Chong,  
441 M.K.C., Wang, M.H., Yeoh, E.K., 2021. Superspreading and heterogeneity in transmission of SARS,  
442 MERS, and COVID-19: A systematic review. *Computational and Structural Biotechnology Journal*  
443 19, 5039–5046.

444 Wang, L., Didelot, X., Yang, J., Wong, G., Shi, Y., Liu, W., Gao, G.F., Bi, Y., 2020. Inference of  
445 person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early  
446 outbreak phase. *Nature Communications* 11, 5006.

447 Woolhouse, M.E.J., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., Hagan, P., Hii,  
448 J.L.K., Ndhlovu, P.D., Quinnell, R.J., Watts, C.H., Chandiwana, S.K., Anderson, R.M., 1997.  
449 Heterogeneities in the transmission of infectious agents: Implications for the design of control  
450 programs. *Proceedings of the National Academy of Sciences* 94, 338–342.

451 Wright, S., 1931. Evolution in Mendelian populations. *Genetics* 16, 97–159.

452 Zhang, J., Palacios, J.A., 2024. Multiple merger coalescent inference of effective population size. *arXiv*  
453 , 2407.14976.