

1 The epidemic lambda-coalescent model

2 Xavier Didelot<sup>1,2,\*</sup>, Ian Roberts<sup>2</sup>, ...

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

4

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

6

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

8 Running title: Epidemic lambda-coalescent

9 Keywords: lambda-coalescent model; pathogen phylogenetics; multiple mergers; offspring distribution

# 1 Introduction

Superspreading in infectious disease epidemiology (Lloyd-Smith et al. 2005). For example SARS-CoV-2 superspreading (Wang et al. 2020; Lemieux et al. 2021; Gómez-Carballa et al. 2021). Coalescent model (Kingman 1982a,b). Work by Li and Fraser (Li et al. 2017; Fraser and Li 2017). Lambda-coalescent models (Pitman 1999; Sagitov 1999; Donnelly and Kurtz 1999). Beta-coalescent (Schweinsberg 2003) is a specific type of Lambda-coalescent. Was used in (Hoscheit and Pybus 2019) and (Menardo et al. 2021). David’s paper on inference of multiple mergers while dating a pathogen phylogeny (Helekal et al. 2024).

## 2 General case

Discrete time  $t$ . Non-overlapping generations of infected individuals. At time  $t$  there are  $N_t$  infected individuals. Each of them creates a number  $s_{t,i}$  of secondary infections at time  $t + 1$ , following the offspring distribution  $\alpha_t(s)$ . The mean of this distribution is the basic reproduction number  $R_t$  and the variance is  $V_t$ . We have:

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

Let  $p_{k,t}$  be the probability that  $k$  individuals at time  $t + 1$  have the same infector at time  $t$ .

### 2.1 Inclusive coalescence probability

Inclusive coalescence probability  $p_{k,t}(N_t, N_{t+1})$  is the probability that exactly  $k$  randomly-sampled individuals from generation  $t + 1$  find a common ancestor in generation  $t$  (*conditioning on population sizes  $N_t$  and  $N_{t+1}$* ).

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

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

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

33 To-do:

34 • Define  $S_t^{-(1)} = (S_{t,2}, \dots, S_{t,N_t})$  notation

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

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

## 37 2.2 Exclusive coalescence probability

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

42 Given full information about offspring counts of the parents of sampled individuals at the present,

43  $\mathbf{x}_t = (x_{t,1}, \dots, x_{t,N_t})$ , we have

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

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

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

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

## 49 2.3 Complementarity of exclusive coalescence probabilities

50 Note that for any offspring distribution we have:

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

$$\sum_{k=1}^n \binom{n-1}{k-1} p_{nkt} = \sum_{k=1}^n \binom{n-1}{k-1} \frac{N_t}{\binom{n}{k}} \mathbb{P}[X_1 = k | \sum_{i=1}^{N_t} X_i = n] \quad (7)$$

$$= \sum_{k=1}^n N_t \frac{k}{n} \mathbb{P}[X_1 = k | \sum_{i=1}^{N_t} X_i = n] \quad (8)$$

$$= \frac{N_t}{n} \sum_{k=0}^n k \mathbb{P}[X_1 = k | \sum_{i=1}^{N_t} X_i = n] \quad (\text{lower limit } k=0 \text{ makes no difference overall})$$

$$= \frac{N_t}{n} \mathbb{E}[X_1 | \sum_{i=1}^{N_t} X_i = n] \quad (9)$$

$$= p_{1,t}(N_t, n) = 1 \quad (\text{see below})$$

<sup>51</sup> Taken from stackexchange (link)

$$\begin{aligned} \mathbb{E}[X_1 | \sum_{i=1}^{N_t} X_i = n] &= \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{E}[X_i | \sum_{i=1}^{N_t} X_i = n] && (X_i \text{ i.i.d.}) \\ &= \frac{1}{N_t} \mathbb{E}[\sum_{i=1}^{N_t} X_i | \sum_{i=1}^{N_t} X_i = n] && (\text{Swap expectation, sum } (possibly \text{ questionable?})) \\ &= \frac{n}{N_t} && (10) \end{aligned}$$

### <sup>52</sup> 3 Poisson case

<sup>53</sup> Here the offspring distribution is  $\alpha_t = \text{Poisson}(R_t)$ . In this case, we have

$$\sum_{i=1}^{N_t} S_{t,i} \sim \text{Poisson}(N_t R_t), \quad (11)$$

<sup>54</sup> and conditional distribution

$$\mathbb{P}\left[S_{t,1} = s \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] = \frac{\mathbb{P}\left[S_{t,1} = s, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]}{\mathbb{P}\left[\sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]} \quad (12)$$

$$= \frac{\alpha_t(s) \mathbb{P}\left[\sum_{i=2}^{N_t} S_{t,i} = N_{t+1} - s\right]}{\mathbb{P}\left[\sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]} \quad (13)$$

$$= \frac{\frac{R_t^s e^{-R_t}}{s!} \cdot \frac{((N_t - 1)R_t)^{N_{t+1}-s}}{(N_{t+1} - s)!}}{\frac{(N_t R_t)^{N_{t+1}} e^{-N_t R_t}}{N_{t+1}!}} \quad (14)$$

$$= \binom{N_{t+1}}{s} \left(\frac{1}{N_t}\right)^s \left(1 - \frac{1}{N_t}\right)^{N_{t+1}-s}, \quad (15)$$

that is

$$S_{t,1} \left| \left( \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right) \right. \sim \text{Binomial} \left( N_{t+1}, \frac{1}{N_t} \right). \quad (16)$$

The inclusive probability of coalescence for  $k$  lines is given by (3) with the falling factorial moments

$$\mathbb{E} \left[ \binom{S_{t,1}}{k} \left| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right. \right] = \frac{1}{k!} \mathbb{E} \left[ \frac{S_{t,1}!}{(S_{t,1} - k)!} \left| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right. \right] = \frac{1}{k!} \frac{N_{t+1}!}{(N_{t+1} - k)!} \left( \frac{1}{N_t} \right)^k, \quad (17)$$

taken from Potts (1953).

Consequently, the inclusive probability of coalescence for two lines is

$$p_{2,t} = \frac{1}{N_t}, \quad (18)$$

and more generally the inclusive probability of coalescence for  $k$  lines is

$$p_{k,t} = \frac{1}{N_t^{k-1}}. \quad (19)$$

The inclusive probability of coalescence for two lines is analogous to the Wright-Fisher model, with each individual at generation  $t + 1$  selecting a parent uniformly at random from generation  $t$ .

The exclusive probability of coalescence for two lines from a sample of  $n$  ( $n \geq 2$ ) is

$$p_{n,2,t} = \frac{(N_t - 1)^{n-2}}{N_t^{n-1}}, \quad (20)$$

and more generally the exclusive probability of coalescence for  $k$  lines from a sample of  $n$  ( $n \geq k$ ) is

$$p_{n,k,t} = \frac{(N_t - 1)^{n-k}}{N_t^{n-1}}. \quad (21)$$

## 4 Negative-Binomial case

Here the offspring distribution is Negative-Binomial with mean  $R_t$  and variance  $V_t$ . The parameters of this distribution are  $r = R_t^2/(V_t - R_t)$  and  $p = R_t/V_t$ .

The inclusive probability of coalescence for two lines is:

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

68 The inclusive probability of coalescence for  $k$  lines is:

$$p_{k,t} = \prod_{i=1}^{k-1} \frac{r+i}{N_t r + i} = \frac{B(N_t r + 1, r + k)}{B(r + 1, N_t r + k)} \quad (23)$$

69 where  $B(x, y)$  denotes the Beta function defined as  $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$ .

70 The exclusive probability of coalescence for  $k$  lines is:

$$p_{nkt} = \frac{N_t B(k + r, n - k + N_t r - r)}{B(r, N_t r - r)} \quad (24)$$

## 71 5 Example

72 Let the offspring distribution have a mean of  $R_t = 2$ . In the Poisson case the offspring distribution  
 73 is Poisson(2). We consider NegBin cases with the same mean and varying dispersion parameters  $r$ .  
 74 When  $r$  is high the dispersion is low and the NegBin behaves almost like a Poisson. See Figure 1.

## 75 6 Lambda-coalescent

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

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

79 The Beta( $2 - \alpha, \alpha$ )-coalescent model (Schweinsberg 2003) has a single parameter  $\alpha \in [0, 2]$  and is  
 80 defined as:

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

81 from which we can deduce that:

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

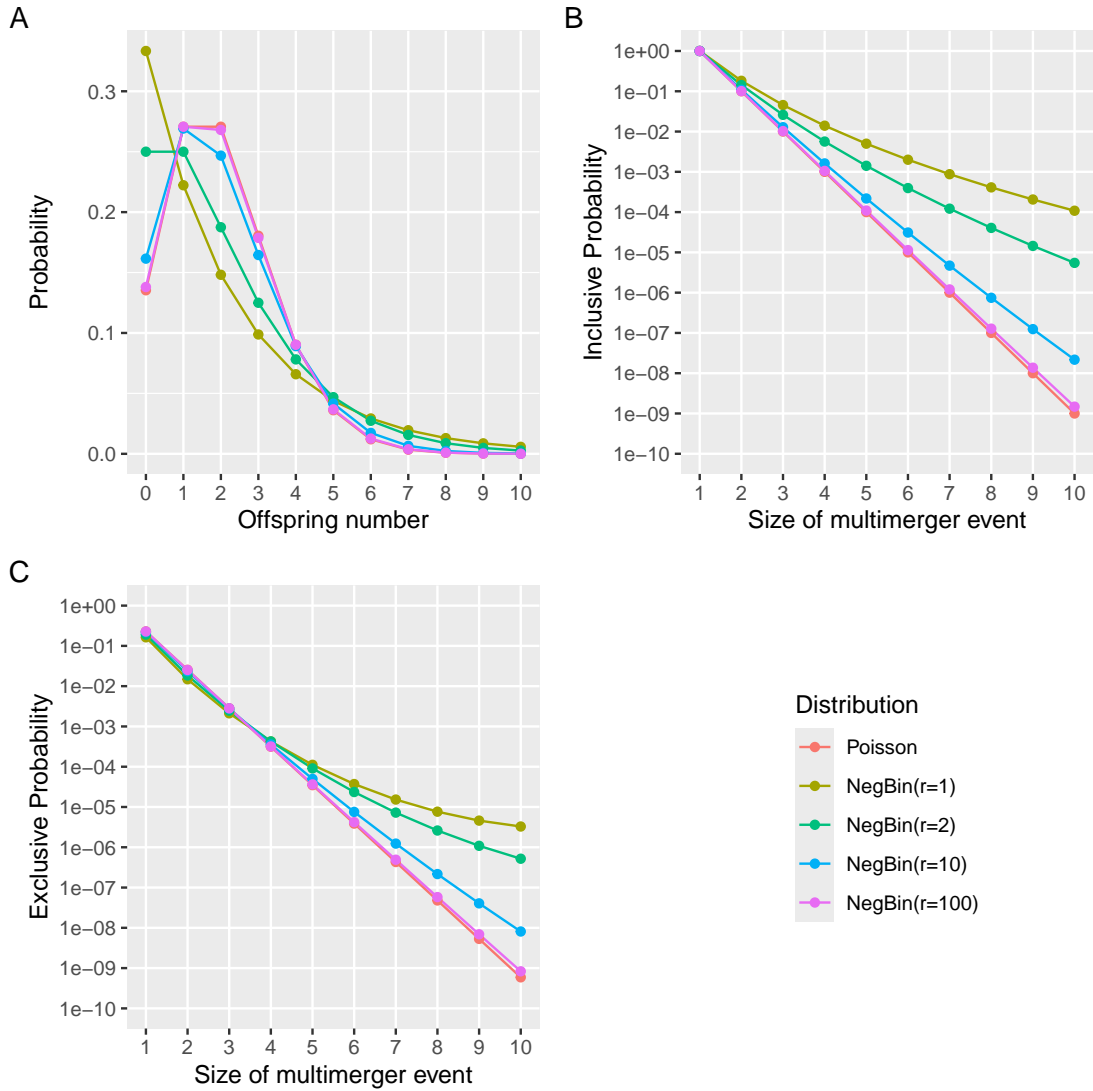


Figure 1: (A) Offspring distribution. (B) Inclusive probability of coalescence. (C) Exclusive probability of coalescence.

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

## 7 Implementation

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



88 repository.

## 89 **8 Discussion**

## 90 **Acknowledgements**

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

## References

- Donnelly, P., Kurtz, T.G., 1999. Particle Representations for Measure-Valued Population Models. *The Annals of Probability* 27. doi:10.1214/aop/1022677258.
- Fraser, C., Li, L.M., 2017. Coalescent models for populations with time-varying population sizes and arbitrary offspring distributions. *bioRxiv* , 10.1101/131730doi:10.1101/131730.
- Gómez-Carballa, A., Pardo-Seco, J., Bello, X., Martínón-Torres, F., Salas, A., 2021. Superspreading in the emergence of COVID-19 variants. *Trends in Genetics* 37, 1069–1080. doi:10.1016/j.tig.2021.09.003.
- Helekal, D., Koskela, J., Didelot, X., 2024. Inference of multiple mergers while dating a pathogen phylogeny. *bioRxiv* , 2023.09.12.557403doi:10.1101/2023.09.12.557403.
- Hoscheit, P., Pybus, O.G., 2019. The multifurcating skyline plot. *Virus Evolution* 5, 1–10. doi:10.1093/ve/vez031.
- Kingman, J., 1982a. The coalescent. *Stochastic Processes and their Applications* 13, 235–248. doi:10.1016/0304-4149(82)90011-4.
- Kingman, J.F.C., 1982b. On the genealogy of large populations. *Journal of Applied Probability* 19, 27–43. doi:10.2307/3213548.
- Lemieux, J.E., Siddle, K.J., Shaw, B.M., Loreth, C., Schaffner, S.F., Gladden-Young, A., Adams, G., Fink, T., Tomkins-Tinch, C.H., Krasilnikova, L.A., DeRuff, K.C., Rudy, M., Bauer, M.R., Lagerborg, K.A., Normandin, E., Chapman, S.B., Reilly, S.K., Anahtar, M.N., Lin, A.E., Carter, A., Myhrvold, C., Kembell, M.E., Chaluvadi, S., Cusick, C., Flowers, K., Neumann, A., Cerrato, F., Farhat, M., Slater, D., Harris, J.B., Branda, J.A., Hooper, D., Gaeta, J.M., Baggett, T.P., O’Connell, J., Gnirke, A., Lieberman, T.D., Philippakis, A., Burns, M., Brown, C.M., Luban, J., Ryan, E.T., Turbett, S.E., LaRocque, R.C., Hanage, W.P., Gallagher, G.R., Madoff, L.C., Smole, S., Pierce, V.M., Rosenberg, E., Sabeti, P.C., Park, D.J., MacInnis, B.L., 2021. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* 371, eabe3261. doi:10.1126/science.abe3261.
- Li, L.M., Grassly, N.C., Fraser, C., 2017. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Molecular Biology and Evolution* 34, 2982–2995. doi:10.1093/molbev/msx195.
- Lloyd-Smith, J., Schreiber, S., Kopp, P., Getz, W., 2005. Superspreading and the effect of individual variation on disease emergence. *Nature* 438, 355–9. doi:10.1038/nature04153.
- Menardo, F., Gagneux, S., Freund, F., 2021. Multiple Merger Genealogies in Outbreaks of *Mycobacterium tuberculosis*. *Molecular Biology and Evolution* 38, 290–306. doi:10.1093/molbev/msaa179.
- Pitman, J., 1999. Coalescents with multiple collisions. *The Annals of Probability* 27, 1870–1902.
- Potts, R.B., 1953. Note on the Factorial Moments of Standard Distributions. *Australian Journal of Physics* 6, 498–499. URL: <https://www.publish.csiro.au/ph/ph530498>, doi:10.1071/ph530498. publisher: CSIRO PUBLISHING.
- Sagitov, S., 1999. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* 36, 1116–1125. doi:10.1239/jap/1032374759.
- Schweinsberg, J., 2003. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications* 106, 107–139. doi:10.1016/S0304-4149(03)00028-0.

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