

## Research



**Cite this article:** Robert A, Kucharski AJ, Gastañaduy PA, Paul P, Funk S. 2020 Probabilistic reconstruction of measles transmission clusters from routinely collected surveillance data. *J. R. Soc. Interface* **17**: 20200084.  
<http://dx.doi.org/10.1098/rsif.2020.0084>

Received: 6 February 2020

Accepted: 8 June 2020

### Subject Category:

Life Sciences—Mathematics interface

### Subject Areas:

computational biology

### Keywords:

transmission tree reconstruction, Markov chain Monte Carlo, Bayesian statistics, measles

### Author for correspondence:

Alexis Robert

e-mail: alexis.robert@lshtm.ac.uk

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.5025773>.

# Probabilistic reconstruction of measles transmission clusters from routinely collected surveillance data

Alexis Robert<sup>1,2</sup>, Adam J. Kucharski<sup>1,2</sup>, Paul A. Gastañaduy<sup>3</sup>, Prabasaj Paul<sup>4</sup> and Sebastian Funk<sup>1,2</sup>

<sup>1</sup>Department of Infectious Disease Epidemiology, and <sup>2</sup>Centre for the Mathematical Modelling of Infectious Disease, London School of Hygiene and Tropical Medicine, London, UK

<sup>3</sup>Division of Viral Diseases, and <sup>4</sup>Division of Healthcare Quality Promotion, Centers for Disease Control and Prevention, Atlanta, GA, USA

AR, 0000-0002-4516-2965; AJK, 0000-0001-8814-9421; PAG, 0000-0003-2690-4799; PP, 0000-0002-4096-6171; SF, 0000-0002-2842-3406

Pockets of susceptibility resulting from spatial or social heterogeneity in vaccine coverage can drive measles outbreaks, as cases imported into such pockets are likely to cause further transmission and lead to large transmission clusters. Characterizing the dynamics of transmission is essential for identifying which individuals and regions might be most at risk. As data from detailed contact-tracing investigations are not available in many settings, we developed an R package called *o2geosocial* to reconstruct the transmission clusters and the importation status of the cases from their age, location, genotype and onset date. We compared our inferred cluster size distributions to 737 transmission clusters identified through detailed contact-tracing in the USA between 2001 and 2016. We were able to reconstruct the importation status of the cases and found good agreement between the inferred and reference clusters. The results were improved when the contact-tracing investigations were used to set the importation status before running the model. Spatial heterogeneity in vaccine coverage is difficult to measure directly. Our approach was able to highlight areas with potential for local transmission using a minimal number of variables and could be applied to assess the intensity of ongoing transmission in a region.

## 1. Introduction

Establishing who infected whom during an outbreak can help inform the design and evaluation of control measures [1–5]. Transmission links can be reconstructed through contact-tracing investigations, whereby cases are asked their movements and contacts during their infectious period. Given that contact-tracing investigations are not always carried out due to the logistical effort and cost involved, inference methods have been developed to use epidemiological data to estimate the probability that a transmission event occurred between any given pair of cases [6–12]. This makes it possible to establish probabilistic transmission trees that link all observed cases. The ensemble of cases belonging to the same transmission tree is called a transmission cluster.

Wallinga & Teunis [2] first developed a likelihood-based estimation procedure to reconstruct probabilistic transmission trees from a given distribution of generation times and observed symptom-onset dates of each case. Since then, genomic, spatial or contact data have been used to supplement the timing of symptoms, which helped identify determinants of transmission, mixing behaviour, individual dispersion, evaluate control measures, anticipate future developments of outbreaks and study viral evolutionary patterns [5,8,9,13–17].

As sequencing of pathogens has become more common, the use of such data to infer transmission trees has increased. Methods developed to add genetic distance to a Wallinga–Teunis algorithm, where cases with lower genetic distance are more likely to be grouped in the same transmission group, showed it substantially increased the accuracy of the reconstructed transmission trees [8,18–21].

The utility of sequence data depends on the characteristics of the pathogen [22,23]. Based on the highly variable 450 nucleotides region of the N gene (N-450) of the measles virus genome, eight measles genotypes have been detected since 2009 [24,25]; these genotype designations are helpful in linking cases, as linked cases must be infected by a virus of the same genotype [25]; however, the diversity of measles genotypes is decreasing [26]. It has been suggested that further sequencing the M-F non-coding region, or full genome sequencing, could help identify measles virus transmission trees, but so far, extended sequencing during measles outbreaks has been scarce [27,28]. In addition, the evolutionary rate of measles virus is very low [29]; therefore, samples from unrelated cases can be very close genetically and genetic sequences from measles cases are not usually indicative of direct transmission links [27,28].

As measles is highly infectious, under-immunized communities (also called pockets of susceptibles) resulting from local heterogeneity in vaccine coverage can lead to large, long-lasting outbreaks [30–34]. Detecting these pockets of susceptibles can be challenging, as historical local values of coverage throughout a given country are rarely available. The number of cases in the transmission trees resulting from each importation during outbreaks, also called the cluster size distribution, will depend both on individual factors (e.g. age of the imported case which might affect contact patterns) and community factors (e.g. the history of coverage in the area) [35,36]. The size of a cluster can, therefore, reflect the level of susceptibility of individuals directly and indirectly connected to the imported case [37,38].

Here, we introduced a model combining age, location, genotype and rash onset date of cases to reconstruct probabilistic transmission trees. We chose these features to make the model applicable to a wide range of settings as they are commonly reported and informative on transmission. We wrote the R package *o2geosocial* to conduct inference on individual-level data using this model. It is based on the package *outbreaker2* and is designed for outbreaks with partial sampling of cases, or uninformative genetic sequences, such as measles outbreaks [9,39]. We used the likelihood of transmission links between different cases to estimate their importation status. We compared the inferred importation status and cluster size distribution to the transmission clusters identified via contact tracing during measles outbreaks in the USA between 2001 and 2016.

## 2. Methods

### 2.1. Presentation of the algorithm

Transmission trees are used to represent who infected whom during an outbreak. They are directed acyclic graphs, where nodes are the reported cases and edges show the connection between them. The root of each transmission tree is an imported case, i.e. a case who was infected in a different transmission setting. The cases placed in the same transmission tree form a

**Table 1.** Table of notations of all variables and distributions defined in the methods.

parameter	symbol
onset date	$t_i, t_j$
infection date	$T_i$
age	$\alpha_i, \alpha_j$
tree	$\tau_j$
genotype	$g_i, g_\tau$
region	$r_i, r_j$
number of generations	$\kappa_{ji}$
spatial parameters	$a, b, c$
conditional report ratio	$\rho$
connectivity	$n_{r_i r_j}$
population	$m_{r_i}, m_{r_j}$
distance	$d_{r_i r_j}$
parameter set	$\theta$
importation threshold	$\lambda$
generation time distribution	$w(t_i - t_j)$
latent period distribution	$f(t_i - T_i)$
age contact probability	$a(\alpha_i, \alpha_j)$
genotype probability	$G(g_i, g_\tau)$
probability of missing generation	$p(\kappa_{ji} \rho)$
spatial probability	$s(r_i, r_j   a, b)$
log-likelihood of connection between $i$ and $j$	$L_{ji}(t_i, t_j, \theta)$
individual log-likelihood	$L_i(t_i, j, t_j, \theta)$

transmission cluster. We estimated the number of cases per cluster (cluster size distribution) and the importation status of the cases from probabilistic transmission trees inferred using routinely collected epidemiological variables.

We used a Metropolis–Hastings algorithm with Markov chain Monte Carlo (MCMC) to classify a set of cases into a set of transmission trees with associated probabilities quantified using a Bayesian model to combine the epidemiological features of the cases. At every iteration of the MCMC algorithm, we proposed a new set of model parameters, infection dates and connections between cases. These three elements formed a tree proposal. We computed the ratio between the posterior probability of this proposal and the current posterior probability. The posterior probability (up to a multiplicative constant which would cancel out when calculating the ratio) was calculated from the likelihood of the trees, and the prior probability of the parameters. The log-likelihood of each tree was equal to the sum of the log-likelihoods of each case. All the notations are defined in table 1.

#### 2.1.1. Likelihood function and parameter definition

In a tree proposal, each case  $i$  was assigned an infector  $j$  and an infection date  $t_i$ . We computed the log-likelihood of each case,  $L_i(t_i, j, t_j, \theta)$  to calculate the likelihood of the tree. The log-likelihood of  $i$  was split in two: (i) the log-probability density of observing the onset date  $T_i$  if case  $i$  had been infected at time  $t_i$   $\log(f(t_i - T_i))$  and (ii) the log-likelihood of connection between  $i$  and  $j$   $L_{ji}(t_i, t_j, \theta)$ , with  $\theta$  the parameter set of the model (2.1):

$$L_i(t_i, j, t_j, \theta) = \log(f(t_i - T_i)) + L_{ji}(t_i, t_j, \theta), \quad (2.1)$$

The function  $f$  represents the distribution of the incubation period. The log-likelihood of connection  $L_{ji}$  was computed from five components reflecting the age group, genotype, location, inferred date of infection of cases  $i$  and  $j$ , and the report ratio (2.2). We allowed for an indirect link between cases due to unreported individuals,  $\kappa_{ji}$  corresponds to the number of generations between  $i$  and  $j$ . If  $\kappa_{ji} = 1$ ,  $j$  infected  $i$ , whereas if  $\kappa_{ji} = 2$ , an unreported case infected by  $j$  infected  $i$ ,  $\kappa_{ji}$  increases with the number of missing links between  $i$  and  $j$

$$L_{ji}(t_i, t_j, \theta) = \log(p(\kappa_{ji}|\rho) \times w^{(\kappa_{ji})}(t_i - t_j) \times a^{(\kappa_{ji})}(\alpha_i, \alpha_j) \times G(g_i, g_{\tau_j}) \times s^{(\kappa_{ji})}(r_i, r_j | a, b)). \quad (2.2)$$

We calculated the temporal probability of transmission between  $i$  and  $j$  from the number of days between  $t_i$  and  $t_j$  and the distribution of the generation time of the disease  $w(t)$ . This probability was quantified by  $w^{(\kappa_{ji})}(t_i - t_j, \kappa_{ji})$ ,  $w^{(\kappa_{ji})} = w * w * \dots * w$ , where  $*$  is the convolution operator applied  $\kappa_{ji}$  times. We used a geometric distribution  $p(\kappa_{ji}|\rho)$  to quantify the probability of observing  $\kappa_{ji}$  missing generation between  $i$  and  $j$ , given the conditional report ratio  $\rho$ . The conditional report ratio quantifies the probability of missing generations between two connected reported cases. Entire missing clusters, cases infected after the last cases or cases infected before the ancestor of a cluster would not interfere in the connection between two cases and, therefore, would not affect the value of the conditional report ratio. The conditional report ratio can be higher than the overall report ratio of an outbreak. The 'ancestor' is the earliest identified case in a cluster.

$a(\alpha_i, \alpha_j, \kappa_{ji})$  was defined as the probability of transmission between age groups  $\alpha_i$  and  $\alpha_j$ . This probability corresponds to the proportion of contacts to the age group  $\alpha_i$  that originated from  $\alpha_j$  and can be deduced from studies such as POLYMOD [36]. We defined  $G(g_i, g_{\tau_j})$  as the probability of observing the pathogen genotype  $g_i$  in case  $i$  in the tree  $\tau_j$  containing case  $j$ . There can only be one measles virus genotype per transmission tree, or cases with unreported genotype. The genotype  $g_{\tau_j}$  is the genotype contained in the tree  $\tau_j$  and is known if at least one case in  $\tau_j$  had a reported genotype

$$G(g_i, g_{\tau_j}) = \begin{cases} 1 & \text{if } g_i \text{ unknown} \\ 1 & \text{if } g_{\tau_j} \text{ unknown} \\ 1 & \text{if } g_i \text{ and } g_{\tau_j} \text{ both known and } g_i = g_{\tau_j} \\ 0 & \text{otherwise} \end{cases} \quad (2.3)$$

In (2.3), if  $G(g_i, g_{\tau_j}) = 0$ , then the connection between  $i$  and  $j$  is impossible, and (2.1) and (2.2) are equal to  $\log(0) = -\infty$ .

$s(r_i, r_j, \kappa_{ji})$  was defined as the probability of connection from  $r_j$  to  $r_i$ , regions of residency of  $i$  and  $j$  (2.4). We used an exponential gravity model to quantify the connectivity of the different geographical units [40]. This approach showed good performance at modelling short-distance commuting, and was easy to parametrize [40–44]. In the simplest form of the exponential gravity model, the number of connections between  $r_i$  and  $r_j$  is proportional to the product of the origin population  $m_{r_j}$ , the destination population  $m_{r_i}$  and an exponential decrease of the distance between  $r_i$  and  $r_j$   $d_{r_j r_i} \cdot n_{r_j r_i} \propto e^{-a \times d_{r_j r_i}} \times m_{r_j}^b \times m_{r_i}^c$ , with  $a$ ,  $b$  and  $c$  parameters adjusting for the impact of distance and population. From this definition, we deduced  $s(r_j, r_i)$ , the spatial probability of transmission from  $i$  to  $j$

$$s(r_i, r_j) = \frac{n_{r_j r_i}}{\sum_h n_{hr_i}} = \frac{e^{-a \times d_{r_j r_i}} \times m_{r_j}^b \times m_{r_i}^c}{\sum_h e^{-a \times d_{hr_i}} \times m_h^b \times m_{r_i}^c} = \frac{e^{-a \times d_{r_j r_i}} \times m_{r_j}^b}{\sum_h e^{-a \times d_{hr_i}} \times m_h^b}. \quad (2.4)$$

Only the parameters  $a$  and  $b$  were required to compute the spatial probability of transmission. If  $r_i = r_j$ , then (2.4) becomes:  $s(r_i, r_j) = m_{r_i}^b / \sum_h m_h^b$ . Other distributions than the exponential decrease can be used in this framework if transmission follows a different pattern.

The parameters  $\rho$ ,  $a$  and  $b$  were estimated. At each iteration of the MCMC, the log-likelihood of the trees was equal to the sum of all individual log-likelihoods  $L_i$  from equation (2.1). The log-posterior density of the proposed trees was calculated by summing the overall log-likelihood of the trees and the log-priors of the parameters.

### 2.1.2. Tree proposals

We used a Metropolis–Hastings algorithm with MCMC to sample from the posterior distribution of parameters and transmission trees. To do this, we developed a set of proposal tree updates. These updates were accepted with acceptance probability as defined by the Metropolis–Hastings algorithm [45]. We used eight types of tree proposal to ensure good mixing. Each proposal conserved the overall number of trees, with a maximum of one unique genotype reported per tree.

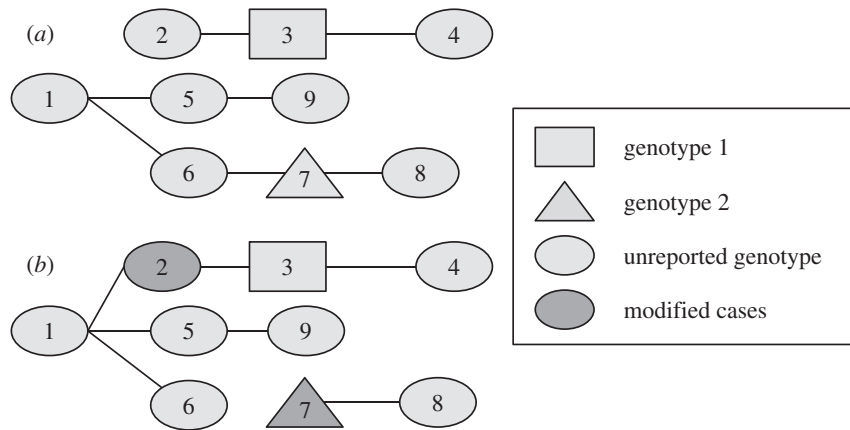
Five of the proposals had already been implemented in the *outbreaker2* package and were adapted to this setting: (i) change the number of generations between two cases; (ii) change the conditional report ratio  $\rho$ ; (iii) change the time of infection; (iv) change the infector of a case (if the case is not the ancestor of a tree); (v) swap infector–infectee (if none is the ancestor of a tree).

We added two proposals to change  $a$  and  $b$ , the spatial kernel parameters. For each proposal, the probability of transmission between every geographical unit was recalculated with the new values. The distance matrix had to be computed for each number of generations between cases, which considerably slowed down the algorithm. As we could not use sequence data, assessing whether a case was isolated or whether it was connected to a reported infector with two missing generations would be very challenging using our model alone. Therefore, we limited the maximal number of missing generations to 1 when  $a$  or  $b$  were estimated ( $\max(\kappa_{ji}) = 2$ ). Finally, the last proposal was designed to change the ancestor of the tree while conserving the overall number of trees (figure 1).

### 2.1.3. Inference of importation status and cluster

Unrelated measles cases stemming from different importations and different regions can be part of the same dataset. Grouping cases and excluding unrealistic transmission links reduces the number of possible trees and speeds up the MCMC runs. To do so, we listed each case's potential infectors using three criteria: (i) the potential infectors must be of the same genotype as the case, or have unreported genotype, (ii) the location of potential infectors must be less than  $\gamma$  km away from the case, and (iii) the potential infectors must have been reported later than  $\delta$  days before the case. This threshold should be determined from the maximum plausible generation time of the disease. The spatial threshold  $\gamma$  should be defined according to the relevance of long-distance transmissions. Cases with no potential infector were considered as importations. Otherwise, they were grouped together with (i) their potential infectors and (ii) cases with common potential infectors.

After grouping the cases, we estimated their importation status and the cluster size distribution using two runs of MCMC (figure 2). The first run was shorter and aimed at removing the most unlikely connections among each group, as they can reflect unrealistic estimates for incubation periods or generation times and corrupt the estimation of the date of infection. We defined a reference threshold  $\lambda$ , whereby if the individual value of log-likelihood  $L_i$  was worse than  $\lambda$ , then the connection between  $i$  and their infector was considered unlikely. In *Outbreaker2*,  $\lambda$  was a



**Figure 1.** Example of the change of ancestors. (a) The initial tree and (b) the new tree proposed after the movement. Initially, there are two ancestors (cases 1 and 2) in a group of nine cases. Cases 3 and 7 have different genotypes and cannot be part of the same tree, the genotypes of the other cases are not reported. The date of infection is in increasing order (1 is the first case, 9 is the last). Therefore, 1 is the only potential infector for 2. One new ancestor was randomly drawn to conserve the number of trees. In this example, 7 is the new ancestor (6 was the only other possibility). The ratio of the posterior densities of (a,b) were then used to determine whether to accept or reject the proposal, according to the Metropolis–Hastings algorithm. This movement ensures good mixing of the potential ancestors of the transmission clusters.

relative value, defined from a quantile of the individual log-likelihoods. In *o2geosocial*,  $\lambda$  can be a relative value or an absolute value, chosen from the number of components of the likelihood. For each sample saved from the short run, we computed the number of unlikely connections  $n$ . If there was no iteration where all connections were better than  $\lambda$ ,  $\min(n)$  new importations were added to the initial tree for the long run (figure 2).

Finally, we ran a long MCMC chain and obtained samples from the posterior distribution. After removing the burn-in period and thinning the chain, we deleted the unlikely transmission links in each iteration and identified transmission clusters. Therefore, unlike the previous versions of *outbreaker2*, the number of importations in each sample can vary and the individual probability of being an importation can be computed (figure 2).

## 2.2. Validation case study: measles outbreaks in the USA between 2001 and 2016

### 2.2.1. Data

To evaluate the performance of the model, we inferred the transmission clusters from a dataset that also included information on whether measles cases were part of a cluster based on contact-tracing investigations. Measles cases in the USA are reported by healthcare providers and clinical laboratories to their corresponding health department. Each case is investigated by local and state health departments classified according to standard case definitions [46], and linked into clusters epidemiologically (e.g. by establishing a direct contact or a shared location between cases, or when cases are part of a specific community where an outbreak is occurring). Cases are considered internationally imported if at least part of the exposure period (7–21 days before rash onset) occurred outside the USA and rash occurred within 21 days of entry into the USA, with no known exposure to measles in the USA during the exposure period.

Confirmed measles cases are routinely reported by state health departments to the CDC. A total of 2098 measles cases were reported in the USA between January 2001 and December 2016. The number of annual cases did not exceed 700 cases during this time period (figure 3; electronic supplementary material, figure S1). The importation status, 5-year age group, onset date, county and state of residence were fully reported for 2077 cases. The 21 cases with missing data were discarded. Twenty-five per cent of the cases were classified as importations. Thirty-nine per cent of the cases had their genotype reported.

Among cases with complete data, 737 independent clusters, containing 1–380 cases, were reconstructed through contact-tracing investigations. Not every identified case could be linked to an importation, and some transmission clusters contained multiple imported cases (e.g. when related individuals travel together to a foreign country and were infected there). Out of the 737 reference clusters, 38 had several cases classified as importations, 256 had none identified.

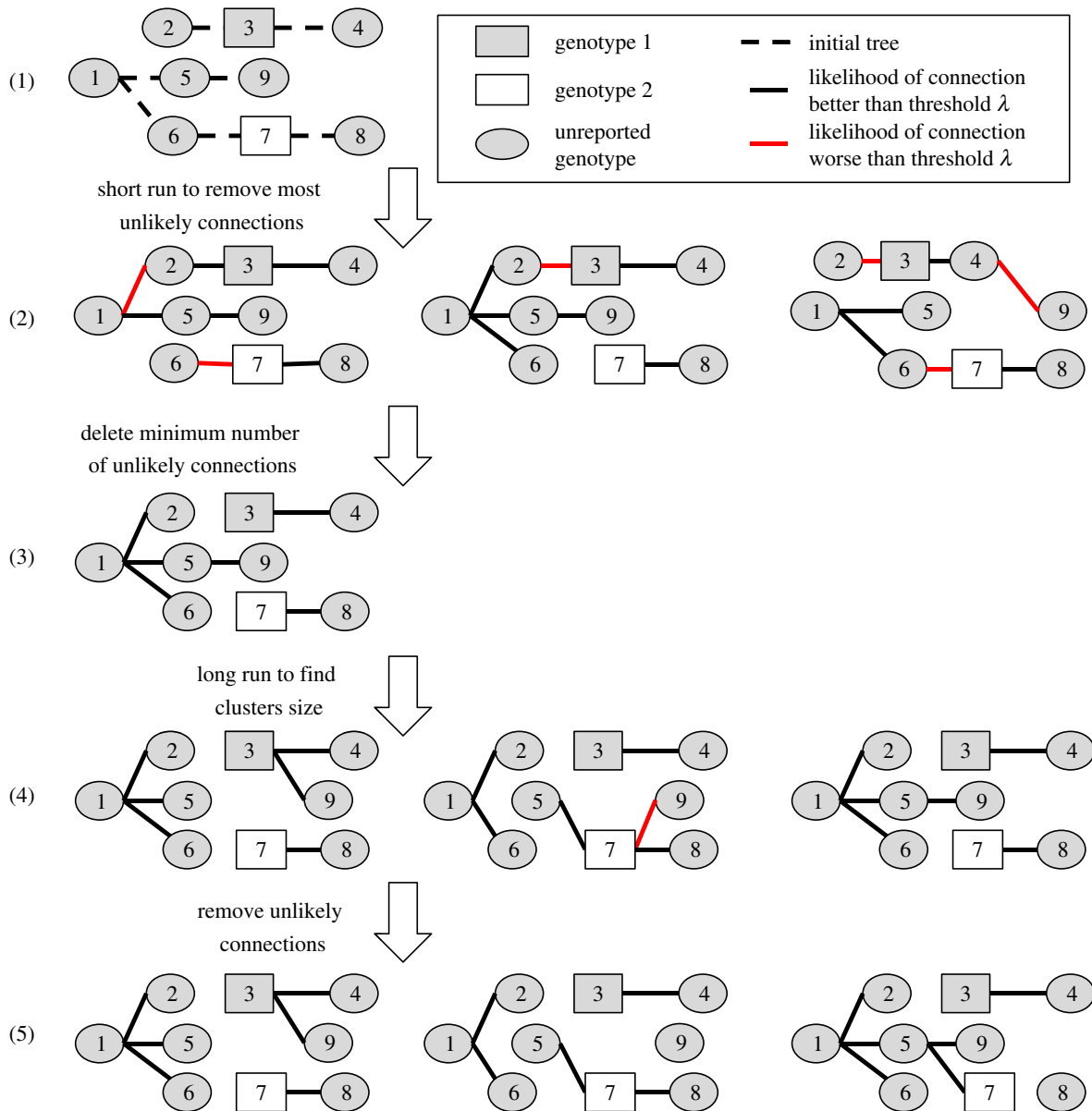
### 2.2.2. Model and parameters

The distributions and priors used in the studies are listed in table 2. As no studies quantifying the probability of age-specific contacts have been carried out in the USA, we used the estimates from the POLYMOD study in the UK [36]. The incubation period and the generation time of measles were taken from previous studies [47–49]. We used the population centroid of each county to compute the distance matrix [50]. We used a beta distribution as the prior of the conditional report ratio [8]. The mean of the prior distribution was calculated using the number of clusters whose first case was not classified as an imported case, meaning the investigations were not able to trace back to the first case imported. As there was no prior information on the possible values of the spatial parameters  $a$  and  $b$ , we used uniform distributions between 0 and 5.

For pre-clustering of cases, we set the temporal threshold  $\delta$  to 30 days, which is above the 97.5% upper quantile of the generation time with a missing generation. We were interested in local transmission to describe the impact of an imported case on a community. But we only had information on the county of residency for each case. Counties are large geographical units: the average county land area is 2911 km<sup>2</sup> and the maximum values reach 50 000 km<sup>2</sup>. Therefore, we set the spatial threshold  $\gamma$  to 100 km to exclude long-distance transmission, while still allowing for cross-county transmission.

Finally, we tested several relative and absolute importation thresholds  $\lambda$ . Absolute values were calculated from a factor  $k$ , multiplied by the number of components in  $L_i$ , excluding the binary genetic component. Tested values were  $k=0.05$  ( $\lambda = \log(0.05)*5 = -15$ ) and  $k=0.1$  ( $\lambda = -11$ ). Connections were considered unlikely if the log-likelihood was worse than  $\lambda$ . Relative values were quantiles of all recorded log-likelihoods in the sampled trees (table 2).





**Figure 2.** Estimating importation status and cluster size distributions in two MCMC runs. Step 1: initial tree obtained after pre-clustering, with the minimum number of importations (here 2, as there are two reported genotypes). Step 2: samples from the first short run, with red lines showing connections worse than the arbitrary threshold  $\lambda$ . Step 3: initial tree for the final run, with one more importation than in step 1, which corresponds to the minimum number of unlikely transmissions at step 2. Step 4: samples from the long run. Step 5: final trees used to compute cluster size distribution and importation status of each case. Case 7 is an importation in one-third of the final samples, whereas case 3 is an importation in all of them.

### 2.2.3. Inference of importation status

Using the contact-tracing investigations, we considered three different initial distributions of the importation status. In scenario 1, there was no inference of the importation status of cases, and the first case of each epidemiological cluster was classified as importation (ideal importation). In scenario 2: there was no inference of the importation status of cases, and all cases identified as importation in the contact-tracing investigations were classified as importations (epidemiological importation). Finally, in scenario 3, the importation status of cases was inferred, using different thresholds  $\lambda$ , and using no prior information on the importation status of cases or the importation status from the contact-tracing investigations.

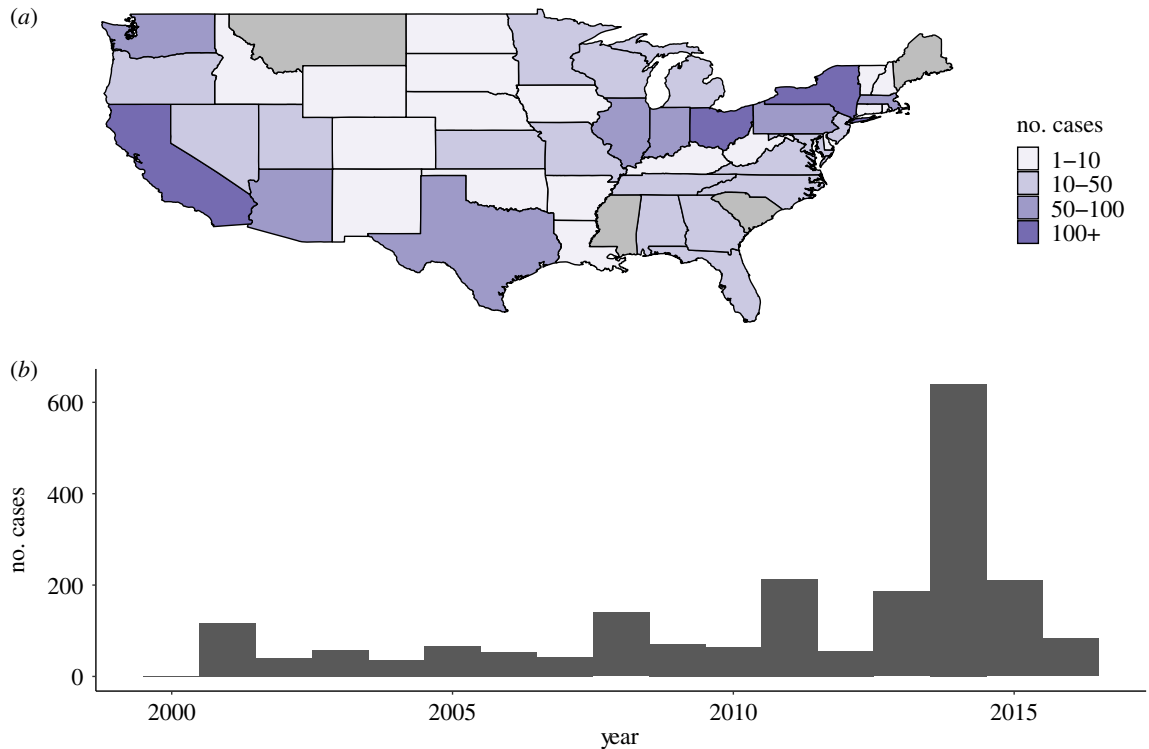
### 2.2.4. Inference of clusters

In order to compare the inferred and reference clusters, we calculated for each case  $i$ : (i) the proportion of cases from the same reference cluster as  $i$  that were inferred with  $i$  (sensitivity) and (ii) the proportion of cases in the same inferred cluster as  $i$  that were part of the reference cluster (precision). These values

were calculated at every iteration, and the median values were used to evaluate the fit obtained with different values of  $\lambda$ . We also compared the inferred cluster size distribution to the reference data. The credibility intervals for each case are reported in electronic supplementary material, figure S2.

## 3. Results

We clustered 2077 measles cases reported in the USA between January 2001 and December 2016 using their onset date, age groups, location and genotype. Using the contact-tracing investigations, we considered three different initial importation status distributions: (i) only the ancestors of each epidemiological cluster (first case of each cluster) were importations (ideal importation), (ii) all cases classified as importation in the contact-tracing investigations were importations (epidemiological importation), (iii) no prior information on importation status of cases. The importation



**Figure 3.** (a) Number of cases per state and (b) annual number of cases reported in the USA between 2001 and 2016. Alaska and Hawaii are not shown in (a).

**Table 2.** Values of parameters used to cluster cases declared in the USA.

parameter	symbol	distribution
incubation period	$f(t)$	gamma, mean = 11.5, s.d. = 2.24
generation time	$w(t)$	normal, mean = 11.7, s.d. = 2.0
conditional report ratio	$\rho$	prior: beta distribution, mean = 0.65, s.d. = 0.15
spatial parameter 1	$a$	prior: uniform distribution
spatial parameter 2	$b$	prior: uniform distribution
spatial pre-clustering	$\gamma$	fixed: 100 km
temporal pre-clustering	$\delta$	fixed: 30 days
importation threshold	$\lambda$	absolute: $5 \times \log 0.05 = -15$ $5 \times \log 0.1 = -11$ relative: 5%

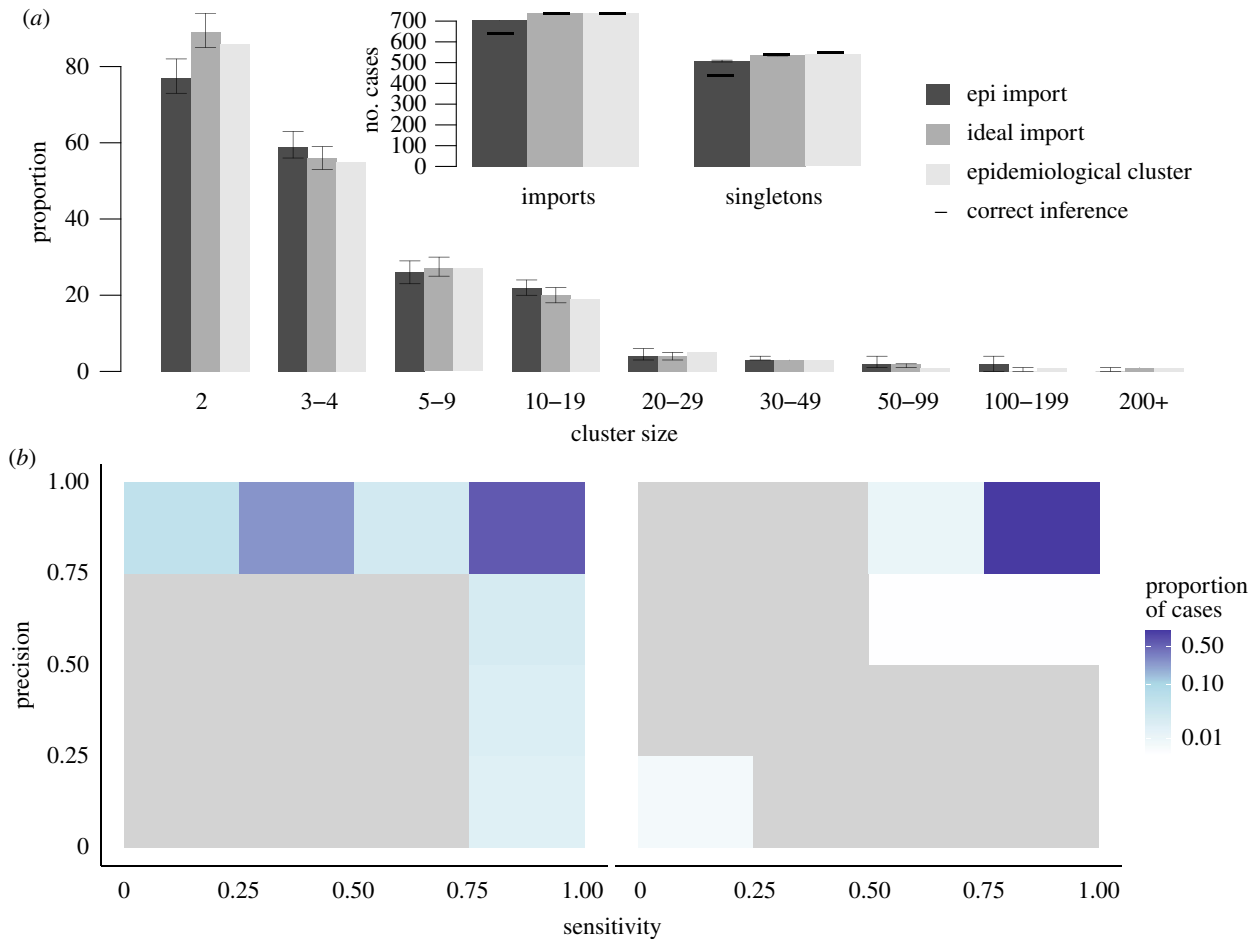
status of the cases was, therefore, not probabilistically inferred in scenarios 1 and 2. The length of the short preliminary run was 30 000 iterations and the main run was 70 000 iterations. For each run, the trace of the posterior distribution shows the convergence of the algorithm (electronic supplementary material, figure S3).

In scenario 1, we did not infer the importation status of cases. The inferred cluster size distribution matched the contact-tracing investigations (figure 4a); 98% of the reference singletons were also isolated in the inferred cluster. For 94% (95% credibility interval: 91–98%) of cases, the inferred cluster

had a sensitivity and precision above 75%, meaning more than 75% of the cases in the inferred cluster were in the reference cluster, and more than 75% of the cases in the reference cluster were in the inferred cluster (figure 4b). For 80% (78–93%) of cases, the inferred clusters were a perfect match with the reference clusters. The cluster size distribution stratified by state was similar to the contact-tracing investigations (electronic supplementary material, figure S4). Therefore, when each ancestor was considered as an importation, the inferred clusters were very close to the reference ones.

In scenario 2, we used the importation status distribution of cases reported in the contact-tracing investigations (539 importations). Pre-clustering highlighted 165 cases with no potential infector, which were also classified as importations. We observed discrepancies between the inferred cluster size distribution and the reference one: among the 704 cases inferred as importation, 61 (9%) were not importations in the reference cluster. Furthermore, 94 cases were the ancestor of a reference cluster and were not classified as importations in the inferred clusters (13%). The overall cluster size distribution matched the reference distribution, but 111 reference singletons were inferred as part of transmission clusters (figure 4a; electronic supplementary material, figure S5). Although the precision of the inferred cluster was above 75% for 93% (88–93%) of the cases, 31% (6–39%) had a sensitivity score below 0.5, meaning they were classified with less than half of the cases from their reference clusters (figure 4c). The discrepancies observed in this scenario are due to inconsistencies between the importation status distribution and the clustering of cases in the contact-tracing investigations, as reference clusters that gathered several importations were split into different inferred clusters.

In scenario 3, we used different threshold  $\lambda$  to infer the importation status of cases. We tested  $\lambda = -15$ ,  $\lambda = -10$  (absolute value), and  $\lambda = 95$ th centile of all recorded log-likelihoods (relative value). For each case  $i$ , if the log-likelihood  $L_i$  was



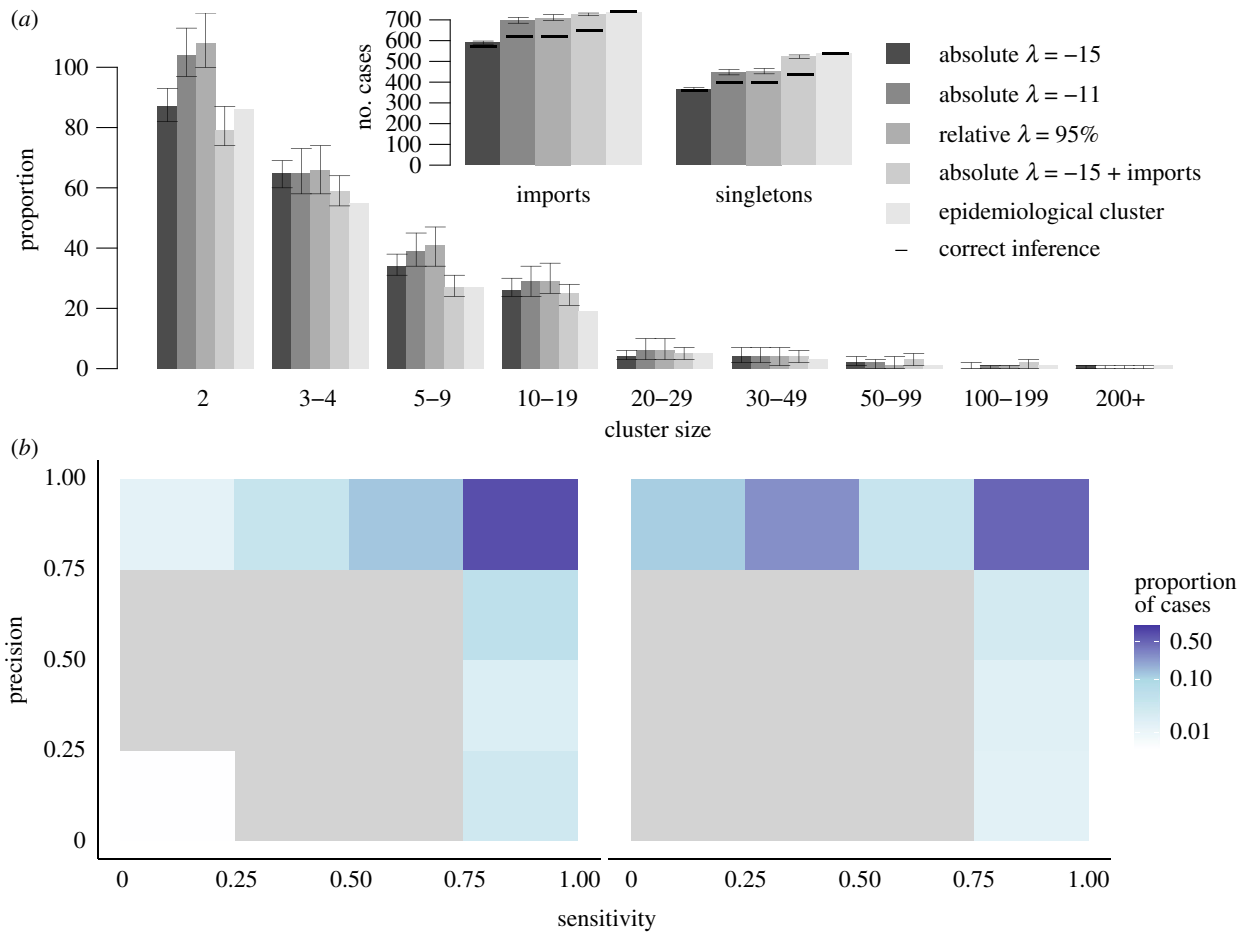
**Figure 4.** Description of transmission clusters inferred using prior knowledge on importation status of cases. (a) Cluster size distribution for scenarios 1 and 2 (grey and dark grey), compared to the reference clusters (light grey). Arrows represent the 95% credibility intervals of each estimate. Only clusters containing at least two cases are represented. Inset: Number of importations and number of isolated cases (singletons) in scenarios 1 and 2, and in the reference clusters. For each scenario, the horizontal dark line represents the number of importations that are also importations in the reference clusters, same for singletons. (b) Heatmap representing the precision and sensitivity of the clusters for each case in scenario 1, cases are classified in a category depending on the proportion of their reference cluster that were inferred in the same cluster ( $x$ -axis) and the proportion of mismatches in the inferred cluster. The same for scenario 2.

worse than  $\lambda$ , the connection between the case and its infector was removed and the case was considered imported. Firstly, using an absolute factor  $\lambda = -15$ , 586 (581–593) cases were classified as importations, and 361 (355–369) of them were singletons. These numbers are much lower than the reference dataset that contains 737 clusters, and 539 singletons (figure 5a; electronic supplementary material, figure S6). However, very few cases inferred as importations or singletons were not classified as such in the reference dataset (15 (10–22) misclassified importations, 4 (0–14) misclassified singletons), and the cluster size distribution for clusters including two cases and more was very similar to the reference one. The precision of the reconstructed cluster was very high (above 75% for 88% (85–93%) of cases) (figure 5b). Overall, the algorithm was not able to accurately identify importations and singletons as the threshold was too low to eliminate some unrealistic connections, but the inferred larger clusters matched their reference counterparts.

We then observed the impact of increasing  $\lambda$  on the inferred cluster size distribution. Runs obtained using an absolute threshold with  $\lambda = -11$  and 95% relative threshold yielded very similar results. The number of cases inferred as importations was higher than in previous runs, while all remaining links showed good connection between cases. The number of importations was closer to the reference dataset, and the number of

singletons was greater than the reference. Nevertheless, 11% (10–12%) of the inferred importations was not classified as importation in the reference clusters. Furthermore, the number of two-case chains was overestimated, and bigger clusters were likely to be split because of the removal of weaker connections. Therefore, increasing  $\lambda$  did not improve the cluster size distribution, as many importations in the reference clusters were not identified and the number of mismatches increased (electronic supplementary material, figure S7).

Finally, we combined prior information and inference of importation status to create a scenario where the importation status of only a proportion of the cases is known, because of disparities in the contact-tracing investigations. This scenario is relevant for a dataset combining different outbreaks scattered across a large area or a long period of time. Cases considered as importations in the contact-tracing investigations were set as importations, and we inferred the importation status of the remaining cases. We used a low threshold to remove the least likely transmission links ( $\lambda = -15$ ). Including prior information led to some misclassification of importation status due to the inconsistencies between the epidemiological importation status and the reference clusters. As in scenario 2, some cases were classified with only part of their reference clusters because clusters with several importations were split into different clusters. Indeed, the sensitivity



**Figure 5.** Description of transmission clusters generated with inferred importation status of cases. (a) Cluster size distribution for different value of threshold in scenario 3 (sorted by shades of grey), compared to the reference clusters (light grey). Arrows represent the 95% credibility intervals of each estimate. Only clusters containing at least two cases are represented. Inset: Number of importations and number of isolated cases (singletons). For each scenario, the horizontal dark line represents the number of importations that are also importations in the reference clusters, same for singletons. (b) Heatmap representing the precision and sensitivity of the clusters for each case in scenario 3, with a 5% relative threshold, cases are classified in a category depending on the proportion of their reference cluster that were inferred in the same cluster. (c) Same when importation status is taken from the contact-tracing investigations and inferred using a 5% relative threshold.

score of 34% (7–51%) of cases was below 0.5. Nevertheless, the cluster size distribution observed in the simulation was the closest to the reference clusters. There were 725 (719–731) clusters, 89% of importations were also ancestors of reference clusters and the number of singletons matched the reference clusters (figure 5a–c). The inferred clusters of 88% (86–94%) of the cases had a precision score of 1, showing they were clustered without any false positives. Despite discrepancies in several states (Massachusetts, Ohio), the cluster size distribution stratified by state showed good agreement with the reference clusters (electronic supplementary material, figure S8).

The conditional report ratio in the transmission chains  $\rho$  and the spatial parameters  $a$  and  $b$  was estimated in each scenario. The parameter estimates did not depend on the prior importation status distribution or the value of  $\lambda$ .  $\rho$  was consistently estimated above 90%, showing a low number of missing generations between cases (electronic supplementary material, figure S9). High values of  $\rho$  show that most of the reported cases could be connected without missing generations. This is not representative of the overall report ratio, which is usually much lower [51].

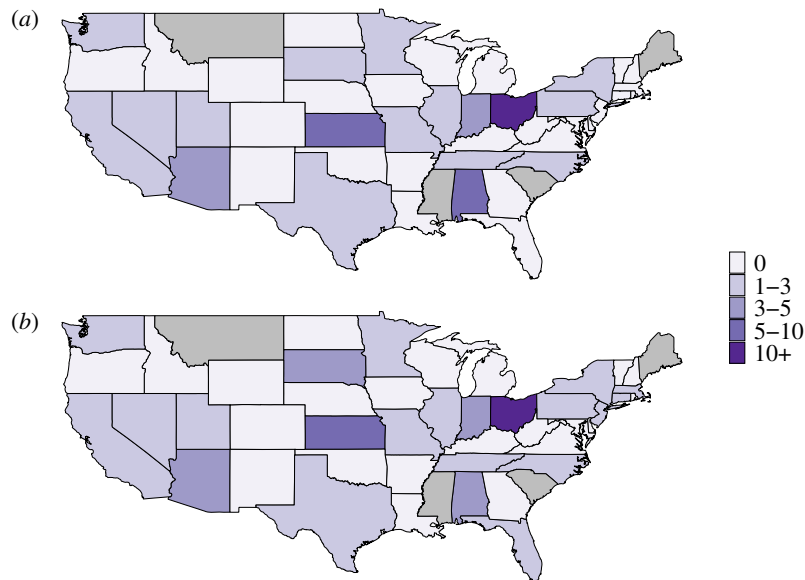
There was little variation in the estimates of the spatial parameters between the different scenarios. The population parameter  $a$  was estimated between 0.6 and 1 for every scenario, and the distance parameter  $b$  was between 0.08 and 0.12. In every scenario, more than 80% of the inferred transmission

were between cases distant of less than 10 km, and few long-distance transmissions were recorded (50–100 km); hence, although most of the reconstructed connections were between cases from the same county, the algorithm was able to identify clusters spreading over several counties or states (electronic supplementary material, figure S10).

We highlighted the added value of including the spatial distance between cases in the likelihood by comparing the cluster size distribution inferred by selecting certain components of  $L_i$  (electronic supplementary material, figure S11). The credibility intervals were much wider when the distance between cases is not part of the likelihood, and the number of chains containing 2–10 cases was overestimated. The important impact of the spatial component of likelihood was also due to the widespread American territory, and could be lower in a different setting.

We used the ratio of the number of importations over the number of subsequent cases per state to evaluate the intensity of transmission in each state between 2001 and 2016 (figure 6). The maps obtained in scenario 1 (ideal scenario) or in scenario 3 (estimation of importation, with epidemiological importations and  $\lambda = -15$ ) were very similar. We only observed minor differences, for example, in South Dakota and in Massachusetts, where the ratios were higher in scenario 3. The highest ratio (31.8 in scenario 1) was observed in Ohio, and is mostly due to a 383 case outbreak in 2014





**Figure 6.** Ratio of the number of importations over the number of subsequent cases in each state in (a) scenario 1 (ideal importations) and (b) scenario 3 with epidemiological importations and  $\lambda = -15$ . Grey states represent states that did not report any case.

[32]. We observed major differences between the incidence map (figure 2a) and the ratio per state. Indeed, although 403 cases were reported in California (highest number in the USA), importations caused on average 1.32 subsequent cases in scenario 1 (1.60 in scenario 3), showing a high proportion of reported cases were inferred as importations.

Similarly, we used the inferred transmission chain to compute the inferred reproduction number in each state. According to the model, about 60% cases did not cause future transmission, and about 5% caused more than five subsequent cases (electronic supplementary material, figure S12). These numbers were consistent in each run. The geographical distribution of reproduction number was very similar to the importation–subsequent cases ratio (electronic supplementary material, figure S13).

## 4. Discussion

We developed the R package *o2geosocial* to classify measles cases into transmission clusters and estimate their importation status using routinely collected surveillance data (genotype, age, onset date and location of the cases). As recently observed during the 2018–2019 measles outbreak in New York, delays in childhood vaccination, local susceptibility and increased contacts can lead to large outbreaks following importations [52,53]. Therefore, we were interested in highlighting the effect of imported cases on communities and we focused on short distance transmission to identify areas where they repeatedly caused subsequent transmission chains. Although this is not predictive of future transmission, it highlights communities with potential for large transmission clusters.

We compared the inferred transmission clusters to the contact-tracing investigations of 2077 confirmed measles cases reported in the USA between 2001 and 2016. We were able to produce reliable estimates of known transmission clusters using epidemiological features with only few misclassifications. Estimating the importation status of cases without prior knowledge was challenging and caused uncertainty on the results. We tested different threshold  $\lambda$  to eliminate unlikely transmissions, and we were able to identify most of the imported cases.

Nevertheless, if several cases were imported in the same region at a similar time, we could not find all of them without discarding valid transmission events, and increasing the number of false positives. When we used the importation status as defined in the contact-tracing investigations without probabilistic inference (scenarios 1 and 2), the reconstructed clusters were similar to the reference ones. Results were also conclusive when we combined prior information and importation inference. The reconstruction of transmission greatly depends on the epidemiological investigations to identify measles importations in a community.

We used the genotype to censor connections between cases when it was reported, as there can be only one reported genotype per transmission cluster. Using a simulated dataset (*toy\_outbreak\_long* in *o2geosocial*), we explored the impact of increasing the proportion of genotyped cases on clustering and observed it could help identify the number of concurrent transmission trees when multiple genotypes are co-circulating. Moreover, we introduced a spatial component to the likelihood of connection between cases using an exponential gravity model. Previous studies showed this model was able to capture short-distance dynamics better than other gravity models, and was easy to parametrize. Introducing the spatial component greatly improved the precision and the sensitivity of the reconstructed clusters (electronic supplementary material, figure S11), and the parameter estimates were robust in the different scenarios.

The final results on the clustering of the 2077 cases using *o2geosocial* were obtained in 7 h for each run of 100 000 iterations on a standard desktop computer (Intel Core i7, 3.20 GHz 6 cores), which is much faster than previous implementations of *outbreaker* and *outbreaker2*. With the addition of the pre-clustering step, whereby we reduced the number of potential infectors for each case, the algorithm ran faster. For smaller chains (50 000 iterations), 4 h were needed to estimate the importation status and cluster the cases. The code for the package and the analysis developed in this project is shared on Github (<https://github.com/alxsrobert/o2geosocial> and <https://github.com/alxsrobert/datapaperMO>), with an illustrative toy dataset, and can be used to analyse recent outbreaks where contact-tracing investigations were not carried out.

Although the results obtained are promising, it should be noted that the dynamics of measles transmission in the USA are likely to be very specific to this location. Indeed, there were less than 700 annual cases between 2001 and 2016. These cases were scattered across a large area, which made the pre-clustering of cases very efficient as we focused on short-distance transmission. In smaller or more endemic settings, the number of potential infectors per cases after the pre-clustering step might be higher, which would increase the running time.

Furthermore, as the location of each case was deduced from the population centroid of counties, we assumed that the distance between cases from the same county was effectively zero. American counties are large and widespread geographical units that can include more than 1 million individuals. For future use of *o2geosocial*, more accurate information on the location of cases could improve cluster inference by identifying multiple importations in a given county. Because cases are reported by the state of residency, we had to ignore that cases may have been out of the reported county or state during their incubation and infectious period, which has been seen during some outbreaks, such as the 2015 'Disney outbreak' in California [54].

We did not include prior information on the local susceptibility of the different areas affected in *o2geosocial*, and these could be estimated using historical values of local coverage. However, protocols to estimate local vaccination coverage can differ in time and space and be difficult to compare, or unavailable at the local level. Furthermore, these estimates are cross-sectional in nature, and might not take into account catch-up vaccination campaigns, or immunity induced by previous outbreaks. Local seroprevalence surveys could identify pockets of susceptibles, but they have not been carried out on a subnational scale in most countries [55].

There has been no national quantitative analysis of age-specific contact patterns carried out in the USA, so we relied on a contact matrix between age groups available for Great Britain from the POLYMOD study [36]. Nevertheless, little variation in the contact rates between age groups has been observed between European countries, and a previous projection of the social contact matrix in the USA yielded similar results [56]. POLYMOD data were probably the most reliable source of information we could use to deduce an estimate of the contact matrix in the USA.

## 5. Conclusion

Heterogeneity in immunity can cause large outbreaks in countries with high national vaccine coverage, and identifying

potential foyers of transmission in post-elimination settings is key for outbreak prevention and control. We have presented a method for estimating the cluster size distribution of past measles outbreaks from routinely collected surveillance data. We found that adding prior knowledge on the importation status of cases improved the inference of the transmission clusters. Although the method was able to identify a proportion of importations, epidemiological investigations on the history of travel and exposure reduced uncertainty on the clustering of cases. We believe these investigations are needed to produce reliable estimates of past transmission clusters. In lieu of the importation status, if multiple genotypes are co-circulating, increasing the proportion of genotyped cases could help discard potential connections and find imported cases. Even with limited information, this method was able to infer probabilistic transmission clusters in a fast and efficient way.

**Ethics.** This study is a secondary analysis of data collected as part of routine surveillance of measles outbreaks in the USA. The research was approved by the London School of Hygiene and Tropical Medicine Research Ethics Committee (reference number 15735).

**Data accessibility.** The package we developed is publicly available on Github (<https://github.com/alxsrobert/o2geosocial>), along with the code used to analyse the data and generate the figures (<https://github.com/alxsrobert/datapaperMO>). Combinations of variables in the surveillance data used to validate this algorithm may contain sensitive personally identifiable health information which are subject to the Privacy Act and cannot be shared publicly. A toy dataset was attached to the *o2geosocial* package (in *o2geosocial/data*). The script *analysis\_generated\_data.R* in the *datapaperMO* repository generates toy datasets with different parameters (distance kernel, number of cases, reproduction numbers etc.) and can be used to re-run the model and test its performance.

**Authors' contributions.** A.R., S.F. and A.J.K. developed the method and the analysis plan. P.A.G. and P.P. provided data for the study and gave feedback on the analysis plan. A.R. implemented the analysis, wrote the code and ran the model. A.R. interpreted the results, with contributions from S.F., A.J.K. and P.A.G. A.R. wrote the first draft and the supplementary material. A.R., S.F., A.J.K., P.A.G. and P.P. contributed to the manuscript, all authors approved the final version.

**Competing interests.** We declare we have no competing interests.

**Funding.** A.R. was supported by the Medical Research Council (MR/N013638/1). S.F. was supported by a Wellcome Trust Senior Research Fellowship in Basic Biomedical Science (210758/Z/18/Z). A.J.K. was supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (206250/Z/17/Z).

**Acknowledgements.** We acknowledge Thibaut Jombart for technical support and feedback on the analysis plan.

**Disclaimer.** The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention, US Department of Health and Human Services.

## References

1. Ferguson NM, Donnelly CA, Anderson RM. 2001 Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* **413**, 542–548. (doi:10.1038/35097116)
2. Wallinga J, Teunis P. 2004 Different epidemic curves for severe acute respiratory syndrome reveal. *Am. J. Epidemiol.* **160**, 509–516.
3. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005 Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359. (doi:10.1038/nature04153)
4. Faye O, Boëlle P-Y, *et al.* 2015 Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. *Lancet Infect. Dis.* **15**, 320–326. (doi:10.1016/S1473-3099(14)71075-8)
5. Ypma RJF, van Ballegoijen WM, Wallinga J. 2013 Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195**, 1055–1062. (doi:10.1534/genetics.113.154856)
6. Wallinga J, Lipsitch M. 2007 How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604. (doi:10.1098/rspb.2006.3754)
7. Cauchemez S, Ferguson NM. 2012 Methods to infer transmission risk factors in complex outbreak data.

- J. R. Soc. Interface* **9**, 456–469. (doi:10.1098/rsif.2011.0379)
8. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. 2014 Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10**, e1003457. (doi:10.1371/journal.pcbi.1003457)
  9. Campbell F, Cori A, Ferguson N, Jombart T. 2019 Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput. Biol.* **15**, e1006930. (doi:10.1371/journal.pcbi.1006930)
  10. Haydon DT, Chase-Topping M, Shaw DJ, Matthews L, Friar JK, Wilesmith J, Woolhouse MEJ. 2003 The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proc. R. Soc. B* **270**, 121–127. (doi:10.1098/rspb.2002.2191)
  11. Cauchemez S, Boëlle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, Hedley AJ, Anderson RM, Valleron A-J. 2006 Real-time estimates in early detection of SARS. *Emerg. Infect. Dis.* **12**, 110–113. (doi:10.3201/eid1201.050593)
  12. Heijne JCM, Rondy M, Verhoef L, Wallinga J, Kretzschmar M, Low N, Koopmans M, Teunis PFM. 2012 Quantifying transmission of norovirus during an outbreak. *Epidemiology* **23**, 277–284. (doi:10.1097/EDE.0b013e3182456ee6)
  13. Kendall M, Ayabina D, Colijn C. 2016 Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees. *Stat. Sci.* **33**, 70–85. (doi:10.1214/17-STS637)
  14. Worby CJ, O'Neill PD, Kyraios T, Robotham JV, De Angelis D, Cartwright EJP, Peacock SJ, Cooper BS. 2016 Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Ann. Appl. Stat.* **10**, 395–417. (doi:10.1214/15-AOAS898)
  15. Lau MSY, Marion G, Streftaris G, Gibson G. 2015 A systematic Bayesian integration of epidemiological and genetic data. *PLoS Comput. Biol.* **11**, e1004633. (doi:10.1371/journal.pcbi.1004633)
  16. Spada E, Sagliocca L, Sourdis J, Garbuglia AR, Poggi V, De Fusco C, Mele A. 2004 Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *J. Clin. Microbiol.* **42**, 4230–4236. (doi:10.1128/JCM.42.9.4230-4236.2004)
  17. Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, Soubeyrand S. 2014 A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B* **281**, 20133251. (doi:10.1098/rspb.2013.3251)
  18. Gire SK *et al.* 2014 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372. (doi:10.1126/science.1259657)
  19. Carroll MW *et al.* 2015 Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature* **524**, 97–101. (doi:10.1038/nature14594)
  20. Ruan YJ *et al.* 2003 Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* **361**, 1779–1785. (doi:10.1016/S0140-6736(03)13414-9)
  21. Pybus OG, Rambaut A. 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550. (doi:10.1038/nrg2583)
  22. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Edward C. 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332. (doi:10.1126/science.1090727)
  23. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. 2018 When are pathogen genome sequences informative of transmission events? *PLoS Pathog.* **14**, e1006885. (doi:10.1371/journal.ppat.1006885)
  24. Rota PA *et al.* 2011 Global distribution of measles genotypes and measles molecular epidemiology. *J. Infect. Dis.* **204**, S514–S523. (doi:10.1093/infdis/jir118)
  25. Hiebert J, Severini A. 2014 Measles molecular epidemiology: what does it tell us and why is it important? *Can. Commun. Dis. Rep.* **40**, 257–260.
  26. Brown KE, Rota PA, Goodson JL, Williams D, Abernathy E, Takeda M, Mulders MN. 2019 Genetic characterization of measles and rubella viruses detected through global measles and rubella elimination surveillance, 2016–2018. *Morb. Mortal. Wkly. Rep.* **68**, 587–591. (doi:10.15585/mmwr.mm6826a3)
  27. Gardy JL *et al.* 2015 Whole-genome sequencing of measles virus genotypes H1 and D8 during outbreaks of infection following the 2010 Olympic Winter Games reveals viral transmission routes. *J. Infect. Dis.* **212**, 1574–1578. (doi:10.1093/infdis/jiv271)
  28. Penados AR, Myers R, Hadeb B, Aladin F, Brown KE. 2015 Assessment of the utility of whole genome sequencing of measles virus in the characterisation of outbreaks. *PLoS ONE* **10**, 1–16. (doi:10.1371/journal.pone.0143081)
  29. World Health Organisation. 2012 Measles virus nomenclature update: 2012. *Wkly. Epidemiol. Rec.* **87**, 73–80. (doi:10.1016/j.actatropica.2012.04.013)
  30. Hagemann C, Streng A, Kraemer A, Liese JG. 2017 Heterogeneity in coverage for measles and varicella vaccination in toddlers—analysis of factors influencing parental acceptance. *BMC Public Health* **17**, 724. (doi:10.1186/s12889-017-4725-6)
  31. Glasser JW, Feng Z, Omer SB, Smith PJ, Rodewald LE. 2016 The effect of heterogeneity in uptake of the measles, mumps, and rubella vaccine on the potential for outbreaks of measles: a modelling study. *Lancet Infect. Dis.* **16**, 599–605. (doi:10.1016/S1473-3099(16)00004-9)
  32. Gastañaduy PA *et al.* 2016 A measles outbreak in an underimmunized Amish community in Ohio. *N. Engl. J. Med.* **375**, 1343–1354. (doi:10.1056/NEJMoa1602295)
  33. Woudenberg T, Van Binnendijk RS, Sanders EAM, Wallinga J, De Melker HE, Ruijs WLM, Hahné SJM. 2017 Large measles epidemic in the Netherlands, May 2013 to March 2014: changing epidemiology. *Eurosurveillance* **22**, 1–9. (doi:10.2807/1560-7917.ES.2017.22.3.30443)
  34. Keenan A, Ghebrehewet S, Vivancos R, Seddon D, MacPherson P, Hungerford D. 2017 Measles outbreaks in the UK, is it when and where, rather than if? A database cohort study of childhood population susceptibility in Liverpool, UK. *BMJ Open* **7**, e014106. (doi:10.1136/bmjopen-2016-014106)
  35. Kucharski AJ, Edmunds WJ. 2015 Characterizing the transmission potential of zoonotic infections from minor outbreaks. *PLoS Comput. Biol.* **11**, 1–17. (doi:10.1371/journal.pcbi.1004154)
  36. Mossong J *et al.* 2008 Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, 0381–0391. (doi:10.1371/journal.pmed.0050074)
  37. Blumberg S, Lloyd-Smith JO. 2013 Inference of R0 and transmission heterogeneity from the size distribution of stuttering chains. *PLoS Comput. Biol.* **9**, 1–17. (doi:10.1371/journal.pcbi.1002993)
  38. Blumberg S, Enanoria WTA, Lloyd-Smith JO, Lietman TM, Porco TC. 2014 Identifying postelimination trends for the introduction and transmissibility of measles in the United States. *Am. J. Epidemiol.* **179**, 1375–1382. (doi:10.1093/aje/kwu068)
  39. Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. 2018 outbreaker2: A modular platform for outbreak reconstruction. *BMC Bioinf.* **19**, 1–8. (doi:10.1186/s12859-018-2330-z)
  40. Lenormand M, Bassolas A, Ramasco JJ. 2016 Systematic comparison of trip distribution laws and models. *J. Transp. Geogr.* **51**, 158–169. (doi:10.1016/j.jtrangeo.2015.12.008)
  41. Zipf GK. 1946 The P1 P2/D hypothesis: on the intercity movement of persons. *Am. Sociol. Rev.* **11**, 677–686. (doi:10.2307/2087063)
  42. Barthélemy M. 2011 Spatial networks. *Phys. Rep.* **499**, 1–79. (doi:10.1016/j.physrep.2010.11.002)
  43. Xia Y, Bjornstad ON, Grenfell BT. 2004 Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *Am. Nat.* **164**, 267–281. (doi:10.1086/422341)
  44. Lenormand M, Huet S, Gargiulo F, Deffuant G. 2012 A universal model of commuting networks. *PLoS ONE* **7**, e45985. (doi:10.1371/journal.pone.0045985)
  45. Andrieu C, De Freitas N, Doucet A, Jordan MI. 2003 An introduction to MCMC for machine learning. *Mach. Learn.* **50**, 5–43. (doi:10.1023/A:1020281327116)
  46. Centers for Disease Control and Prevention (CDC). 2013 National Notifiable Disease Surveillance System: measles/rubeola 2013. See <https://www.cdc.gov/nndss/conditions/measles/case-definition/2013/> (accessed 23 October 2019).
  47. Lessler J, Reich NG, Brookmeyer R, Perl TM, Nelson KE. 2015 Incubation periods of acute respiratory viral infections: a systematic review. *Lancet Infect. Dis.* **9**, 291–300. (doi:10.1016/S1473-3099(09)70069-6)
  48. Klinkenberg D, Nishiura H. 2011 The correlation between infectivity and incubation period of

- measles, estimated from households with two cases. *J. Theor. Biol.* **284**, 52–60. (doi:10.1016/j.jtbi.2011.06.015)
49. Fine PEM. 2003 The interval between successive cases of an infectious disease. *Am. J. Epidemiol.* **158**, 1039–1047. (doi:10.1093/aje/kwg251)
  50. US Census Bureau. 2010 Centers of Population for the 2010 Census 2010. See <https://www.census.gov/geographies/reference-files/2010/geo/2010-centers-population.html> (accessed 22 August 2019).
  51. Woudenberg T, Woonink F, Kerkhof J, Cox K, Ruijs WLM. 2018 The tip of the iceberg: incompleteness of measles reporting during a large outbreak in The Netherlands in 2013–2014. *Epidemiol. Infect.* **146**, 716–722. (doi:10.1017/S0950268818002698)
  52. Gastañaduy PA *et al.* 2018 Impact of public health responses during measles outbreak in an amish community in Ohio: modeling the dynamics of transmission. *Am. J. Epidemiol.* **187**, 2002–2010. (doi:10.1093/aje/kwy082)
  53. Patel M *et al.* 2019 National update on measles cases and outbreaks—United States, January 1–October 1, 2019. *MMWR Morb. Mortal. Wkly. Rep.* **68**, 893–896. (doi:10.15585/mmwr.mm6840e2)
  54. Zipprich J, Winter K, Hacker J, Xia D, Watt J, Harriman K. 2015 Measles outbreak—California, December 2014–February 2015. *Ann. Emerg. Med.* **64**, 82–83. (doi:10.1016/j.annemergmed.2015.04.002)
  55. Durrheim D. 2018 Measles elimination, immunity, serosurveys, and other immunity gap diagnostic tools. *J. Infect. Dis.* **218**, 341–343. (doi:10.1093/infdis/jiy138)
  56. Prem K, Cook AR, Jit M. 2017 Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS Comput. Biol.* **13**, e1005697. (doi:10.1371/journal.pcbi.1005697)