

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

5 ³ 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:

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

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:

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

and the conditional distribution:

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

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

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

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

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

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:

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

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 $t+1$. The
 116 inclusive coalescent probability in Equation 12 can also be obtained conceptually by considering that
 117 among the k lines, the first one has an ancestor with probability one, and the remaining $k-1$ need to
 118 have the same ancestor among a set of N_t from which they choose uniformly at random so that the
 119 probability of picking the same ancestor is $1/N_t$. The exclusive coalescent probability in Equation 14
 120 can be derived likewise by considering that in addition to the above, each of the $n-k$ other lines need
 121 to choose a different ancestor, which happens with probability $(N_t - 1)/N_t$.

122 Figure 1 illustrates the inclusive and exclusive coalescence probabilities for the Poisson case for a set
 123 of size $k = 1$ to $k = 10$ amongst a total of $n = 10$ observed lines, in a population of size $N_t = 10$,
 124 $N_t = 20$ 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 $r =$

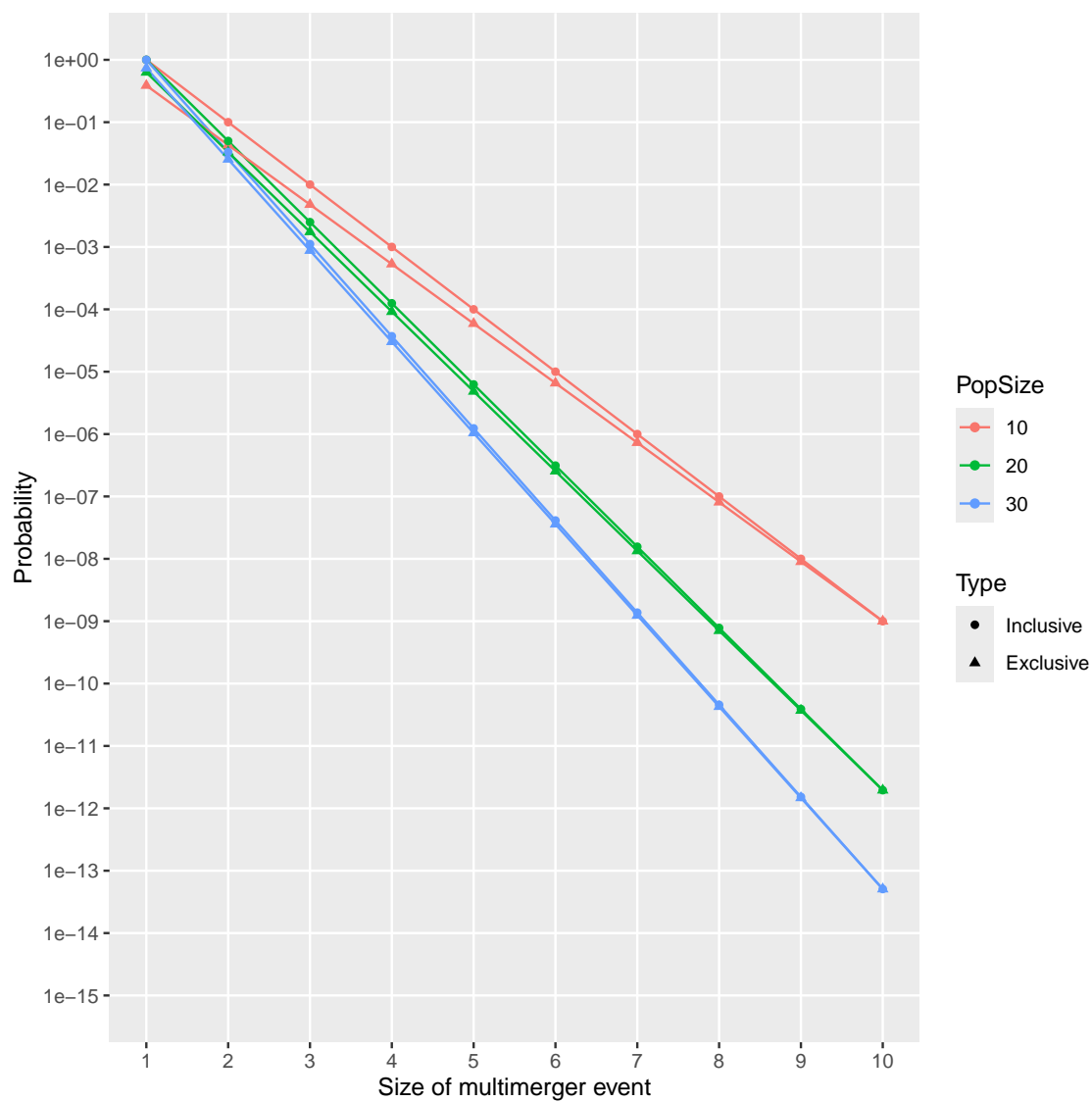


Figure 1: Inclusive and exclusive coalescence probabilities for the Poisson case.

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 size
149 $N_t = 20$. Several Negative-Binomial offspring distributions are compared, all of which have the same
150 mean $R_t = 2$, and with the dispersion parameter equal to $r = 0.1$, $r = 1$, $r = 10$ and $r = 100$. When
151 $r = 1$ the Negative-Binomial reduces to a Geometric distribution. When r is high (for example $r = 100$
152 as shown in Figure 2) the dispersion is low and the Negative-Binomial case behaves almost like the
153 Poisson case. When r is lower the dispersion of the offspring distribution increases, so that both the
154 inclusive and exclusive probabilities of larger multimerger events increase.

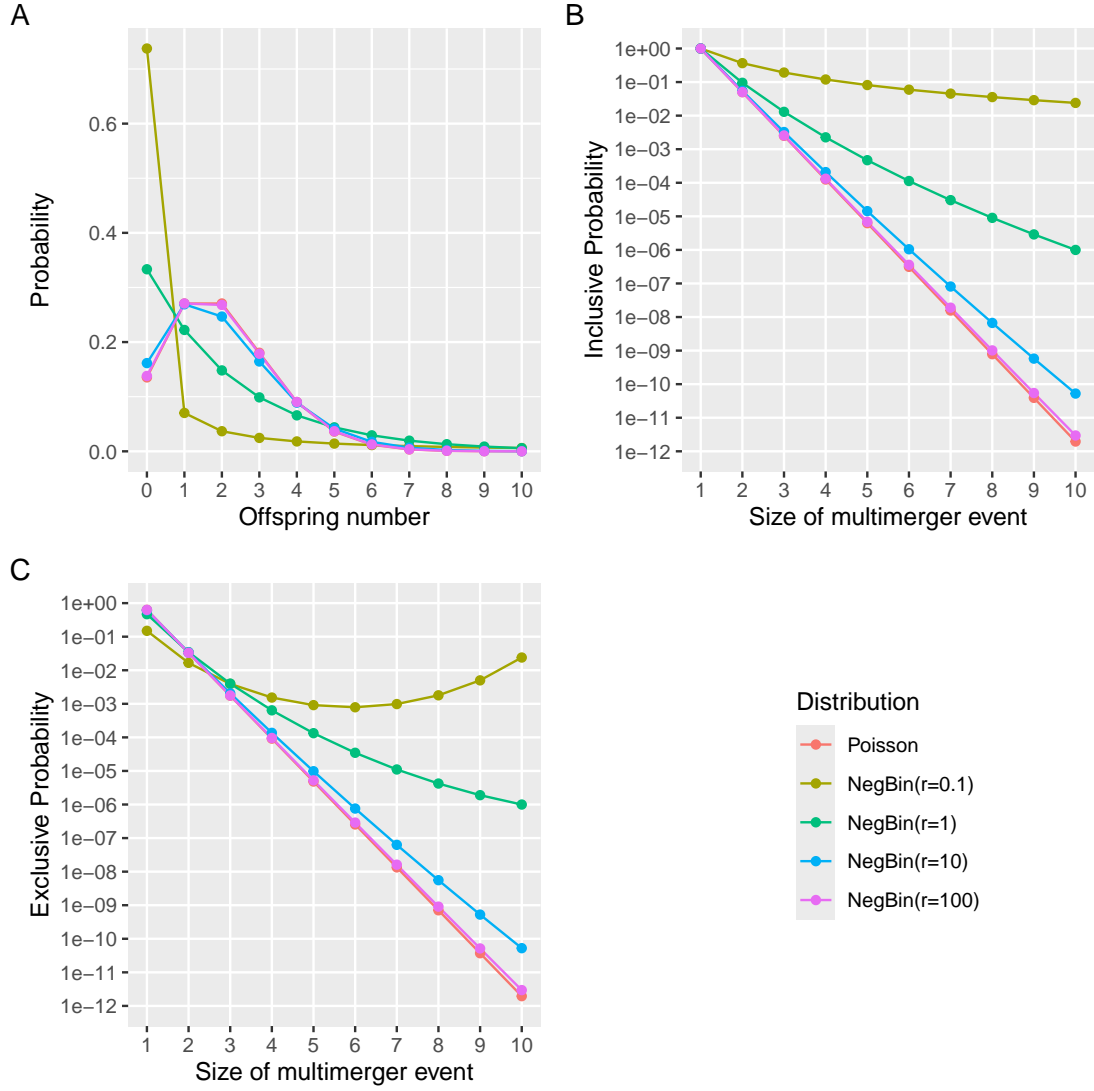


Figure 2: (A) Offspring distributions with mean $R_t = 2$. (B) Inclusive probability of coalescence for $N_t = 20$ and $n = 10$. (C) Exclusive probability of coalescence for $N_t = 20$ and $n = 10$.

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 results and several previous studies. In the Poisson case, from Equations 12 and 14 we can see that both inclusive and exclusive probabilities are of order $\mathcal{O}(N_t^{1-k})$. We can therefore ignore events with $k > 2$ and retain only the events with $k = 2$ which occur with probability:

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

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 probability:

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

Koelle and Rasmussen (2012) 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 $N_e = N/\sigma^2$ 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 Kingman (1982a) to derive the backward-in-time coalescent model from the forward-in-time Cannings exchangeable models (Cannings 1974). 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 $p_{2,t} = V_t/N_t$.

Volz (2012) 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 and $I(t)$ the prevalence. Setting in this formula the prevalence as $I(t) = N_{t+1} = N_t R_t$ and the incidence as $f(t) = R_t N_{t+1} = R_t^2 N_t$ 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 (offspring) 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.

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 . This formula was used to estimate the dispersion parameter of a Negative-Binomial offspring distribution from genetic data (Li et al. 2017). In our notation, their formula is equivalent to the inclusive coalescence probability for two lineages:

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

In the Poisson case we have $\sigma^2 = R$ so that Equation 25 simplifies to $1/N_t$ which agrees with Equation 23. In the Negative-Binomial case we have $\sigma^2/R = 1/p = (r + R)/r$ 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 $r = R^2/(\sigma^2 - R)$ in Equation 24 we obtain the Equation 25.

6 Definition of a new lambda-coalescent model

The coalescent model (Kingman 1982a,b) describes the ancestry of a sample from a large population evolving according to many forward-in-time models such as the Wright-Fisher model (Wright 1931; Fisher 1930), the Moran model (Moran 1958) and the Cannings exchangeable model (Cannings 1974). 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 multimergers, making them unsuitable to represent the ancestry of outbreaks considered in this study. However, the lambda-coalescent models are an extension of the coalescent model that do allow multimergers (Pitman 1999; Sagitov 1999; Donnelly and Kurtz 1999).

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

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

The beta-coalescent (Schweinsberg 2003) is a specific type of lambda-coalescent that has been used recently in several studies analysing genetic data from infectious disease agents (Hoscheit and Pybus

201 2019; Menardo et al. 2021; Helekal et al. 2025; Zhang and Palacios 2024). The beta-coalescent model
 202 has a single parameter $\alpha \in [0, 2]$ and is defined as:

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

203 By combining Equations 26 and 27 we can deduce that:

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

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

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

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

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

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

217 Figure 3 compares this distribution for $n = 10$ in the beta-coalescent with parameter $\alpha \in \{0.5, 1, 1.5\}$
 218 and for the omega-coalescent with parameters $N_t \in \{10, 20, 30\}$ and $r \in \{0.1, 1, 10\}$. In the beta-

219 coalescent, the distribution shifts towards more larger multimerger events as the parameter α decreases.
 220 In the omega-coalescent a wider range of behaviours is obtained when varying the two parameters N_t
 221 and r . For a given value of N_t , decreasing the value of r results in more larger events. Conversely, for
 222 a given value of r we can see that increasing the value of N_t reduces the probability of larger events.

223 Genealogies can be simulated from the omega-coalescent model defined in Equation 29 using the
 224 same algorithm as for other lambda-coalescent models (Pitman 1999). Figure 4 shows examples of
 225 trees simulated for a sample of size $n = 20$, constant population size $N_t = 30$ and dispersion parameter
 226 $r \in \{0.1, 1, 10, 100\}$. It is already clear from these single realisations that the lower values of r result
 227 in trees with more larger multimerger events and lower time to the most recent common ancestor, but
 228 to quantify these properties we need to consider many trees.

229 Figure 5 shows summary statistics for 10,000 trees simulated in the same conditions as the individual
 230 trees shown in Figure 4. As the dispersion parameter increases from $r = 0.1$ to $r = 100$ multimerger
 231 events become less and less likely and large. Simultaneously, the time to the most recent common
 232 ancestor increases, as well as the stemminess of the tree (ie the proportion of branch lengths in non-
 233 terminal branches).

234 7 Parameter inference

235 Consider a genealogy T with n leaves and c coalescent nodes, with $t_0 = 0$ the sampling time, t_1, \dots, t_c
 236 the times of the coalescent nodes in increasing order and k_i the number of lineages coalescing at time
 237 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
 238 lambda-coalescent model, the genealogy T has likelihood:

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

239 Note that in Equation 31 the term $\binom{n_i}{k_i}$ term from the coalescent rate cancels out with its reciprocal
 240 from the probability of sampling k_i specific lineages to coalesce within a set of n_i . Estimating the
 241 lambda measure in general is a difficult problem (Koskela 2018; Miró Pina et al. 2023). Here however
 242 we focus on estimation under the omega-coalescent model, where the $\lambda_{n, k}$ terms are given by Equation
 243 29. There are therefore two parameters to estimate which have direct and important biological

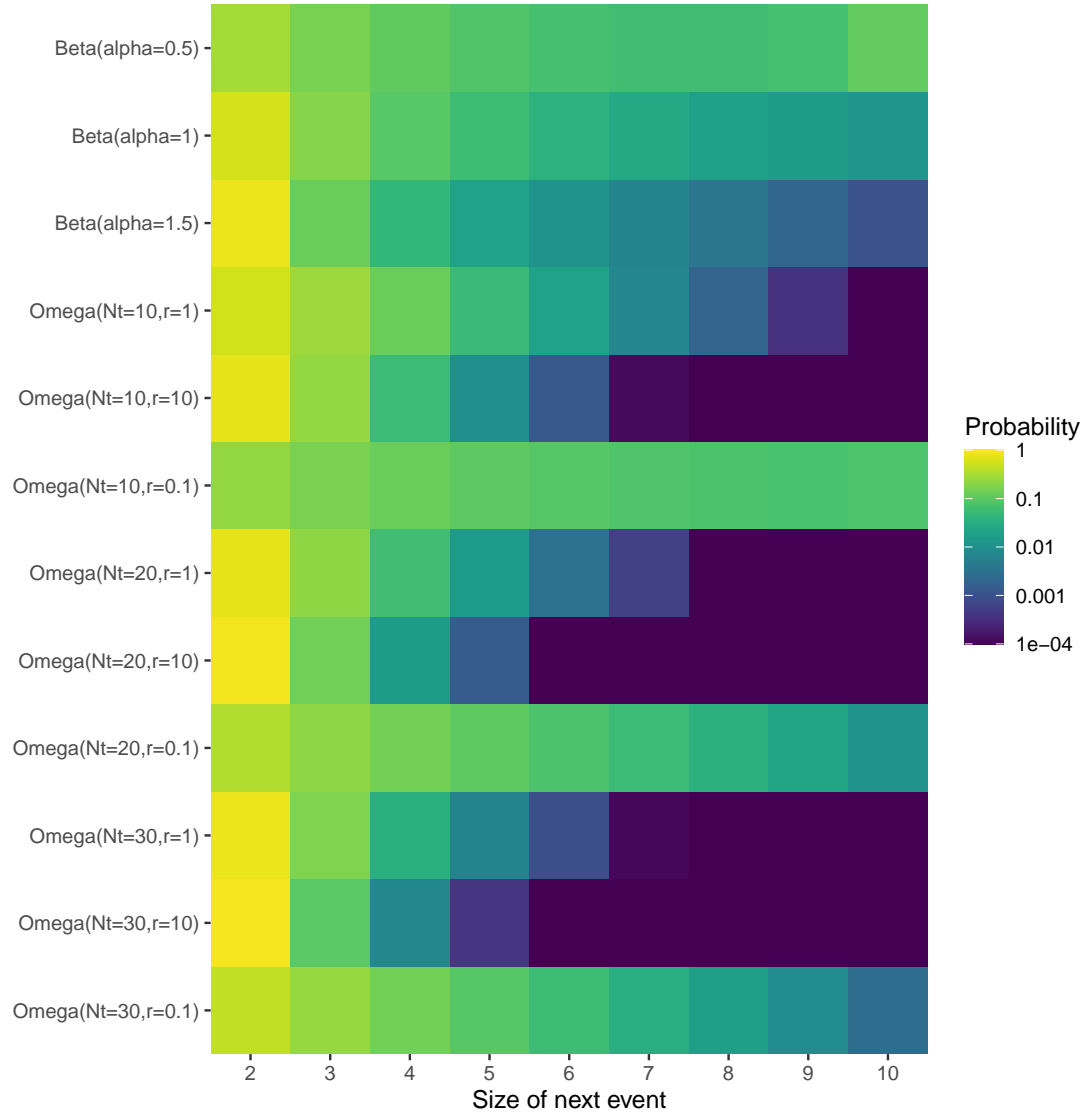


Figure 3: Distribution of the size of the next event among a set of $n = 10$ lineages, compared between the beta-coalescent and the omega-coalescent model with various parameters.

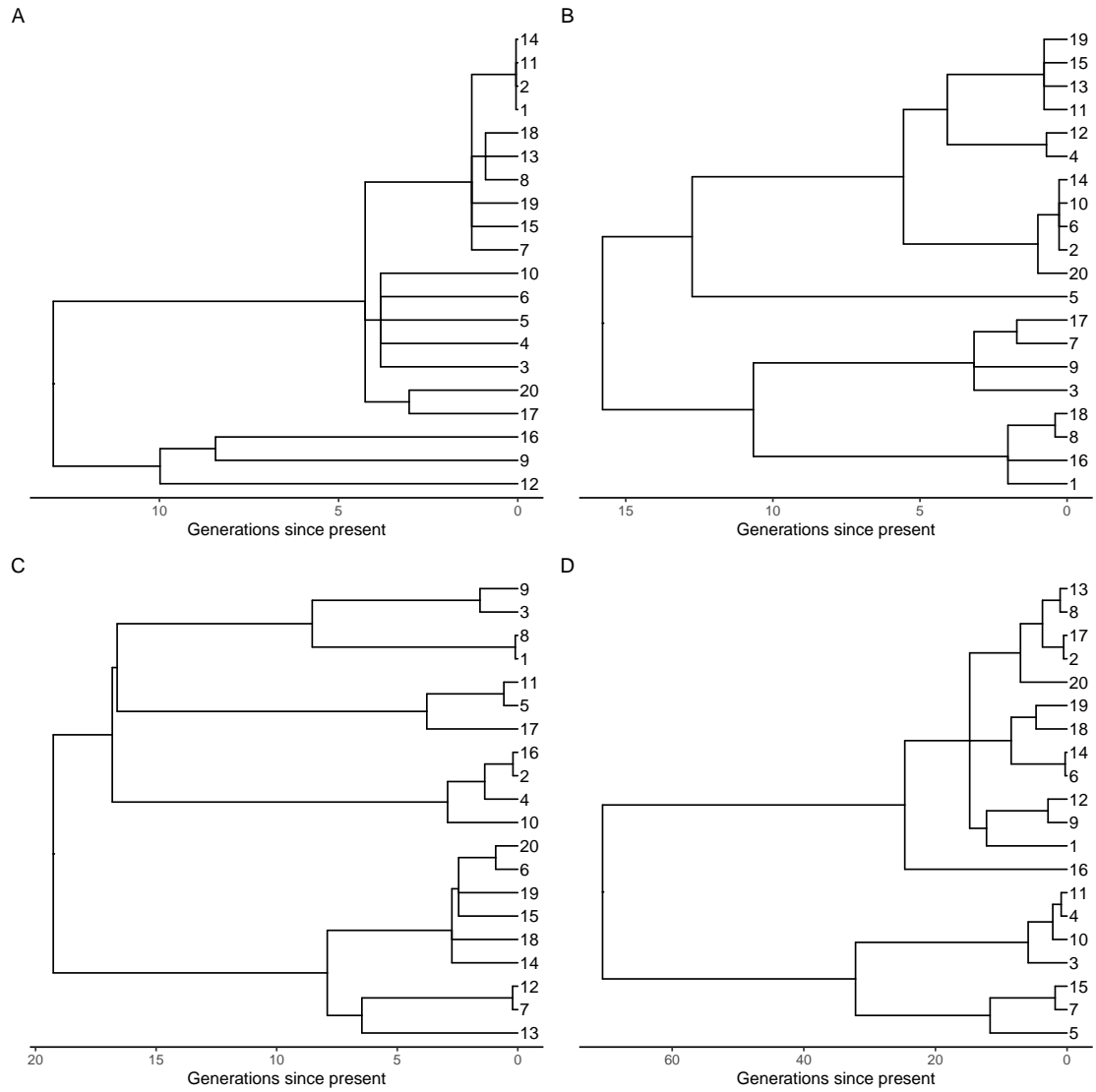


Figure 4: Example of trees simulated under the omega-coalescent with $r = 0.1$ (A), $r = 1$ (B), $r = 10$ (C) and $r = 100$ (D).

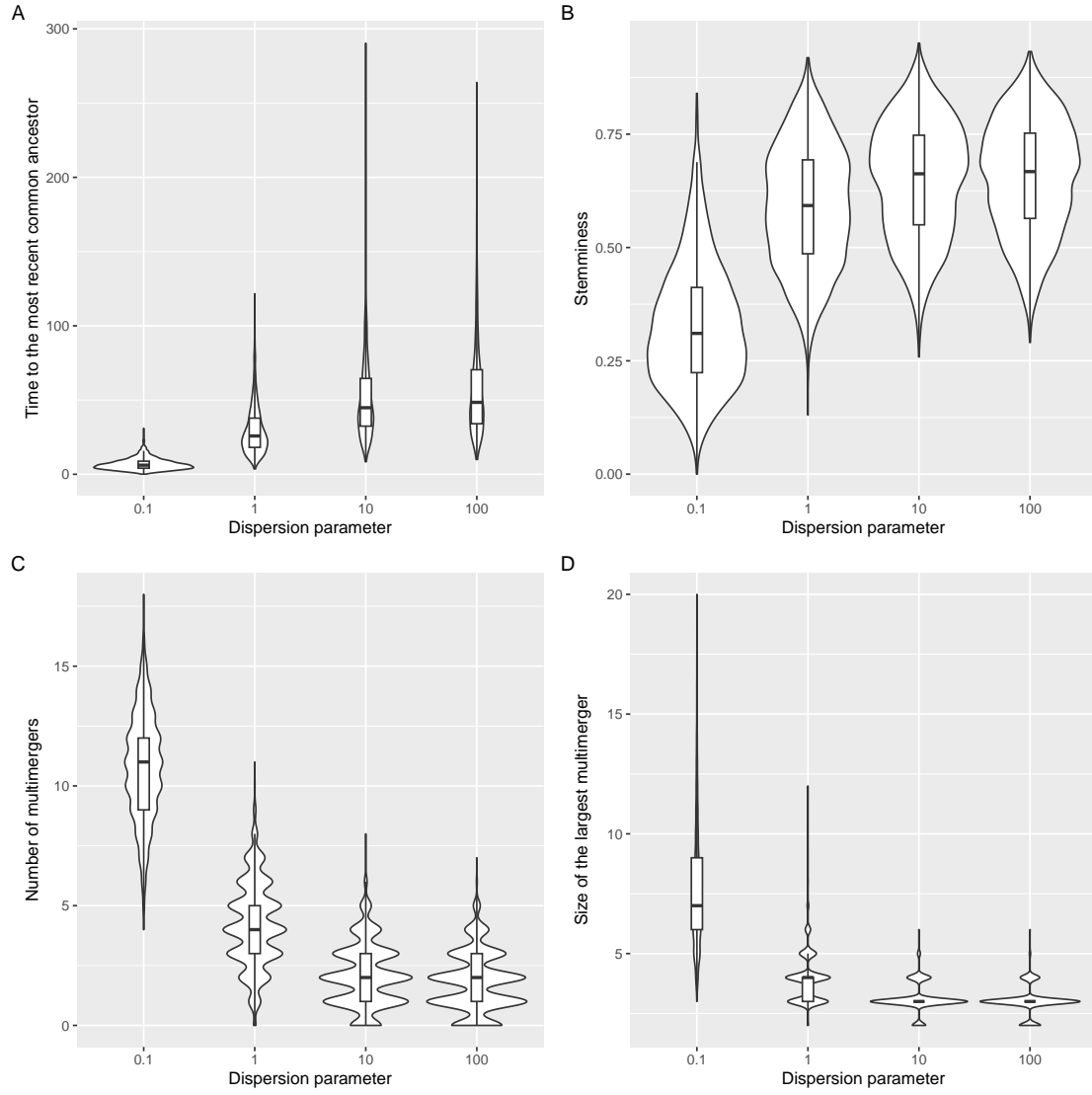


Figure 5: Summary statistics for trees simulated under the omega-coalescent with $r = 0.1$, $r = 1$, $r = 10$ and $r = 100$, namely the time to the most recent common ancestor (A), stemminess (B), number of multimerers (C) and the size of the largest multimerger (D).

244 meaning: the effective population size N_t (which remains constant) and the dispersion parameter
 245 r of the Negative-Binomial offspring distribution. We perform estimation simply by maximising the
 246 likelihood in Equation 31, using the Brent algorithm (Brent 1971) when estimating a single parameter
 247 and the L-BFGS-B algorithm when (Byrd et al. 1995) estimating both parameters.

248 We simulated 100 genealogies from the omega-coalescent model each of which had $n = 100$ leaves, with
 249 parameter N_e drawn uniformly at random between 100 and 500 and parameter r drawn uniformly at
 250 random between 0.01 and 2. If we assume knowledge of the dispersion parameter, then estimating
 251 the population size works really well (Figure 6A). Conversely we obtain good result when estimating
 252 the dispersion parameter given a known population size (Figure 6B). However, attempting to estimate
 253 both parameters at the same time performed significantly less well (Figures 6C and D). To illustrate
 254 the cause of this, we consider a simulation for which the true N_t was 200 and the true r was 0.5, and
 255 we construct the likelihood surface (Figure 6E). This shows a strong inverse tradeoff between the two
 256 parameters, which explains why one can be estimated given the other, but not jointly.

257 8 Implementation

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

263 9 Discussion

264 The omega-coalescent could be extended to allow temporally offset leaves following work on the
 265 coalescent (Drummond et al. 2003) and the beta-coalescent (Hoscheit and Pybus 2019). It could
 266 also be defined in a varying population size following the same approach as previously described for
 267 the coalescent (Griffiths and Tavare 1994; Pybus et al. 2000; Ho and Shapiro 2011) and the beta-
 268 coalescent (Hoscheit and Pybus 2019; Zhang and Palacios 2024). This could be even more useful
 269 for the omega-coalescent than for the beta-coalescent since in the omega-coalescent the probability of
 270 multimerger events of various size depends explicitly on the population size (see for example Figure

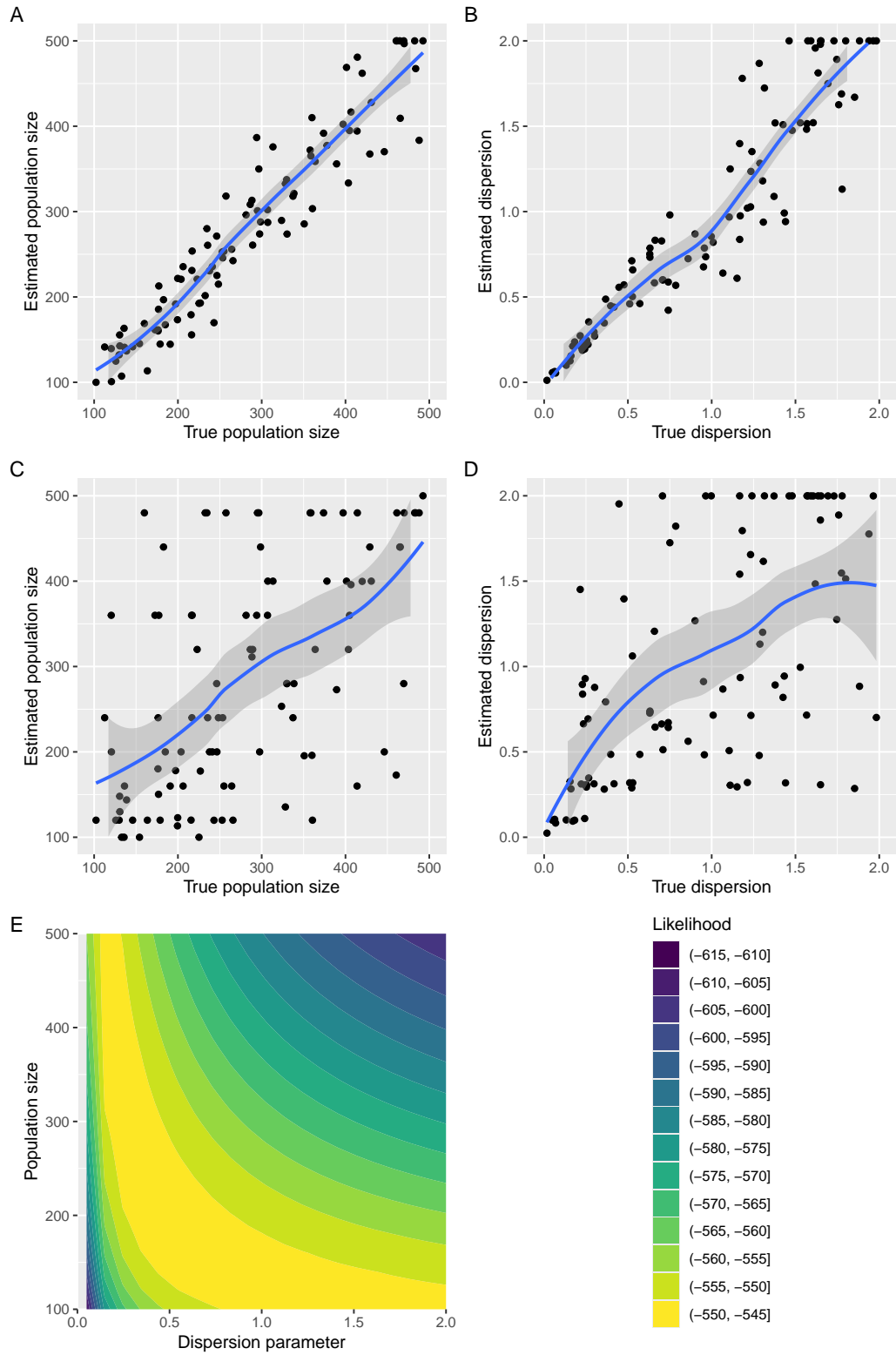


Figure 6: Maximum likelihood estimation of parameters. (A) Estimation of the population size given the dispersion parameter. (B) Estimation of the dispersion parameter given the population size. (C and D) Joint estimation of both the population size and dispersion parameters. (E) Example of likelihood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

271 3).

272 We compared the omega-coalescent only to the beta-coalescent (Schweinsberg 2003) but it could also
273 be compared to the Dirac coalescent aka psi-coalescent (Eldon and Wakeley 2006), the Durrett-
274 Schweinsberg coalescent (Durrett and Schweinsberg 2005) or the extended Beta-coalescent (Helekal
275 et al. 2025).

276 The xi-coalescent models admit multiple simultaneous mergers (Schweinsberg 2000).

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

281 Acknowledgements

282 We acknowledge funding from the National Institute for Health Research (NIHR) Health Protection
283 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.
- 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. 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.
- 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. doi:10.1016/j.spa.2005.04.009.
- Eldon, B., Wakeley, J., 2006. Coalescent Processes When the Distribution of Offspring Number Among Individuals Is Highly Skewed. *Genetics* 172, 2621–2633. doi:10.1534/genetics.105.052175.
- 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.

313 Fisher, R.A., 1930. The genetical theory of natural selection. Clarendon Press. doi:10.5962/bh1.
 314 title.27468.

315 Fraser, C., Li, L.M., 2017. Coalescent models for populations with time-varying population sizes and
 316 arbitrary offspring distributions. bioRxiv , 10.1101/131730doi:10.1101/131730.

317 Fraser, C., Riley, S., Anderson, R.M., Ferguson, N.M., 2004. Factors that make an infectious
 318 disease outbreak controllable. Proceedings of the National Academy of Sciences 101, 6146–6151.
 319 doi:10.1073/pnas.0307506101.

320 Gómez-Carballa, A., Pardo-Seco, J., Bello, X., Martínón-Torres, F., Salas, A., 2021. Superspreading
 321 in the emergence of COVID-19 variants. Trends in Genetics 37, 1069–1080. doi:10.1016/j.tig.
 322 2021.09.003.

323 Grassly, N.C., Fraser, C., 2008. Mathematical models of infectious disease transmission. Nature
 324 Reviews Microbiology 6, 477–87. doi:10.1038/nrmicro1845.

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

327 Hall, M., Woolhouse, M., Rambaut, A., 2015. Epidemic Reconstruction in a Phylogenetics Framework:
 328 Transmission Trees as Partitions of the Node Set. PLOS Computational Biology 11, e1004613.
 329 doi:10.1371/journal.pcbi.1004613.

330 Heledkal, D., Koskela, J., Didelot, X., 2025. Inference of multiple mergers while dating a pathogen
 331 phylogeny. Systematic Biology , in press.

332 Ho, F., Parag, K.V., Adam, D.C., Lau, E.H.Y., Cowling, B.J., Tsang, T.K., 2023. Accounting for the
 333 Potential of Overdispersion in Estimation of the Time-varying Reproduction Number. Epidemiology
 334 34, 201–205. doi:10.1097/EDE.0000000000001563.

335 Ho, S.Y.W., Shapiro, B., 2011. Skyline-plot methods for estimating demographic history from
 336 nucleotide sequences. Molecular Ecology Resources 11, 423–434. doi:10.1111/j.1755-0998.2011.
 337 02988.x.

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

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

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

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

346 Kingman, J.F.C., 1982b. On the genealogy of large populations. *Journal of Applied Probability* 19,
347 27–43. doi:10.2307/3213548.

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

350 Koskela, J., 2018. Multi-locus data distinguishes between population growth and multiple merger
351 coalescents. *Statistical Applications in Genetics and Molecular Biology* 17, 1–24. doi:10.1515/
352 sagmb-2017-0011.

353 Kucharski, A.J., Althaus, C.L., 2015. The role of superspreading in Middle East respiratory syndrome
354 coronavirus (MERS-CoV) transmission. *Eurosurveillance* 20. doi:10.2807/1560-7917.ES2015.20.
355 25.21167.

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

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

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

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

374 Miró Pina, V., Joly, É., Siri-Jégousse, A., 2023. Estimating the Lambda measure in multiple-merger
375 coalescents. *Theoretical Population Biology* 154, 94–101. doi:10.1016/j.tpb.2023.09.002.

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

378 Paradis, E., Schliep, K., 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary
379 analyses in R. *Bioinformatics* 35, 526–528. doi:10.1093/bioinformatics/bty633.

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

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

384 Pybus, O.G., Rambaut, A., Harvey, P.H., 2000. An integrated framework for the inference of viral
385 population history from reconstructed genealogies. *Genetics* 155, 1429–1437. doi:10.1073/pnas.
386 88.5.1597.

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

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

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

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

398 Stein, R.A., 2011. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*
399 15, e510–e513. doi:10.1016/j.ijid.2010.06.020.

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

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

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

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

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

Woolhouse, M.E.J., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., Hagan, P., Hii, J.L.K., Ndhlovu, P.D., Quinnell, R.J., Watts, C.H., Chandiwana, S.K., Anderson, R.M., 1997. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proceedings of the National Academy of Sciences* 94, 338–342. doi:10.1073/pnas.94.1.338.

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

Zhang, J., Palacios, J.A., 2024. Multiple merger coalescent inference of effective population size. *arXiv*, 2407.14976doi:10.48550/arXiv.2407.14976, arXiv:2407.14976.