

Detecting and quantifying heterogeneity in susceptibility using contact tracing data

Beth M. Tuschhoff, David A. Kennedy

Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA

Abstract

The presence of heterogeneity in susceptibility, differences between hosts in their likelihood of becoming infected, can fundamentally alter disease dynamics and public health responses, for example, by changing the final epidemic size, the duration of an epidemic, and even the vaccination threshold required to achieve herd immunity. Yet, heterogeneity in susceptibility is notoriously difficult to detect and measure, especially early in an epidemic. Here we develop a method that can be used to detect and estimate heterogeneity in susceptibility given contact by using contact tracing data, which is typically collected early in the course of an outbreak. This approach provides the capability, given sufficient data, to estimate and account for the effects of this heterogeneity before they become apparent during an epidemic. It additionally provides the capability to analyze the wealth of contact tracing data available for previous epidemics and estimate heterogeneity in susceptibility for disease systems in which it has never been estimated previously. The premise of our approach is that highly susceptible individuals become infected more often than less susceptible individuals, and so individuals not infected after appearing in contact networks should be less susceptible than average. This change in susceptibility can be detected and quantified when individuals show up in a second contact network after not being infected in the first. To develop our method, we simulated contact tracing data from artificial populations with known levels of heterogeneity in susceptibility according to underlying discrete or continuous distributions of susceptibilities. We analyzed this data to determine the parameter space under which we are able to detect heterogeneity and the accuracy with which we are able to estimate it. We found that our power to detect heterogeneity increases with larger sample sizes, greater heterogeneity, and intermediate fractions of contacts becoming infected in the discrete case or greater fractions of contacts becoming infected in the continuous case. We also found that we are able to reliably estimate heterogeneity and disease dynamics. Ultimately, this means that contact tracing data alone is sufficient to detect and quantify heterogeneity in susceptibility.

1. Introduction

At the outset of an epidemic, public health responses depend on estimates of the final epidemic size, the peak number of cases, the timing of the peak, and the herd immunity threshold. Compartmental models such as the susceptible-infected-recovered (SIR) model are commonly used to model infectious disease dynamics and predict outcomes, but there are limitations to this approach (Keeling and Danon, 2009; Roberts et al., 2015; Tolles and Luong, 2020; Dhar, 2020). Namely, SIR models tend to oversimplify the complexity of disease dynamics, resulting in discrepancies between the model predictions and epidemic data (Keeling and Danon, 2009). One of the simplifying assumptions of the standard SIR model is that all host individuals are the same. However, this is often false: individuals can be heterogeneous in many ways (Woolhouse et al., 1997; VanderWaal and Ezenwa, 2016) including with regard to their likelihood of becoming infected, hereafter referred to as heterogeneity in susceptibility (Dwyer et al., 1997).

Heterogeneity in susceptibility can have a large impact on infectious disease dynamics (Dwyer et al., 1997; Gomes et al., 2014; Langwig et al., 2017; Gomes et al., 2022). Increased amounts of heterogeneity in susceptibility result in a lower peak number of cases, different timing of the peak, smaller final epidemic size, and lower herd immunity thresholds (Aguas et al., 2020; Gomes et al., 2022; Montalbán et al., 2022). As a result, disease control programs (Anderson and May, 1984) and epidemiological models (Dwyer et al., 1997; Langwig et al., 2017; King et al., 2018; Gomes et al., 2019) may need to account for heterogeneity in

45 susceptibility if they are to be optimally useful. Accurate early predictions of disease dynamics could give
46 policy makers critical information to make decisions, but heterogeneity in susceptibility is notoriously difficult
47 to measure (Elder et al., 2008). Moreover, the effects of heterogeneity in susceptibility are typically small
48 during the earliest phases of epidemics and only become apparent later, making it even more challenging to
49 estimate heterogeneity in susceptibility in real time and account for its effects. It would therefore be useful to
50 develop new methods for quantifying the degree of heterogeneity in host susceptibility early in epidemics.

51 Existing methods to quantify heterogeneity in susceptibility are not adequate for estimation in real time
52 because they rely on using data that is either collected later in epidemics or that typically cannot be collected
53 due to ethical or logistical constraints. Dwyer et al. (1997), Ben-Ami et al. (2010), and Langwig et al. (2017)
54 used laboratory dose-response and field transmission experiments to estimate heterogeneity in susceptibility,
55 but these experimental methods are not feasible for application in real time or for human epidemics in
56 general due to time constraints and ethical concerns. Gomes et al. (2019) compared disease incidence across
57 municipalities in several countries to construct Lorenz curves and fit susceptibility risk distributions, but
58 this method requires a substantial amount of data that would not be available early in an epidemic. Smith
59 et al. (2005) and Corder et al. (2020) used morbidity data to fit models and estimate heterogeneity, but this
60 method cannot be used until later in an epidemic when there is sufficient data to fit curves. Gomes et al.
61 (2022) also used curve fitting with mortality data that could be implemented once at least four months of
62 data were available, but their method is heavily dependent on the underlying model and assumptions. With
63 the recent increased interest in real-time estimation, Anderson et al. (2023) developed a method to estimate
64 within-household heterogeneity in susceptibility, but this is not the same as the population-level heterogeneity
65 that drives population-level disease dynamics. Here we develop a novel method to identify and quantify
66 host heterogeneity in susceptibility using contact tracing data, which can be collected early in an epidemic.
67 Contact tracing is often performed to mitigate the spread of pathogens that are otherwise difficult to control
68 (Eames and Keeling, 2003; Hossain et al., 2022), and therefore, our method should not require the collection
69 of any data beyond that which would already be collected for other purposes.

70 Contact tracing typically takes one of two forms: forward and backward. Forward contact tracing attempts
71 to find all the contacts of an infected person to whom the disease could transmit. This is done by identifying
72 infected individuals and all their known contacts. The contacts are then quarantined and monitored for
73 disease. For any contact that is infected, the process is repeated with their contacts. Backward contact
74 tracing attempts to identify the contact of an infected person from whom the disease transmitted. In practice,
75 both methods can be employed simultaneously in an effort to maximize the effectiveness of contact tracing
76 efforts (Bradshaw et al., 2021), and the data on infected individuals and their contacts are typically recorded.
77 When done thoroughly, contact tracing data provide information about the infection status of individuals
78 that have been in contact with an infected individual. As we will explain, when contact tracing data tracks
79 specific individuals through multiple exposure events, it can be used to quantify heterogeneity in susceptibility
80 given contact through the method that we develop here.

81 Our method uses the fact that average susceptibility decreases over time in a population with heterogeneity
82 in susceptibility (Fig 1). This is because individuals with high susceptibility are more likely to be infected than
83 individuals with low susceptibility for a given exposure level. Individuals that show up in a second contact
84 tracing network, after not being infected in the first, should therefore have a lower risk of infection than
85 individuals that show up in a network for the first time. In the rest of the paper, we establish our method and
86 analyze its effectiveness for two cases: a population with two discrete susceptibility levels and a population
87 with continuous variation in susceptibility. Notably, the selection of these two cases is arbitrary, and our
88 method is flexible enough that it could be employed for any distribution of heterogeneity in susceptibility.

89 2. Methods and Results

90 Our method to detect and quantify heterogeneity in susceptibility exploits the change in average suscep-
91 tibility over multiple exposure events that would be expected to occur if a population had heterogeneity
92 in susceptibility (Fig 1). Given contact with an infectious individual, individuals with high susceptibility
93 are more likely to be infected than those with low susceptibility. This creates a selection process in which
94 susceptibility should on average decline in a heterogeneous host population following each exposure event.
95 This change in average susceptibility provides a way to identify and estimate the level of heterogeneity early

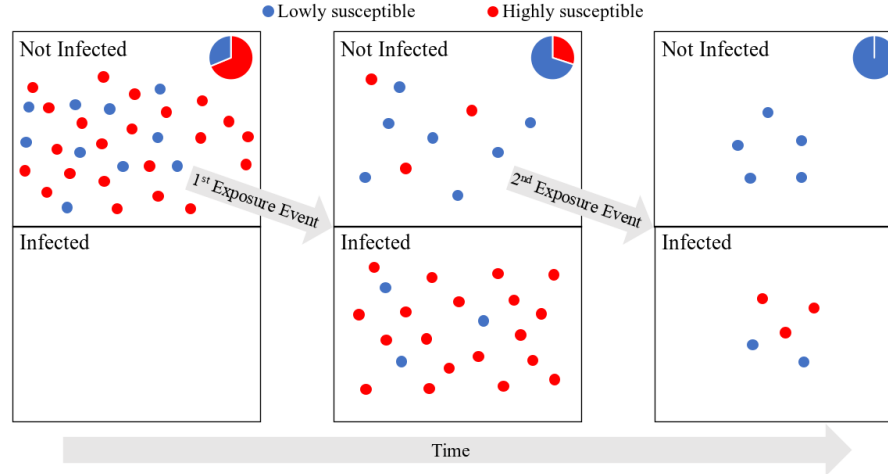


Figure 1: Average susceptibility decreases over exposure events in a heterogeneous population. The figure depicts individuals infected and not infected over two exposure events in a heterogeneous population with more susceptible (red) and less susceptible (blue) individuals. The pie charts show the composition of the not infected population. Average susceptibility in the not infected population decreases after each exposure event as the more susceptible individuals are primarily infected. Note that if the population lacked heterogeneity in susceptibility, all individuals would be either red or blue, and thus, susceptibility would not change.

96 in an epidemic despite the seemingly small effects of heterogeneity at the beginning of epidemics. Notably, no
97 change in average susceptibility should occur in a population that lacks heterogeneity in susceptibility.

98 This method employs contact tracing data. With contact tracing data, there are multiple contact networks
99 that are each composed of an infected individual and the known contacts the infected individual had during
100 their infectious period. This means each contact network is a set of exposure events where contacts are
101 exposed to a pathogen and have a chance of being infected. In order for our method to work, there must
102 be individuals that show up in at least two separate contact networks such that they are exposed but not
103 infected in the first of these networks. At the start of the second exposure event, these individuals would
104 have been previously exposed but not infected (henceforth called focal individuals). This contact network
105 must also contain naive contacts: individuals that have not been previously exposed to the pathogen. The
106 basis of our method is to compare the fraction of naive individuals and focal individuals that become infected
107 in the second contact network; if there is no heterogeneity in susceptibility, focal individuals should have the
108 same susceptibility as naive individuals, whereas if there is heterogeneity in susceptibility, focal individuals
109 should on average be less susceptible than naive individuals. This difference in susceptibility arises due to the
110 selection process for infection of more susceptible individuals (Fig 1).

111 To compute the number of naive and focal individuals infected, there must be data on which specific
112 individuals are infected and which individuals are showing up in a contact network for a second time, which
113 would be available for example if individuals were identifiable between contact networks. There must also
114 be a sufficient sample size to detect heterogeneity in susceptibility. Here we explore the effects of sample
115 size, level of heterogeneity, and infection probability on our ability to detect and quantify heterogeneity in
116 susceptibility.

117 We apply this method to two underlying models describing the distribution of individuals' susceptibilities.
118 In one underlying model (discrete case), it is assumed that the population is composed of two host types
119 where each host type has a different susceptibility or probability of being infected given contact. Discrete
120 susceptibility types might be expected when heterogeneity in susceptibility is predominantly accounted for by
121 a small number of factors that create groups in the population with distinct susceptibilities. For example,
122 genetic polymorphisms could be selected for that increase resistance to a pathogen, resulting in populations
123 containing a mixture of individuals with and without the mutation such as was seen for HIV (Huang et al.,
124 1996). Likewise, prior exposure, whether natural or vaccine-induced, to a pathogen or related pathogen

125 could create more resistant subpopulations such as with milkmaids not developing smallpox after contracting
126 cowpox (Barquet and Domingo, 1997). Behaviors like handwashing and mask wearing (Larson, 1999; Van der
127 Sande et al., 2008) or host nutritional status (Chandra, 1979) could also produce approximately binary
128 outcomes for susceptibility to infection.

129 In the other underlying model (continuous case), it is assumed that the population is composed of hosts
130 with a continuous range of susceptibilities such that each host’s probability of being infected given contact
131 is unique. This situation might be expected when there is a complex combination of factors dictating
132 heterogeneity in susceptibility or when the cause of heterogeneity is a trait that continuously varies across
133 individuals. For instance, variability in gene expression, which could be affected by epigenetics, copy number
134 variations, and sequence polymorphisms, is associated with disease susceptibility (Li et al., 2010). In addition,
135 some of the factors that lead to discrete variation in susceptibility could also have a continuous effect such
136 as the degree of cleanliness achieved by handwashing (Larson, 1999) or continuous variation in nutrients.
137 Beyond a complex combination of factors, there could also be situations where a continuously varying trait
138 like body mass (Dobner and Kaser, 2018) or the level of antibodies induced in an immune response (Plotkin,
139 2008) explains the heterogeneity in susceptibility in the population.

140 2.1. Methods

141 Our method is comprised of two parts: detecting heterogeneity in susceptibility and quantifying it if
142 present. The former is a hypothesis testing problem, and the latter is a parameter estimation problem. For
143 the detection of heterogeneity, we test the hypothesis that there is heterogeneity in susceptibility against the
144 null hypothesis that there is homogeneity in susceptibility.

145 2.1.1. Detection of heterogeneity in susceptibility

146 We consider F contact networks that each contain $N_i - 1$ naive individuals and one focal individual
147 where i is the set of contact networks. For simplicity, we assume N_i are equal for all i and thus drop the
148 subscript. Note that this assumption can be easily relaxed. We therefore have a total of $F(N - 1)$ naive
149 individuals and F focal individuals. We first compute the fractions of naive, focal, and total individuals
150 infected. The fractions of naive and focal individuals infected are estimates for the probability of a naive
151 or focal individual being infected (p_n and p_f respectively). The fraction of total individuals infected is an
152 estimate for the average probability of being infected (\bar{p}). We then calculate the log-likelihood of the data
153 (numbers of individuals infected) under each hypothesis as a sum of the log-likelihoods for the number of
154 each type of individual infected where

$$L_{\text{hom}} = \ln [P(x_n | F(N - 1), \bar{p})] + \ln [P(x_f | F, \bar{p})] \quad (1)$$

$$L_{\text{het}} = \ln [P(x_n | F(N - 1), p_n)] + \ln [P(x_f | F, p_f)]. \quad (2)$$

155 L_{hom} is the log-likelihood of the data under the null hypothesis that there is homogeneity in susceptibility,
156 so we assume all individuals have the same probability of being infected, regardless of their number of
157 exposures to the pathogen ($p_n = p_f = \bar{p}$). L_{het} is the log-likelihood under the alternative hypothesis that
158 there is heterogeneity in susceptibility, so we assume naive and focal individuals have different probabilities
159 of being infected due to the infection selection process that occurs when heterogeneity is present ($p_n \neq p_f$).
160 These log-likelihoods are calculated identically regardless of whether the heterogeneity is discrete or continuous.
161 $P(x|n, p)$ is the probability of observing x individuals infected out of n individuals exposed with probability p of
162 being infected and is distributed according to a binomial distribution. The number of naive individuals infected
163 has distribution $\text{Binom}(n = F(N - 1), p_n)$, and the number of focal individuals infected has distribution
164 $\text{Binom}(n = F, p_f)$. x_n and x_f are the numbers of naive and focal individuals infected respectively where
165 $x_n \in [0, F(N - 1)]$ and $x_f \in [0, F]$. p_n , p_f and \bar{p} are estimated from the data as $p_n = \frac{x_n}{F(N-1)}$, $p_f = \frac{x_f}{F}$, and
166 $\bar{p} = \frac{x_f + x_n}{FN}$. The log-likelihoods of the data under each hypothesis were compared using a likelihood ratio test
167 with one degree of freedom and significance level $\alpha = 0.05$.

168 Here, we simulated data to test our method. To do so, we first set parameters dictating the sample size
169 and heterogeneity present in the population. Then, we simulated initial exposure events with N individuals

170 in each network and kept uninfected individuals as our focal individuals. For each focal individual, we then
 171 simulated a second exposure event with that focal individual and $N - 1$ naive individuals. The susceptibilities
 172 of the naive individuals were drawn randomly from the same heterogeneity distribution set for the starting
 173 population. We recorded the fraction of each type of individual (i.e. focal or naive) infected in the second
 174 exposure events and calculated the log-likelihood of the simulated data under our two hypotheses. Then, we
 175 compared the hypotheses with a likelihood ratio test. We ran 1,000 simulations for each set of parameters to
 176 determine our statistical power to detect heterogeneity in susceptibility with that parameter combination.
 177 All simulations and data analysis were performed in R version 4.0.3 (R Core Team, 2020).

178 For the discrete case, we simulated data using two types of individuals (denoted A and B), but we note
 179 that the aforementioned factors could potentially be combined to result in more than two distinct groupings,
 180 and similar methods could be applied for these situations. At the beginning of each simulation, we set the
 181 probability of being infected for each type of individual, p_A and p_B , where $p_A \in [0, 1]$ and $p_B \in [0, p_A]$. We
 182 also set the fraction of the starting population that is type A (f_A) where $f_A \in [0, 1]$. All three parameters p_A ,
 183 p_B , and f_A affect the level of heterogeneity in susceptibility in the population.

184 We later calculated the coefficient of variation of the risk of being infected for this discrete case (C_d) and
 185 the expected fraction of naive individuals infected (E_d) from p_A , p_B , and f_A to better summarize the results.
 186 The risks of being infected for type A and B individuals, r_A and r_B respectively, are shown below. These
 187 equations are derived from the formula for the probability of being infected $p_i = 1 - e^{-r_i}$, $i = A, B$.

$$r_A = -\ln(1 - p_A) \quad (3)$$

$$r_B = -\ln(1 - p_B) \quad (4)$$

188 The coefficient of variation is defined as the standard deviation divided by the mean. Hence, C_d is the
 189 standard deviation of risk divided by the mean risk (Supplementary information S1) and is given by

$$C_d = \frac{(r_A - r_B)\sqrt{f_A(1 - f_A)}}{r_A f_A + r_B(1 - f_A)} \quad (5)$$

E_d is the same as the mean probability of being infected \bar{p} , which is given by

$$E_d = \bar{p} = p_A f_A + p_B(1 - f_A) \quad (6)$$

190 We additionally defined the sample size for the simulation by setting the number of individuals in each
 191 exposure group N and the number of focal individuals F . For our simulations, we used $N = 5$ and $F = 50$ or
 192 200.

193 For the continuous case, in contrast to the discrete case just discussed, each individual in the population
 194 has a different risk of being infected. Here, we assume that individuals' risks for being infected follow
 195 a gamma distribution, but as in the discrete case, other distributions could be used. We chose to use a
 196 gamma distribution for illustration purposes because it is flexible and has been used to model heterogeneous
 197 populations previously (Dwyer et al., 1997; Langwig et al., 2017).

198 At the beginning of each simulation, we set the parameters k and θ , respectively the shape and scale
 199 of the gamma distribution, that dictate the risk distribution where $k, \theta > 0$. For ease of interpretation, we
 200 present our results with respect to the coefficient of variation of risk for continuous variation C_c and expected
 201 fraction of naive individuals infected E_c . As in the discrete case, the risk r_i for the i th individual being
 202 infected is related to the probability of being infected such that $p_i = 1 - e^{-r_i}$ and thus

$$r_i = -\ln(1 - p_i). \quad (7)$$

203 As it is gamma distributed, the risk distribution has standard deviation $\sigma = \theta\sqrt{k}$ and mean $\mu = k\theta$. So,
 204 C_c can be simplified to

$$C_c = \frac{1}{\sqrt{k}} \quad (8)$$

205 E_c is the same as the mean probability of being infected \bar{p} and is derived in Dwyer et al. (1997) as

$$E_c = \bar{p} = 1 - \frac{S_t}{S_0} = 1 - (1 + \theta)^{-k} \quad (9)$$

206 where S_0 and S_t are the number of susceptible individuals at the beginning and end of an exposure round
207 respectively.

208 We additionally defined the sample size for the simulation by setting the number of individuals in each
209 exposure group N and the number of focal individuals F . As in the discrete case, we use $N = 5$ and $F = 50$
210 or 200.

211 We tested the ability of our method to detect heterogeneity in susceptibility for each potential combination
212 of $f_A, F, C_d \in [0, 3]$ with step size 0.02, and $E_d \in [0.02, 0.98]$ with step size 0.02 in the discrete case and
213 $F, C_c \in [0, 3]$ with step size 0.02, and $E_c \in [0.02, 0.98]$ with step size 0.02 in the continuous case. This was
214 done for 1,000 simulations to compute the statistical power of the method. We did not simulate $E_d = 0, 1$ or
215 $E_c = 0, 1$ because such values preclude heterogeneity in susceptibility. We examined $C_d, C_c \in [0, 3]$ because
216 this captures most of the range of published values for the coefficient of variation of risk we could find: 0.0007
217 to 3.33 (Dwyer et al., 1997, 2000; Smith et al., 2005; Ben-Ami et al., 2008; Elderd et al., 2008; Ben-Ami et al.,
218 2010; Pessoa et al., 2014; Langwig et al., 2017; King et al., 2018; Gomes et al., 2019; Corder et al., 2020;
219 Gomes et al., 2022).

2020 2.1.2. Quantification of heterogeneity in susceptibility

221 Given the detection of heterogeneity in susceptibility, the next question is whether that heterogeneity will
222 substantially impact disease dynamics. To determine whether it will, we need to ask whether contact tracing
223 data is sufficient to estimate the parameters of SIR models that include heterogeneity in susceptibility and
224 whether those parameter estimates accurately capture disease dynamics. To do so, we fit the parameters of
225 our underlying risk distributions using simulated contact tracing data as above. Parameter values used to
226 simulate the contact tracing data for the discrete and continuous heterogeneity cases are provided in Table 1.

227 We generated posterior distributions for both models using Metropolis-Hastings MCMC. In the discrete
228 case, our MCMC chain had length 30,000,000 with a burn-in of 15,000,000 and thinning interval 1,500. For
229 all three parameters, we used flat priors and uniform proposal distributions. Our proposal distributions
230 were $p_A \sim \text{Unif}(0, 1)$, $p_B \sim \text{Unif}(0, p_A)$, and $f_A \sim \text{Unif}(0, 1)$. There is not a simple, analytic likelihood
231 function for the likelihood of the data given a proposed parameter set, so the likelihood was estimated by
232 simulation with Approximate Bayesian Computation (ABC), where the likelihood estimate was determined
233 by comparing the fraction of simulations that provided results that were within a pre-specified error tolerance
234 of the actual data (Beaumont et al., 2002). To do so, we ran 100 simulations of the number of focal and naive
235 individuals infected across F contact networks for a proposed parameter set. We then calculated the fraction
236 of simulations where the number of individuals infected was within a 1% error tolerance of the number
237 infected in the true data. Note that our results are fairly insensitive to this error tolerance (Supplementary
238 information S4). This simulation was done separately for focal and naive individuals. We then computed
239 the overall log-likelihood as a sum of the logs of those fractions. We assessed convergence of the chains by
240 visually inspecting the resulting trace plots and marginal posterior distributions for each parameter. In
241 the continuous case, our MCMC chain had length 600,000 with a burn-in of 200,000 and thinning interval
242 100. We used an exponential prior $\text{Exp}(2)$ for k because known values of C_c suggest that k is likely to be
243 small (Dwyer et al., 1997, 2000; Smith et al., 2005; Ben-Ami et al., 2008; Elderd et al., 2008; Ben-Ami et al.,
244 2010; Pessoa et al., 2014; Langwig et al., 2017; King et al., 2018; Gomes et al., 2019; Corder et al., 2020;
245 Gomes et al., 2022). We used a flat prior for θ for all values $[0, \text{inf})$ and a multivariate lognormal proposal
246 distribution $(k, \theta) \sim \text{MLogNorm}(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.01 & -0.008 \\ -0.008 & 0.05 \end{pmatrix})$. We assessed convergence of the chains by
247 visually inspecting the resulting trace plots and marginal posterior distributions for each parameter (Kennedy
248 et al., 2015).

249 We then used these parameter estimates to generate SIR dynamics. Notably, the system of differential
250 equations describing the discrete and continuous cases differ. For the discrete case, we implemented the
251 following system of ordinary differential equations:

$$\frac{dS_A}{dt} = -\beta_A S_A I \quad (10)$$

$$\frac{dS_B}{dt} = -\beta_B S_B I \quad (11)$$

$$\frac{dI}{dt} = (\beta_A S_A + \beta_B S_B) I - \gamma I \quad (12)$$

252 S_A and S_B are the susceptible individuals of types A and B , and I is the infected individuals where I
 253 includes infected A and infected B individuals such that $I = I_A + I_B$. At the start of each SIR simulation, we
 254 determine the fraction of the population to allocate as A and B from f_A . We also set the basic reproduction
 255 number $R_{0,d} = \frac{\bar{\beta}(S_0+I_0)}{\gamma} = \frac{(p_A f_A + p_B(1-f_A))c(S_0+I_0)}{\gamma}$ at an assumed “true” value where $\bar{\beta}$ is the average
 256 transmission rate and $S_0 + I_0$ is the population size. $R_{0,d}$ is often a reasonably well approximated value,
 257 and it does not change with heterogeneity in susceptibility as initial average susceptibility remains the same
 258 regardless of heterogeneity (Hébert-Dufresne et al., 2020; Shaw and Kennedy, 2021). β_A and β_B are the
 259 transmission rates for types A and B respectively and were calculated as $\beta_A = p_A c$ and $\beta_B = p_B c$ where c is
 260 the contact rate. Note that c was calculated from $R_{0,d}$. γ is the recovery rate and was kept constant between
 261 the types of individuals at an assumed “true” value.

262 For the continuous case, we implemented the following system of ordinary differential equations derived in
 263 Elder et al. (2008):

$$\frac{dS}{dt} = -\beta SI \left(\frac{S}{S_0} \right)^{C_c^2} \quad (13)$$

$$\frac{dI}{dt} = \beta SI \left(\frac{S}{S_0} \right)^{C_c^2} - \gamma I \quad (14)$$

264 S is the number of susceptible individuals where S_0 is the number of susceptible individuals at the
 265 beginning of the simulation, and I is the number of infected individuals. At the start of each simulation,
 266 we set the basic reproduction number $R_{0,c} = \frac{\bar{p}c(S_0+I_0)}{\gamma}$ at an assumed “true” value where \bar{p} is the average
 267 probability of being infected, c is the contact rate, $S_0 + I_0$ is the population size, and γ is the recovery rate.
 268 \bar{p} is computed from the sampled parameters as $\bar{p} = 1 - (1 + \theta)^{-k}$, c was calculated from $R_{0,c}$, and γ was
 269 fixed at an assumed “true” value. β is the transmission rate and was calculated as $\beta = \bar{p}c$.

270 For each case, we randomly sampled 1,000 parameter sets from the posterior distribution to run SIR
 271 model simulations, and we compared this to the dynamics generated by the “true” parameter set used to
 272 generate our contact tracing data. Using these simulations, we determined 95% central credible intervals for
 273 the SIR dynamics for each model by finding the 2.5% and 97.5% percentiles of the 1,000 simulated dynamics
 274 at each time point over the epidemic. For our SIR simulations, we set $R_{0,d} = R_{0,c} = 3$, $S_0 = 20,000$, $I_0 = 10$,
 275 and $\gamma = 0.1$.

Table 1: The 95% CIs, medians, and true values for parameters estimated from MCMC in the discrete and continuous cases with $F = 1000$ and $N = 5$.

	Parameter	95% CI	Median	True
Discrete case	p_A	[0.437,0.958]	0.599	0.748
	p_B	[0.005,0.172]	0.085	0.125
	f_A	[0.102,0.543]	0.321	0.2
	C_d	[0.842,1.845]	1.093	1.3
	E_d	[0.236,0.263]	0.249	0.25
Continuous case	k	[0.364,1.024]	0.584	0.592
	θ	[0.321,1.257]	0.647	0.626
	C_c	[0.988,1.657]	1.309	1.3
	E_c	[0.237,0.269]	0.252	0.25

276 2.2. Results

277 2.2.1. Detection of heterogeneity in susceptibility

278 Figures 2 and 3 illustrate that the sample size, level of heterogeneity, and fraction of individuals infected
 279 affect our power to detect heterogeneity in susceptibility. This is because these factors ultimately affect the
 280 likelihoods used to test for heterogeneity in terms of the difference between the probabilities of infection for
 281 naive and focal individuals (p_n and p_f) and the variability in the likelihood ratio test statistic (Supplementary
 282 information S3). More precisely, these figures show that as the number of focal individuals F increases
 283 from 50 to 200, there is greater power to detect lower levels of heterogeneity (lower values of C_d , C_c). This

284 additionally allows for greater power across a wider range of E_d and E_c . This was to be expected because
 285 higher sample sizes, particularly of the previously exposed, focal individuals, decreases variability in our
 286 estimates of p_n , p_f , and \bar{p} . Notably, changing the total number of hosts in each contact network N had very
 287 little effect on our results (Supplementary information S2).

288 The level of heterogeneity in susceptibility present is described by the coefficient of variation of the risk
 289 distribution C_d or C_c . As C_d and C_c increase, there is more power to detect heterogeneity in susceptibility as
 290 there is more heterogeneity in the population. In the discrete case, for a given C_d , there is also more power
 291 to detect heterogeneity as f_A approaches 0.5. This is because as f_A approaches 0.5, the population is more
 292 evenly split between the two types of individuals, allowing for a greater difference between p_A and p_B and,
 293 therefore, p_n and p_f .

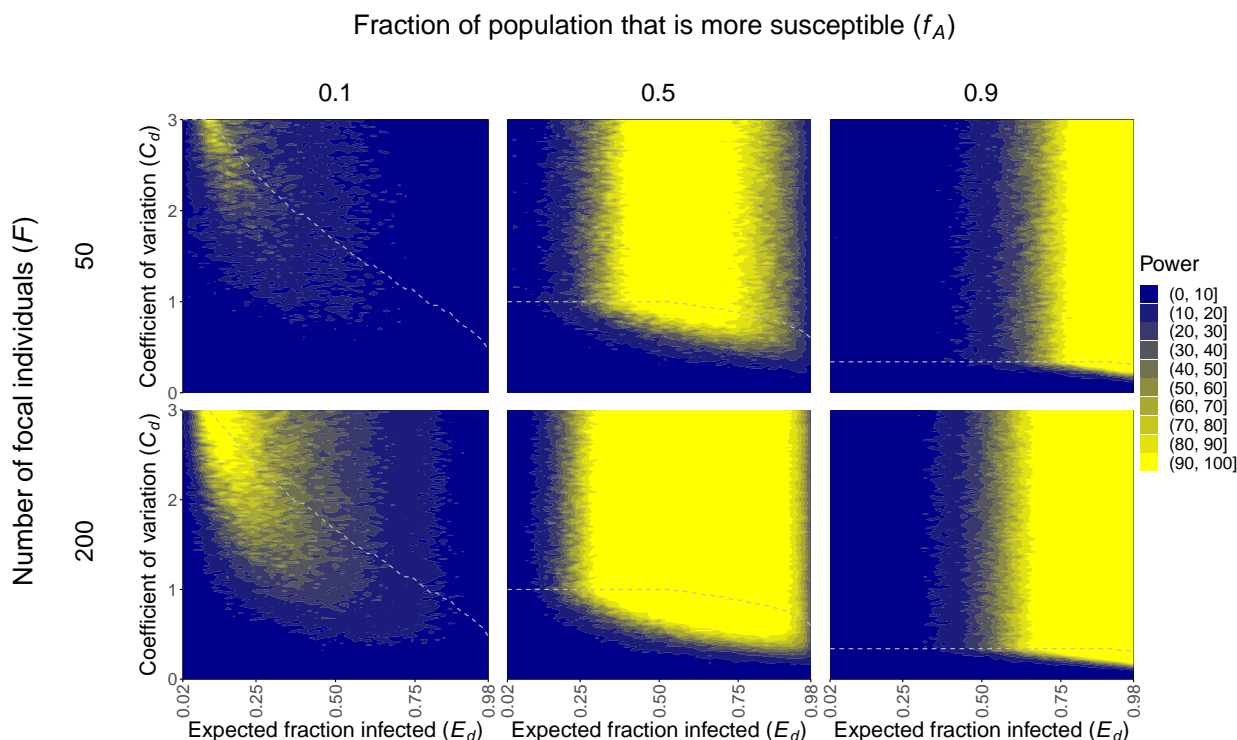


Figure 2: Increased heterogeneity in susceptibility (larger C_d and $f_A \rightarrow 0.5$), intermediate fractions of individuals infected (intermediate E_d), and increased sample sizes (larger F) enhance our power to detect heterogeneity in susceptibility in the discrete case. The plots show the power to detect heterogeneity in susceptibility in the discrete case across different numbers of focal individuals F and fraction of the population that is type A and more susceptible f_A . The areas above the gray dashed lines represent parameter space that gives computationally indistinguishable probabilities of infection p_A and p_B , and therefore power, to the parameter combination with the same E_d and highest C_d below the line. This occurs because risks of infection can be changed to increase C_d without bound, whereas probabilities are bounded. $N = 5$.

294 Lastly, the impact of the expected fraction of naive individuals infected (E_d , E_c) on power differs between
 295 the two underlying models. There is greater power to detect heterogeneity when an intermediate fraction
 296 of individuals is infected in the discrete case and when a greater fraction of individuals is infected in the
 297 continuous case. In the discrete case, E_d is determined by p_A , p_B , and f_A as per equation 6. The only way
 298 to have a large fraction of individuals infected is if both p_A and p_B are large. Hence, when E_d is high, p_A
 299 and p_B must both be close to 1. For similar reasons, when E_d is low, p_A and p_B must both be close to 0.
 300 Even though the risks r_A and r_B associated with these values may have varying levels of heterogeneity, the
 301 individuals themselves will have very similar infection outcomes, making it difficult to detect heterogeneity
 302 in susceptibility. Therefore, heterogeneity in susceptibility is better detected when an intermediate fraction
 303 of individuals is infected in the discrete case. In contrast, power increases in the continuous case with

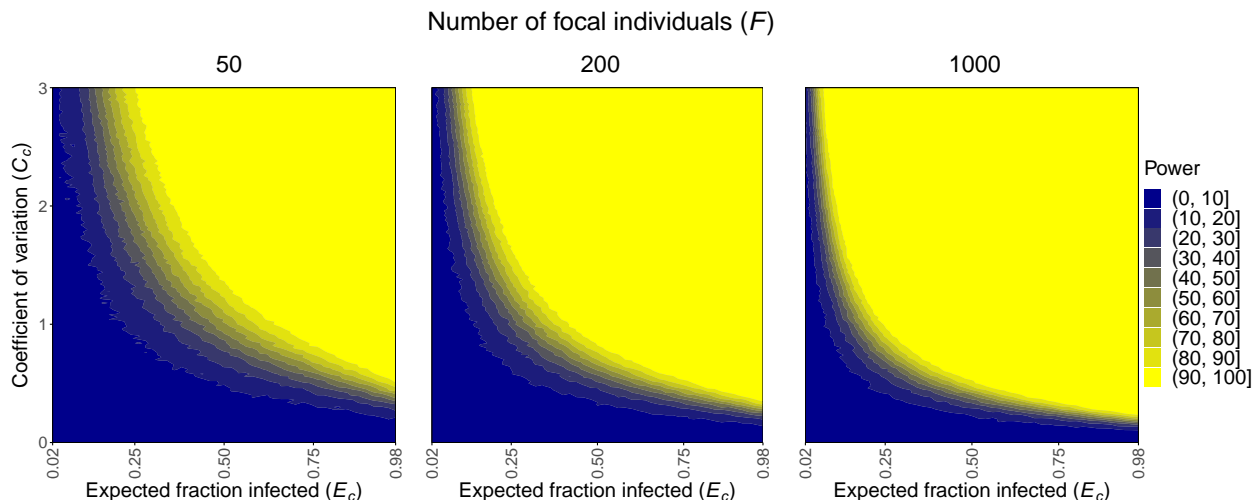


Figure 3: Increased heterogeneity in susceptibility (larger C_c), greater fractions of individuals infected (larger E_c), and increased sample sizes (larger F) enhance our power to detect heterogeneity in susceptibility in the continuous case. The plots show the power to detect heterogeneity in susceptibility in the continuous case across different numbers of focal individuals F . $N = 5$.

304 greater fractions of individuals infected (larger values of E_c). This is because there is more selection for
 305 who is infected as more individuals are infected, so the average population susceptibility will decrease more
 306 drastically, making it easier to detect heterogeneity in susceptibility.

307 2.2.2. Quantification of heterogeneity in susceptibility

308 We then explored the method's ability to estimate model parameters as well as predict the associated SIR
 309 dynamics. We perform this analysis for a particular parameter combination that leads to $C_d = C_c = 1.3$ and
 310 $E_d = E_c = 0.25$. These values were chosen because they represent a biologically realistic scenario based on
 311 previous literature (Dwyer et al., 1997, 2000; De Serres et al., 2000; Rieder, 2003; Smith et al., 2005; Taylor
 312 et al., 2007; Ben-Ami et al., 2008; Elder et al., 2008; Lessler et al., 2009; Ben-Ami et al., 2010; Pessoa et al.,
 313 2014; Ajelli et al., 2015; Langwig et al., 2017; King et al., 2018; Gomes et al., 2019; Corder et al., 2020; Koh
 314 et al., 2020; Gomes et al., 2022). In the discrete case, we used C_d and E_d and set $f_A = 0.2$ to calculate the
 315 true values $p_A = 0.748$ and $p_B = 0.125$. In the continuous case, we used C_c and E_c to calculate the true
 316 values $k = 0.592$ and $\theta = 0.626$.

317 We determined our 95% CIs for parameter estimation of the underlying parameters with $F = 1000$ and
 318 $N = 5$ to be those shown in Table 1. Note that the true values for p_A , p_B , and f_A as well as for k and θ
 319 are captured by these intervals. Admittedly, these parameter estimates are somewhat broad. Upon further
 320 investigation, we found the broad intervals to be due to high correlation in our parameter estimates, indicating
 321 low identifiability (Fig 4, 5). However, acceptable estimates do not span the entire ranges of the parameters
 322 and encapsulate the true parameters, so there is some information about their values in the data. As we will
 323 discuss, this partial identifiability does not hinder us from making precise predictions about the impact of the
 324 heterogeneity in susceptibility on the disease dynamics.

325 Using equations 5, 6, 8, and 9, we calculated and plotted the posterior distributions for C_d and E_d and C_c
 326 and E_c (Fig 6). With $F = 1000$ and $N = 5$, we determined the 95% CIs to be those shown in Table 1, which
 327 capture the true values. In the discrete case, the range of potential estimates for C_d is somewhat broad, but
 328 there is a strong ability to accurately and precisely estimate E_d . However, in the continuous case, there is a
 329 strong ability to accurately and precisely estimate both C_c and E_c . With increasing values of F from 50 to
 330 5000, estimates for C_c and E_c become more precise.

331 We then investigated the SIR dynamics for these parameter sets with different sample sizes (F and N). We
 332 also investigated the dynamics with different error tolerances allowed for ABC in the discrete case. For both
 333 underlying models, with $N = 5$ and $F = 50, 200, 1000$, or 5000, the true dynamics are captured by the 95%
 334 CIs (Fig 7). Additionally, for $F > 200$ in the discrete case and for all F in the continuous case, the estimated
 335 disease dynamics do not overlap those where there is assumed to be no heterogeneity in susceptibility. Hence,

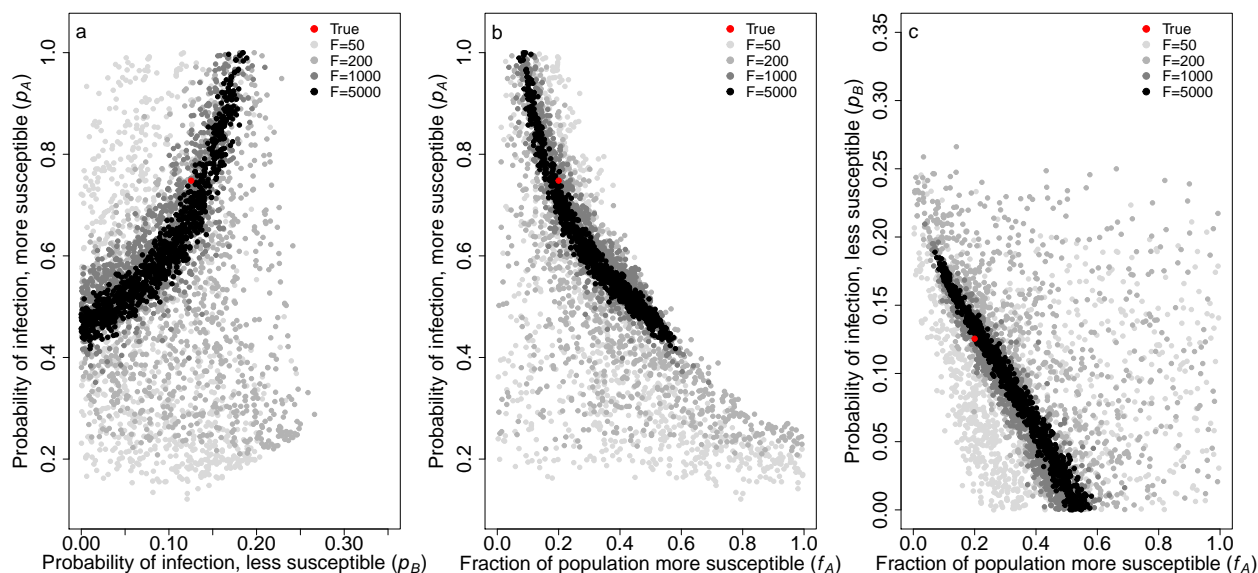


Figure 4: Parameter estimates for p_A , p_B , and f_A in the discrete case capture the true values and are highly correlated. The plots show the correlation in the parameter estimates for a) p_A vs. p_B , b) p_A vs. f_A , and c) p_B vs. f_A with different numbers of focal individuals F . These are the parameters that determine the distribution of individuals' susceptibilities in the discrete case. The red dot represents the true parameters used to generate our simulated data, and the gray dots depict 1,000 parameter sets from our posterior distribution for $F = 50$ (light gray), 200 (medium gray), 1000 (dark gray), and 5000 (black). $p_A = 0.748$, $p_B = 0.125$, $f_A = 0.2$, and $N = 5$.

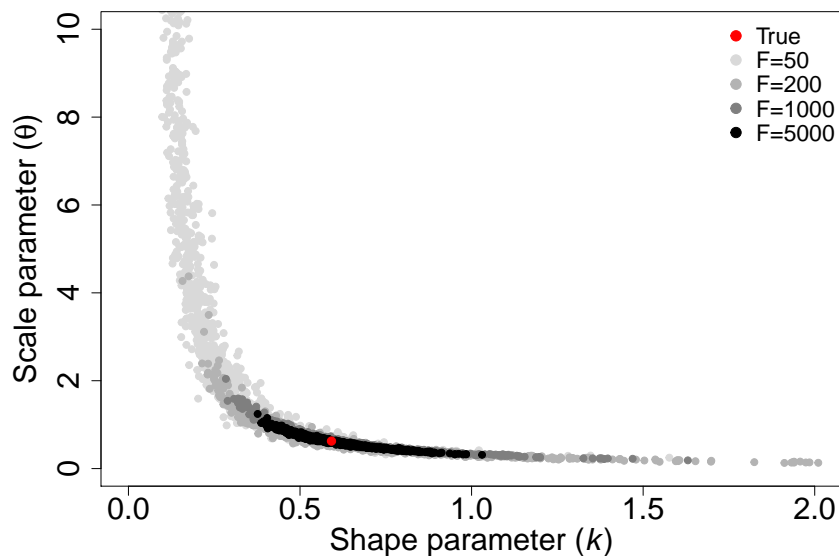


Figure 5: Parameter estimates for k and θ in the continuous case capture the true values and are highly correlated. This plot shows the correlation in the parameter estimates for k and θ that determine the gamma distribution of individuals' susceptibilities in the continuous case with different numbers of focal individuals F . The red dot represents the true parameters used to generate our simulated data, and the gray dots depict 1,000 parameter sets from our posterior distribution for $F = 50$ (light gray), 200 (medium gray), 1000 (dark gray), and 5000 (black). $k = 0.592$, $\theta = 0.626$, and $N = 5$.

336 despite low identifiability in the parameter estimates, we are able to use this method to make accurate and
 337 precise predictions about the effect of heterogeneity in susceptibility on disease dynamics. This is because

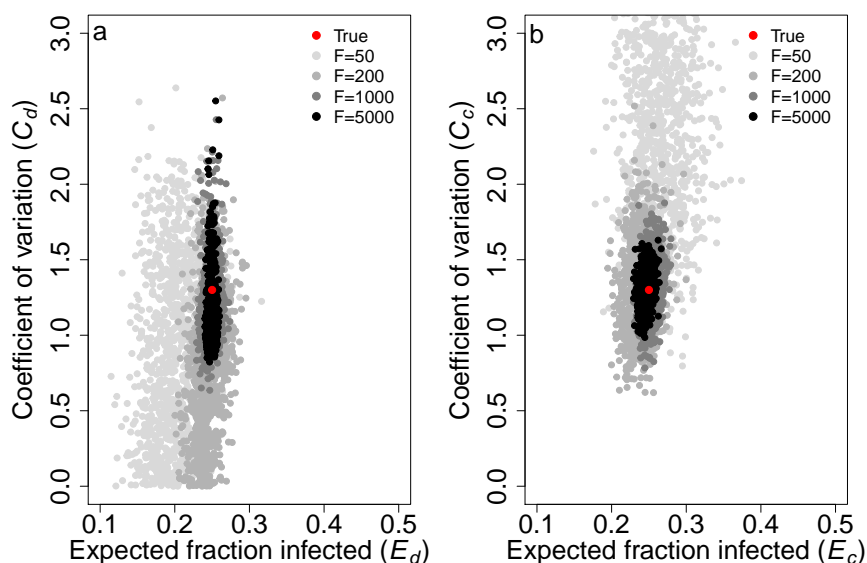


Figure 6: Parameter estimates for the coefficient of variation (C_d , C_c) and expected fraction of naive individuals infected (E_d , E_c) capture the true values and become more precise with increasing numbers of focal individuals F . The plots show the parameter estimates for C and E with different numbers of focal individuals F in a) the discrete case and b) the continuous case. The red dot represents the true parameters used to generate our simulated data, and the gray dots depict 1,000 parameter sets from our posterior distribution for $F = 50$ (light gray), 200 (medium gray), 1000 (dark gray), and 5000 (black). $C_d = C_c = 1.3$, $E_d = E_c = 0.25$, $f_A = 0.2$, and $N = 5$.

338 there is interdependence among the parameters (Figs 4, 5), and so, while individual parameters may be only
 339 partially identifiable, combinations of them can be precisely estimated, leading to relatively precise estimates
 340 of the level of heterogeneity in susceptibility C and the fraction of naive individuals infected E .

341 We found the continuous case provided more accurate and precise predictions of disease dynamics than the
 342 discrete case, but the 95% CIs narrowed with higher sample sizes in both cases (Fig 7). In the discrete case,
 343 as F increased, there was a limit to how narrow the 95% CIs became. $F > 1000$ did not substantially improve
 344 the predicted dynamics relative to those for $F = 1000$. Likewise, the number of non-focal individuals had
 345 relatively little impact on our predicted dynamics, yielding nearly identical results for $N = 5$ and $N = 100$
 346 (Supplementary information S2). In the continuous case, as F increased, the 95% CIs narrowed and converged
 347 around the true dynamics. With $N = 5$ versus $N = 100$, there was not a substantial difference in the 95%
 348 CIs (Supplementary information S2).

349 To assess the accuracy of our ABC method for parameter estimation in the discrete case, we examined
 350 the SIR dynamics with different error tolerances of 10%, 1%, or 0%. We did so with $N = 5$ and $F = 200$ and
 351 1000. Changing the error tolerance did not substantially impact the precision of the 95% CIs in any of the
 352 cases explored (Supplementary information S4).

353 We also attempted to predict disease dynamics with the wrong underlying model of individuals' risks
 354 as it may be unknown which model is correct in a real system. To do so, we generated data under the
 355 discrete case then predicted SIR dynamics assuming the continuous case and vice versa. Notably, the 95%
 356 CIs from the incorrectly assumed underlying models did not capture the true dynamics, meaning that caution
 357 should be taken in ensuring that an accurate model of heterogeneity is assumed before trusting the precise
 358 disease dynamics that would be expected to arise from a given set of parameter estimates (Supplementary
 359 information S5). Nevertheless, we stress that the ability to detect the presence of heterogeneity is independent
 360 of the underlying model and will not be affected by an incorrect model.

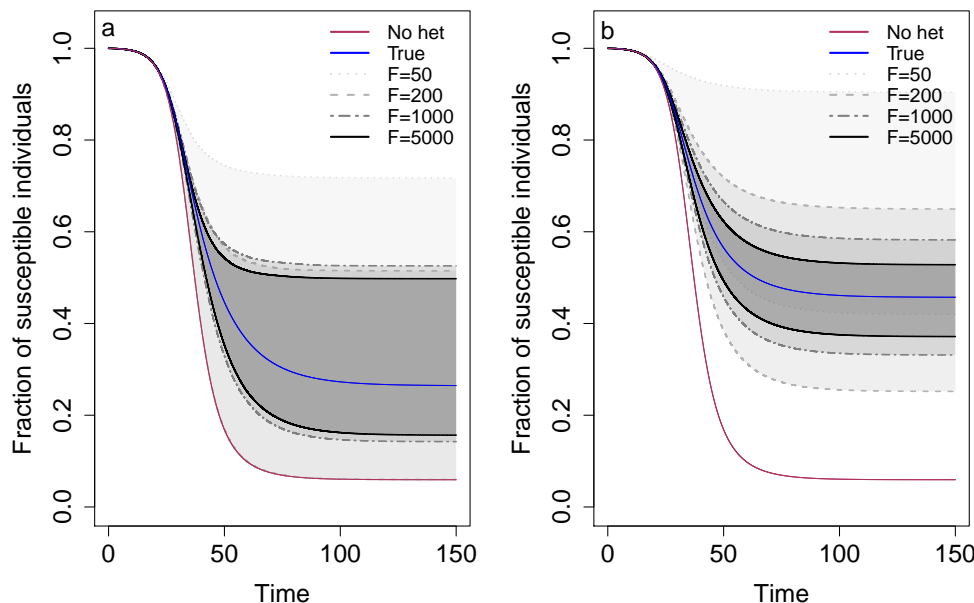


Figure 7: Predicted SIR dynamics capture the true dynamics and the 95% CIs narrow as the number of focal individuals F increases. The plots show the predicted SIR dynamics in a) the discrete case and b) the continuous case with different numbers of focal individuals F . Specifically, the fraction of susceptible individuals $\frac{S}{S_0}$ is shown over the course of an epidemic. Shaded regions represent 95% CIs determined from 1,000 posterior samples for $F = 50$ (light gray), 200 (medium gray), 1000 (dark gray), and 5000 (black). The blue line shows the true dynamics for the parameters used to generate the contact tracing data, and the red line shows the corresponding dynamics if there is homogeneity in susceptibility. $C_d = C_c = 1.3$, $E_d = E_c = 0.25$, $f_A = 0.2$, and $N = 5$.

361 3. Discussion

362 As we saw play out during the COVID-19 pandemic, early epidemiological model predictions of disease
 363 dynamics can be crucial in informing public health policy. There are numerous imperfect assumptions
 364 made by standard SIR models, and a great deal of work has been aimed at trying to improve such models.
 365 Heterogeneity in susceptibility, differences between hosts in their likelihood of becoming infected given contact,
 366 can be critically important to disease dynamics (Dwyer et al., 1997; Gomes et al., 2014; Langwig et al.,
 367 2017; Gomes et al., 2022). However, current methods to estimate this heterogeneity rely on data that is
 368 collected late in an epidemic or is unable to be collected due to ethical or logistical constraints. Here we
 369 have developed a method to detect and estimate heterogeneity using contact tracing data which, in theory,
 370 could allow epidemiologists to incorporate the effects of heterogeneity in susceptibility into their models even
 371 before the effects of such heterogeneity are observable at the population scale. Using a simulation-based
 372 approach, we found that contact tracing data alone has enough information to be used to detect and quantify
 373 heterogeneity in susceptibility. For our method, power to detect heterogeneity increases with larger sample
 374 sizes and greater heterogeneity present as well as intermediate fractions infected in the discrete case (E_d) and
 375 high fractions infected in the continuous case (E_c).

376 Few studies have estimated heterogeneity in susceptibility in any infectious disease systems. Performing
 377 a standard literature search, we were able to find 46 estimates of heterogeneity in susceptibility from only
 378 9 unique systems (Dwyer et al., 1997, 2000; Smith et al., 2005; Ben-Ami et al., 2008; Elder et al., 2008;
 379 Ben-Ami et al., 2010; Pessoa et al., 2014; Langwig et al., 2017; King et al., 2018; Gomes et al., 2019; Corder
 380 et al., 2020; Gomes et al., 2022) with only 6 of those estimates pertaining to 4 human disease systems. While
 381 this list may not be entirely exhaustive, our method may be useful for expanding the set of systems for
 382 which heterogeneity in susceptibility can be detected and estimated. To determine whether our method
 383 is sufficiently powered, we need to know whether the values of the expected fraction infected E and the
 384 coefficient of variation of risk C are in a parameter space where our method would likely be suitable. Of the

385 estimates for C that we found in the literature, 42 (91%) of them were greater than 0.5 and 21 (46%) were
386 greater than or equal to 1.5. With 200 focal individuals ($F = 200$), $f_A = 0.5$, and $C = 1.5$, we have at least
387 80% power to detect heterogeneity in susceptibility when E_d is between 0.28 and 0.92 or when E_c is between
388 0.26 and 0.98. With $F = 1000$ and $C = 1.5$, we have at least 80% power when E_d is between 0.18 and 0.98 or
389 when E_c is between 0.14 and 0.98 (Figs 2, 3). In studies examining contact tracing data, we found secondary
390 attack rates, which provide conservative estimates of E , to often be around 0.2 and sometimes as high as
391 0.733 (De Serres et al., 2000; Rieder, 2003; Taylor et al., 2007; Lessler et al., 2009; Ajelli et al., 2015; Koh
392 et al., 2020). Our method should therefore be sufficiently powered for many systems.

393 The precision in our prediction of SIR dynamics is also affected by the nature of the heterogeneity in
394 susceptibility. Our estimates of how heterogeneity affects disease dynamics are less precise when there are
395 discrete differences in risk between hosts, as opposed to continuous variation in risk (Fig 7). This is because,
396 in addition to C_d and E_d , the fraction of the initial population that is the more susceptible type of individual,
397 f_A , is critical for determining the trajectory of the epidemic. With the same C_d and E_d , the final epidemic
398 size can differ depending on f_A (Supplementary information S6). Hence, the need to estimate the additional
399 parameter f_A in the discrete case with the same data results in wider 95% CIs. However, we can generate
400 narrow 95% CIs and more precise parameter estimates in the discrete case if there is prior knowledge of the
401 parameters p_A , p_B , or f_A (Supplementary information S7).

402 We found that using the correct underlying model is additionally important for accurately predicting
403 disease dynamics, but not for the detection of heterogeneity in the first place. The underlying model used
404 for parameter estimation should therefore be carefully chosen to reflect prior understanding of the potential
405 drivers of heterogeneity in susceptibility in the system. The process for initial detection of heterogeneity in
406 susceptibility is the same regardless of the underlying model (Eqs. 1, 2). Therefore, we can reliably detect
407 heterogeneity in susceptibility without knowledge of the distribution of individuals' risks.

408 One strength of our method is that it allows for estimation of heterogeneity in susceptibility in real time,
409 early in an epidemic with no data other than contact tracing data. Admittedly, the use of this data in real
410 time will depend on the speed with which the necessary data can be collected and communicated, but existing
411 methods to quantify heterogeneity are not adequate for real time usage even with immediate access to the
412 data. Ben-Ami et al. (2010) and Langwig et al. (2017) used experimental dose-response curves to estimate
413 heterogeneity in susceptibility, and Dwyer et al. (1997) used a combination of laboratory dose-response
414 experiments, field transmission experiments, and models fit to mortality data to investigate heterogeneity.
415 Although these experimental methods can provide good estimates of heterogeneity in susceptibility, they
416 are not feasible for application in real time or for human epidemics in general due to time constraints and
417 ethical concerns. Gomes et al. (2019) compared disease incidence across municipalities in several countries to
418 quantify heterogeneity for tuberculosis. This was done by ordering the municipalities by incidence rate and
419 plotting the percentage of cumulative tuberculosis cases versus cumulative population to construct Lorenz
420 curves and thereby fit susceptibility risk distributions. This method, however, requires a considerable amount
421 of data with ten or more years of data used in this study. Smith et al. (2005) and Corder et al. (2020) used
422 malaria morbidity data to fit models of malaria and estimate heterogeneity. This method cannot be used until
423 later in an epidemic when sufficient data is collected to fit curves. Gomes et al. (2022) also used curve fitting
424 with mortality data to estimate heterogeneity in susceptibility for COVID-19. They were able to estimate
425 heterogeneity in real time once at least four months of data were available. While our method is in principle
426 able to estimate heterogeneity in a similar time frame provided robust contact tracing, we also note that their
427 method is heavily dependent on the underlying model and assumptions, and the authors advise not to trust
428 the precision of their estimates. In addition, Gomes et al. were unable to disentangle heterogeneity in contact
429 rate from heterogeneity in underlying susceptibility. Our method estimates heterogeneity in underlying
430 susceptibility, and the remaining heterogeneity in contact rate can be determined from the contact network
431 data. Anderson et al. (2023) used household study data to estimate heterogeneity in susceptibility. While
432 this method is suitable for use in real time, and can be applied to human infectious diseases, the method
433 notably is designed to estimate heterogeneity within households, which is not the same as the population-level
434 heterogeneity that drives population-level disease dynamics.

435 Our method is unable to precisely estimate the individual parameters that define the risk distributions
436 (i.e. p_A , p_B , f_A in the discrete case and k , θ in the continuous case), but our method is able to reliably
437 predict disease dynamics. This seeming paradox arises because the disease dynamics depend on combinations
438 of parameters rather than individual parameters. Notably, our method is substantially better at estimating

439 the composite parameters describing the coefficient of variation of risk C and the expected fraction of naive
440 individuals infected E . Nevertheless, our method does require a substantial amount of data (200 individuals
441 showing up in contact networks for a second time). This requirement could be mitigated by pooling contact
442 network data from multiple locations in order to more quickly collect sufficient data. It may also be possible
443 to combine our method with another, like that of Gomes et al. (2022), to reduce the data required by either
444 method.

445 There are additionally several considerations to address with regard to working with contact tracing data.
446 Perhaps most prominently, contact tracing data tend to be messy and imperfect. Our method as described
447 above assumes perfect data. However, our method can be readily modified to account for imperfect data. We
448 can imagine multiple ways in which contact tracing data may be imperfect. Some important considerations are
449 that: a) individuals may be mislabeled as uninfected when they are infected (false negatives), b) individuals
450 may be mislabeled as infected when they are uninfected (false positives), and c) individuals may be missing
451 from the contact networks despite being contacts (missing contacts). If there are false negatives, our method
452 may overestimate the level of heterogeneity because our estimate of p_f may be biased lower. This is because,
453 assuming infection confers at least partial immunity, focal individuals that were actually infected previously
454 (i.e. false negatives) will be less likely to be infected than focal individuals that were true negatives. To
455 counteract this issue, we developed a version of the method that corrects for false negatives by adjusting
456 the likelihood calculations for both detecting and estimating heterogeneity. For estimating parameters and
457 predicting disease dynamics, adjusting the method to correct for false negatives fixes the issue (Supplementary
458 information S8). For detecting heterogeneity in susceptibility, adjusting the likelihood calculation corrects
459 for the impact of false negatives except when the expected fraction infected E_d is very close to 1. We do
460 not think this will be a major issue as E_d is typically less than 0.5 (De Serres et al., 2000; Rieder, 2003;
461 Taylor et al., 2007; Lessler et al., 2009; Ajelli et al., 2015; Koh et al., 2020). If there are false positives, our
462 method may underestimate the level of heterogeneity because our estimate of p_f may be biased higher. This
463 is because a high false positive rate will have a larger impact on making individuals with a low susceptibility
464 appear infected than those with a high susceptibility that are more likely to be true positives. Hence, focal
465 individuals, which are on average less susceptible, and naive individuals will appear to have more similar
466 infection probabilities. However, false positive rates are often small, close to 1-2% (Yang and Rothman,
467 2004; Cohen et al., 2020), so this issue is not a huge concern for our method unless false positive rates are
468 known to be unusually large. If there are many missing contacts, our method could underestimate the level
469 of heterogeneity because our estimate of p_n may be biased lower. This is because individuals that we believe
470 to be naive but were previously exposed in a first contact network may be less likely to be infected than
471 true naive individuals. These missed individuals may have gained immunity through infection or may be on
472 average less susceptible through the infection selection process. However, there is a low chance of a missed
473 individual from a first contact network showing up in a second contact network that also happens to have
474 a focal individual early in an epidemic. So, missing individuals should have only a negligible effect on the
475 method's performance in these early stages. Later on in an epidemic, this source of bias will become more
476 important to consider. While we have considered these three ways in which contact tracing data may be
477 imperfect, it is highly likely that each set of contact tracing data will have its own set of peculiarities. Note
478 that these peculiarities, if known, can readily be accounted for using our ABC method since any process may
479 be used for simulation. Known imperfections in the data should therefore not bias estimates although they
480 may still reduce power or increase required sample sizes.

481 Another important point is that our method assumes no forms of heterogeneity other than heterogeneity
482 in susceptibility. One other source of heterogeneity is heterogeneity in transmission (Lloyd-Smith et al., 2005).
483 Heterogeneity in transmission is differences between hosts in their likelihood of transmitting a pathogen
484 once infected. If this heterogeneity arises due to variation in the number of contacts that individuals have,
485 then heterogeneity in transmission poses no problems for our method. It would simply mean that each
486 contact network would have a unique value for N . We note that this variation in contact rate is the typical
487 mechanism through which heterogeneity in transmission is assumed to act (Lloyd-Smith et al., 2005). However,
488 if heterogeneity in transmission arises due to differences between hosts in their likelihood of transmission
489 given contact, our method may have less power to detect heterogeneity in susceptibility and may yield less
490 precise or faulty conclusions about the disease dynamics (Supplementary information S9). Our method, in its
491 existing form, is thus not suitable in these cases. This concern can be partially mitigated by performing a
492 goodness of fit test before implementing the method to determine whether there is evidence of heterogeneity

493 in transmission given contact (Supplementary information S9). If there is heterogeneity in transmission,
494 then our method should not be used. A next step in developing this method will be to generalize it to allow
495 for estimation of heterogeneity in susceptibility even when there is heterogeneity in transmission. This is,
496 however, a non-trivial problem because if every individual has a unique force of infection, then the number of
497 parameters to estimate grows at the same rate as the number of focal individuals.

498 There may additionally be heterogeneity in exposure strength among contacts within a network such
499 that individuals experience different forces of infection. This could be due to factors like differences in
500 exposure time or type of contact (e.g., contacts that shared a taxi, were at the same party, etc.). This
501 added heterogeneity may reduce the power of our method to detect heterogeneity in susceptibility as different
502 contact types may provide varying levels of information that our current method disregards. To alleviate
503 the potential impact of this heterogeneity, it may be necessary to break apart contact networks into specific
504 exposure events and either weigh the type of contact differently or only use equivalent contact types.

505 Finally, we note that exposure could change individuals' susceptibilities. Individuals exposed in a first
506 contact network could receive a small dose of the pathogen such that their immune system is stimulated
507 without them becoming infected. This could decrease their susceptibility, meaning that some focal individuals
508 have lower susceptibilities because they developed immunity, not because they were innately less susceptible
509 (Leon and Hawley, 2017). However, this will have the same effect as heterogeneity in susceptibility of slowing
510 down the epidemic and could even be considered a form of heterogeneity in susceptibility.

511 The earliest practice of tracing diseases dates back to the 1500s when doctors would track the spread of
512 syphilis (Cohn, 2018), and the earliest known example of contact tracing dates to 1576 during a bubonic
513 plague pandemic (Cohn, 2009). Since then, the practice of contact tracing has spread, and it is now used
514 widely, ranging from diseases such as influenza to HIV (De Serres et al., 2000; Rieder, 2003; Taylor et al.,
515 2007; Lessler et al., 2009; Ajelli et al., 2015; Koh et al., 2020). Recently, contact tracing data has transitioned
516 from paper copies to electronic databases. Regardless, all of these sources of data could be used with our
517 method provided they include focal individuals that are identifiable between contact networks, specify which
518 individuals are infected, and have a sufficient sample size. Using our method, it should therefore, without
519 collecting any new data, be possible to estimate heterogeneity in susceptibility, in various locations and time
520 periods, for dozens of disease systems in which it has never been estimated previously.

521 **Acknowledgements:** We thank the Read, McGraw, and Kennedy labs for stimulating discussions.

522 References

- 523 Aguas, R., Corder, R.M., King, J.G., Goncalves, G., Ferreira, M.U., Gomes, M.G.M., 2020. Herd immunity
524 thresholds for SARS-CoV-2 estimated from unfolding epidemics. medRxiv .
- 525 Ajelli, M., Parlamento, S., Bome, D., Kebbi, A., Atzori, A., Frasson, C., Putoto, G., Carraro, D., Merler,
526 S., 2015. The 2014 Ebola virus disease outbreak in Pujehun, Sierra Leone: epidemiology and impact of
527 interventions. BMC Medicine 13. 281.
- 528 Anderson, R.M., May, R.M., 1984. Spatial, temporal, and genetic heterogeneity in host populations and
529 the design of immunization programmes. Mathematical Medicine and Biology: A Journal of the IMA 1,
530 233–266.
- 531 Anderson, T.L., Nande, A., Merenstein, C., Raynor, B., Oommen, A., Kelly, B.J., Levy, M.Z., Hill, A.L.,
532 2023. Quantifying individual-level heterogeneity in infectiousness and susceptibility through household
533 studies. Epidemics 44. 100710.
- 534 Barquet, N., Domingo, P., 1997. Smallpox: the triumph over the most terrible of the ministers of death.
535 Annals of Internal Medicine 127, 635–642.
- 536 Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics.
537 Genetics 162, 2025–2035.
- 538 Ben-Ami, F., Ebert, D., Regoes, R.R., 2010. Pathogen dose infectivity curves as a method to analyze the
539 distribution of host susceptibility: a quantitative assessment of maternal effects after food stress and
540 pathogen exposure. The American Naturalist 175, 106–115.
- 541 Ben-Ami, F., Regoes, R.R., Ebert, D., 2008. A quantitative test of the relationship between parasite dose
542 and infection probability across different host–parasite combinations. Proceedings of the Royal Society B:
543 Biological Sciences 275, 853–859.
- 544 Bradshaw, W.J., Alley, E.C., Huggins, J.H., Lloyd, A.L., Esvelt, K.M., 2021. Bidirectional contact tracing
545 could dramatically improve COVID-19 control. Nature Communications 12. 232.
- 546 Chandra, R., 1979. Nutritional deficiency and susceptibility to infection. Bulletin of the World Health
547 Organization 57, 167–177.
- 548 Cohen, A.N., Kessel, B., G, M.M., 2020. Diagnosing SARS-CoV-2 infection: the danger of over-reliance on
549 positive test results. medRxiv .
- 550 Cohn, Jr., S.K., 2009. Plague Disputes, Challenges of the ‘Universals’, in: Cultures of Plague: Medical thinking
551 at the end of the Renaissance. Oxford University Press. doi:10.1093/acprof:oso/9780199574025.003.0007.
- 552 Cohn, Jr., S.K., 2018. Syphilis: Naming and Blaming?, in: Epidemics: Hate and Compassion from the Plague
553 of Athens to AIDS. Oxford University Press. doi:10.1093/oso/9780198819660.003.0006.
- 554 Corder, R.M., Ferreira, M.U., Gomes, M.G.M., 2020. Modelling the epidemiology of residual Plasmodium
555 vivax malaria in a heterogeneous host population: A case study in the Amazon Basin. PLoS Computational
556 Biology 16, e1007377.
- 557 De Serres, G., Shadmani, R., Duval, B., Boulianne, N., Déry, P., Fradet, M.D., Rochette, L., Halperin, S.A.,
558 2000. Morbidity of pertussis in adolescents and adults. The Journal of Infectious Diseases 182, 174–179.
- 559 Dhar, A., 2020. What one can learn from the SIR model, in: Indian Academy of Sciences Conference Series.
- 560 Dobner, J., Kaser, S., 2018. Body mass index and the risk of infection-from underweight to obesity. Clinical
561 Microbiology and Infection 24, 24–28.
- 562 Dwyer, G., Dushoff, J., Elkinton, J.S., Levin, S.A., 2000. Pathogen-driven outbreaks in forest defoliators
563 revisited: building models from experimental data. The American Naturalist 156, 105–120.

- 564 Dwyer, G., Elkinton, J.S., Buonaccorsi, J.P., 1997. Host heterogeneity in susceptibility and disease dynamics:
565 tests of a mathematical model. *The American Naturalist* 150, 685–707.
- 566 Eames, K.T., Keeling, M.J., 2003. Contact tracing and disease control. *Proceedings of the Royal Society B:*
567 *Biological Sciences* 270, 2565–2571.
- 568 Elder, B.D., Dushoff, J., Dwyer, G., 2008. Host-pathogen interactions, insect outbreaks, and natural selection
569 for disease resistance. *The American Naturalist* 172, 829–842.
- 570 Gomes, M.G.M., Ferreira, M.U., Corder, R.M., King, J.G., Souto-Maior, C., Penha-Gonçalves, C., Gonçalves,
571 G., Chikina, M., Pegden, W., Aguas, R., 2022. Individual variation in susceptibility or exposure to
572 SARS-CoV-2 lowers the herd immunity threshold. *Journal of Theoretical Biology* 540, 111063.
- 573 Gomes, M.G.M., Lipsitch, M., Wargo, A.R., Kurath, G., Rebelo, C., Medley, G.F., Coutinho, A., 2014. A
574 missing dimension in measures of vaccination impacts. *PLoS Pathogens* 10, e1003849.
- 575 Gomes, M.G.M., Oliveira, J.F., Bertolde, A., Ayabina, D., Nguyen, T.A., Maciel, E.L., Duarte, R., Nguyen,
576 B.H., Shete, P.B., Lienhardt, C., 2019. Introducing risk inequality metrics in tuberculosis policy development.
577 *Nature Communications* 10, 2480.
- 578 Hébert-Dufresne, L., Althouse, B.M., Scarpino, S.V., Allard, A., 2020. Beyond R_0 : heterogeneity in secondary
579 infections and probabilistic epidemic forecasting. *Journal of the Royal Society Interface* 17, 20200393.
- 580 Hossain, A.D., Jarolimova, J., Elnaïem, A., Huang, C.X., Richterman, A., Ivers, L.C., 2022. Effectiveness
581 of contact tracing in the control of infectious diseases: a systematic review. *The Lancet Public Health* 7,
582 E259–E273.
- 583 Huang, Y., Paxton, W.A., Wolinsky, S.M., Neumann, A.U., Zhang, L., He, T., Kang, S., Ceradini, D., Jin,
584 Z., Yazdanbakhsh, K., et al., 1996. The role of a mutant CCR5 allele in HIV-1 transmission and disease
585 progression. *Nature Medicine* 2, 1240–1243.
- 586 Keeling, M., Danon, L., 2009. Mathematical modelling of infectious diseases. *British Medical Bulletin* 92,
587 33–42.
- 588 Kennedy, D.A., Dukic, V., Dwyer, G., 2015. Combining principal component analysis with parameter
589 line-searches to improve the efficacy of metropolis-hastings mcmc. *Environmental and Ecological Statistics*
590 22, 247–274.
- 591 King, J.G., Souto-Maior, C., Sartori, L.M., Maciel-de Freitas, R., Gomes, M.G.M., 2018. Variation in
592 Wolbachia effects on *Aedes* mosquitoes as a determinant of invasiveness and vectorial capacity. *Nature*
593 *Communications* 9, 1483.
- 594 Koh, W.C., Naing, L., Chaw, L., Rosledzana, M.A., Alikhan, M.F., Jamaludin, S.A., Amin, F., Omar, A.,
595 Shazli, A., Griffith, M., et al., 2020. What do we know about SARS-CoV-2 transmission? A systematic
596 review and meta-analysis of the secondary attack rate and associated risk factors. *PloS One* 15, e0240205.
- 597 Langwig, K.E., Wargo, A.R., Jones, D.R., Viss, J.R., Rutan, B.J., Egan, N.A., Sá-Guimarães, P., Kim, M.S.,
598 Kurath, G., Gomes, M.G.M., Lipsitch, M., 2017. Vaccine effects on heterogeneity in susceptibility and
599 implications for population health management. *mBio* 8, e00796–17.
- 600 Larson, E., 1999. Skin hygiene and infection prevention: more of the same or different approaches? *Clinical*
601 *Infectious Diseases* 29, 1287–1294.
- 602 Leon, A.E., Hawley, D.M., 2017. Host responses to pathogen priming in a natural songbird host. *EcoHealth*
603 14, 793–804.
- 604 Lessler, J., Reich, N.G., Cummings, D.A., of Health, N.Y.C.D., Team, M.H.S.I.I., 2009. Outbreak of 2009
605 pandemic influenza A (H1N1) at a New York City school. *The New England Journal of Medicine* 361,
606 2628–2636.

- 607 Li, J., Liu, Y., Kim, T., Min, R., Zhang, Z., 2010. Gene expression variability within and between human
608 populations and implications toward disease susceptibility. *PLoS Computational Biology* 6, e1000910.
- 609 Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Getz, W.M., 2005. Superspreading and the effect of individual
610 variation on disease emergence. *Nature* 438, 355–359.
- 611 Montalbán, A., Corder, R.M., Gomes, M.G.M., 2022. Herd immunity under individual variation and
612 reinfection. *Journal of Mathematical Biology* 85. 2.
- 613 Pessoa, D., Souto-Maior, C., Gjini, E., Lopes, J.S., Ceña, B., Codeço, C.T., Gomes, M.G.M., 2014. Unveiling
614 time in dose-response models to infer host susceptibility to pathogens. *PLoS Computational Biology* 10,
615 e1003773.
- 616 Plotkin, S.A., 2008. Correlates of vaccine-induced immunity. *Clinical Infectious Diseases* 47, 401–409.
- 617 R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical
618 Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- 619 Rieder, H., 2003. Contacts of tuberculosis patients in high-incidence countries. *The International Journal of*
620 *Tuberculosis and Lung Disease* 7, S333–S336.
- 621 Roberts, M., Andreasen, V., Lloyd, A., Pellis, L., 2015. Nine challenges for deterministic epidemic models.
622 *Epidemics* 10, 49–53.
- 623 Van der Sande, M., Teunis, P., Sabel, R., 2008. Professional and home-made face masks reduce exposure to
624 respiratory infections among the general population. *PLoS One* 3, e2618.
- 625 Shaw, C.L., Kennedy, D.A., 2021. What the reproductive number R_0 can and cannot tell us about COVID-19
626 dynamics. *Theoretical Population Biology* 137, 2–9.
- 627 Smith, D., Dushoff, J., Snow, R., Hay, S., 2005. The entomological inoculation rate and *Plasmodium*
628 *falciparum* infection in African children. *Nature* 438, 492–495.
- 629 Taylor, M.M., Rotblatt, H., Brooks, J.T., Montoya, J., Aynalem, G., Smith, L., Kenney, K., Laubacher, L.,
630 Bustamante, T., Kim-Farley, R., et al., 2007. Epidemiologic investigation of a cluster of workplace HIV
631 infections in the adult film industry: Los Angeles, California, 2004. *Clinical Infectious Diseases* 44, 301–305.
- 632 Tolles, J., Luong, T., 2020. Modeling epidemics with compartmental models. *JAMA* 323, 2515–2516.
- 633 VanderWaal, K.L., Ezenwa, V.O., 2016. Heterogeneity in pathogen transmission: mechanisms and methodology.
634 *Functional Ecology* 30, 1606–1622.
- 635 Woolhouse, M.E., Dye, C., Etard, J.F., Smith, T., Charlwood, J., Garnett, G., Hagan, P., Hii, J., Ndhlovu,
636 P., Quinell, R., et al., 1997. Heterogeneities in the transmission of infectious agents: implications for the
637 design of control programs. *Proceedings of the National Academy of Sciences* 94, 338–342.
- 638 Yang, S., Rothman, R.E., 2004. PCR-based diagnostics for infectious diseases: uses, limitations, and future
639 applications in acute-care settings. *The Lancet Infectious Diseases* 4, 337–348.