Title: Probabilistic reconstruction of measles transmission clusters from routinely collected surveillance data.

Authors:

Alexis Robert[1,2], Adam J. Kucharski[1,2], Paul A. Gastanaduy[3], Prabasaj Paul[4], Sebastian Funk[1,2]

Authors' affiliations:

1. Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

2. Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

3. Division of Viral Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia

4. Division of Nutrition, Physical Activity, and Obesity, Centers for Disease Control and Prevention, Atlanta, Georgia

Abstract

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. Characterising 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 combined age, location, genotype, and onset date of cases in order to probabilistically reconstruct the importation status and transmission clusters within a newly developed R package called *o2geosocial*.

We compared our inferred cluster size distributions to 737 transmission clusters identified through detailed contact-tracing in the United States 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.

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 investigation, 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.

Wallinga and Teunis 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[2]. 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 N-450 sequence data, 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 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 size distribution of transmission trees resulting from each importation during outbreaks (otherwise known as 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 index 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 United States between 2001 and 2016.

Methods

Presentation of the algorithm

Likelihood function and parameter definition

We used a probabilistic model to infer the individual contribution to the log-likelihood L_i of every case included in the list of cases.

$$L_i(t_i, \mathrm{j}, t_j, \theta) = \log(f(t_i - T_i)) + \mathrm{L_{ji}}(t_i, t_j, \theta) \quad (1)$$

L_i was computed from $\mathrm{L_{ji}}(\theta)$, the log-likelihood of case i being infected by case j as a function of infection times t_i and t_j and model parameters θ, and the timing of t_i, the date of infection of i, relative to the date of symptom onset T_i. We defined $f(t_i - T_i)$ as the probability density of observing T_i if case *i* was infected at time t_i (i.e. *f* represents the distribution of incubation periods). The log-likelihood of transmission $\mathrm{L_{ji}}$ was computed from five components reflecting the age group, genotype, location, and inferred date of infection of cases i and j, and the generation time of the disease:

$$\mathrm{L_{ji}}(t_i, t_j, \theta) = \log\left(p(\kappa_{ji}|\rho) * w^{(\kappa_{ji})}(t_i - t_j) * a^{(\kappa_{ji})}(\alpha_i, \alpha_j) * G(g_i, g_{\tau_j}) * s^{(\kappa_{ji})}(r_i, r_j| \, a, b)\right) \quad (2)$$

We allowed for missing generations between cases due to an unreported individual, and κ_{ji} corresponds to the number of generations between i and j. We calculated the temporal probability of transmission between i and j from the number of days between the dates of infection of the two cases t_i and t_j and the generation time of the disease $w(\mathrm{t})$. This probability of infection was quantified by $w^k(t_i - t_j, \kappa_{ji})$, $w^{(\kappa)} = \prod_\kappa w$, where Π is the convolution operator. We used an exponential distribution $p(\kappa_{ji}|\rho)$ to quantify the probability of observing κ_{ji} missing generation between *i* and *j* from the conditional report ratio ρ which quantifies the probability of missing generation between two connected cases in a cluster. It does not correspond to the overall report ratio of an outbreak as entire missing clusters, or unreported cases infected after the last case or before the ancestor of a cluster are not included in ρ. The "ancestor" is the earliest identified case of a cluster.

$a(\alpha_i, \alpha_j, \kappa_{ji})$ was defined as the probability of transmission between age groups α_i and α_j. This probability corresponds to the proportion of contacts to the age group α_i that originated from α_j and can be deduced from studies such as Polymod[36]. We defined $G\left(g_i, g_{\tau_j}\right)$ as the probability of observing the pathogen genotype g_i in case *i* in the tree τ_j containing case *j*. There can only be one measles virus genotype per transmission tree, or cases with unreported genotype.

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

$s(r_i, r_j, \kappa_{ij})$ was defined as the probability of connection from r_j to r_i, counties of residency of i and j. We used a gravity model to quantify the connectivity of the different geographical units. In the simplest form of the gravity approach, the number of connections between two counties k and l is proportional to the product of the origin population m_k, the destination population m_l and a function of the distance between k and l d_{kl} : $\mathrm{p_{kl}} \propto f\left(d_{kl,}a\right) * \mathrm{m_k^b} * \mathrm{m_l^c}$, with a, b, and c parameters adjusting for the impact of distance and population. From this definition, we deduced $s(k,l)$, the probability of transmission from an individual from region k to another from region l:

$$s(k,l) = \frac{p_{kl}}{\sum_h p_{hl}} = \frac{f\left(d_{kl,}a\right) * \mathrm{m_k^b} * \mathrm{m_l^c}}{\sum_h f\left(d_{hl,}a\right) * \mathrm{m_h^b} * \mathrm{m_l^c}} = \frac{f\left(d_{kl,}a\right) * \mathrm{m_k^b}}{\sum_h f\left(d_{hl,}a\right) * \mathrm{m_h^b}}$$

Only the parameters a and b were required to compute the spatial probability of transmission. We used the exponential gravity model ($f(d_{kl}, a) = e^{-a*d_{kl}}$)[40]. This approach showed good performance at modelling short distance commuting, and was easy to parametrise[40–44].

In order to compute the log-posterior densities of the proposed trees, we summed the individual log-likelihoods and added log-priors on the report ratio ρ, which quantified the percentage of cases in the chains reported to the surveillance system; and the spatial parameters a and b (Table 1).

Tree proposals

We used a Metropolis Hastings algorithm with Markov chain Monte Carlo (MCMC) to sample from the posterior distribution of parameters and the 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 ρ; 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 re calculated with the new values. Depending on the number of geographical units, this calculation considerably slowed down the algorithm. Therefore, when a or b were estimated, we limited the maximal number of missing generations to 1 ($\max(\kappa_{ji}) = 2$). Finally, the last proposal was designed to change the ancestor of the tree whilst conserving the overall number of trees (Figure 1).

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 γ km away from the case and iii) the potential infectors must have been reported later than δ days before the case. This threshold should be determined from the maximum plausible generation time of the disease. The spatial threshold γ 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 λ, whereby if the individual value of log-likelihood $\mathrm{L_i}$ was worse than λ, then the connection between i

and their index was considered unlikely. In *Outbreaker2,* λ was a relative value, defined from a quantile of the individual log-likelihoods. In *o2geosocial*, λ 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 where better than λ, *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).

Validation case study: measles outbreaks in the United States between 2001 and 2016

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 United States 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 United States and rash occurred within 21 days of entry into the United States, with no known exposure to measles in the United States during the exposure period.

Confirmed measles cases are routinely reported by state health departments to the CDC. 2,098 measles cases were reported in the United States between January 2001 and December 2016. The number of annual cases did not exceed 700 cases during this time period (Figure 3, Supplement Figure S1). The importation status, 5-year age group, onset date, county, and state of residence were fully reported for

2,077 cases. The 21 cases with missing data were discarded. 25% of the cases were classified as importations. 39% of the cases had their genotype reported. The dataset of 2,077 cases is referred to as "reference dataset" in the results section, and was used to evaluate the performance of the inference method.

Among cases with complete data, 737 independent clusters, containing 1 to 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.

Model and parameters

The distributions and priors used in the studies are listed in Table 1. As no studies quantifying the probability of age-specific contacts have been carried out in the United States, 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, we used uniform distributions as priors for a and b.

For pre-clustering of cases, we set the temporal threshold δ 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 2,911km^2 and the maximum values reach 50,000km^2. Therefore, we set the spatial threshold γ to 100km to exclude long distance transmission, while still allowing for cross-county transmission.

Finally, we tested several relative and absolute importation thresholds λ. 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* ($\lambda = -15$) and k = *0.1* ($\lambda = -11$). Connections were considered unlikely if the log-likelihood was worse than λ. Relative values were quantiles of all recorded log-likelihoods in the sampled trees (Table 1).

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 λ, and using no prior information on the importation status of cases or the importation status from the contact tracing investigations.

In order to compare the inferred and reference clusters, we calculated for each case i) the proportion of the reference cluster correctly inferred (sensitivity) and ii) the proportion of the inferred cluster that was 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 λ. We also used the inferred cluster size distribution to the reference data. The credibility intervals for each case are reported in the Supplement (Supplement Figure S2).

Results

We clustered 2,077 measles cases reported in the United States 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 distribution: 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 status of the cases was therefore not probabilistically inferred in scenario 1 and 2. The short preliminary run was 30,000 iterations and 70,000 iterations. For each run, the trace of the posterior distribution shows the convergence of the algorithm (Supplement 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 (Supplement 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, Supplement 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 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 2.

In scenario 3, the importation status of cases was inferred from a threshold λ. For each case i, if the log-likelihood L_i was worse than λ, the connection between the case and its index was removed and the case was considered imported. Firstly, using an absolute factor $k = 0.05$ ($\lambda = -15$), 586 (581-593) cases were classified as importations, 361 (355-369) of them were singletons. These numbers are much lower than the reference datasets that contains 737 clusters, and 539 singletons (Figure 5A, Supplement Figure S6). We observed very few misclassifications of importation status and singletons (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 λ on the inferred cluster size distribution. Runs obtained using an absolute threshold with $k = 0.10$ ($\lambda = -11$) 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, the 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 λ did not improve the cluster size distribution, as many importations in the reference clusters were not identified and the number of mismatches increased (Supplement Figures S7).

Finally, we combined prior information and inference of importation status. 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 ($k = 0.05$). 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 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 (Supplement Figures S8).

The conditional report ratio in the transmission chains ρ 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 λ. ρ was consistently estimated above 90%, showing a low number of missing generations between cases (Supplement Figure S9). This number is not representative of the overall report ratio, which is usually much lower[51], and does not take into account missing importations in singletons and chains. High values of ρ show that the reported cases can be connected without missing generations.

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 10km, and few long-distance transmissions were recorded (50-100km), 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 (Supplement 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 (Supplement 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 to 10 cases was over estimated. 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 the scenario 1 (ideal scenario) or in scenario 3 (estimation of importation, with epidemiological importations and $k = 0.05$) 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[32]. We observed major differences between the incidence map (Figure 3A) and the ratio per state. Indeed, although 403 cases were reported in California (highest number in the US), 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 5 subsequent cases (Supplement Figure S12). These numbers were consistent in each run. The geographical distribution of reproduction number was very similar to the importation - subsequent cases ratio (Supplement Figure S13).

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 2,077 confirmed measles cases reported in the United States 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 λ 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 (scenario 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 parametrise. Introducing the spatial component greatly improved the precision and the sensitivity of the reconstructed clusters (Supplement Figure S11), and the parameter estimates were robust in the different scenarios.

The final results on the clustering of the 2,077 cases using *o2geosocial* were obtained in 7 hours 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 hours 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 *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 United States 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 United States, 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 United States yielded similar results[56]. POLYMOD data was probably the most reliable source of information we could use to deduce an estimate of the contact matrix in the United States.

Conclusions

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.

Acknowledgements

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

Funding

AR was supported by the Medical Research Council (MR/N013638/1). SF was supported by a Wellcome Trust Senior Research Fellowship in Basic Biomedical Science (210758/Z/18/Z). AJK was supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (206250/Z/17/Z).

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.

Figure legends

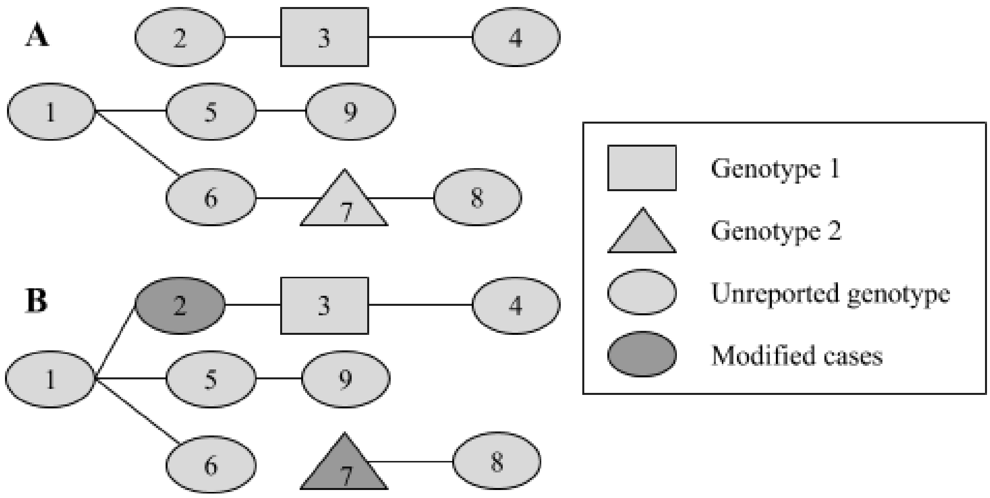


Figure 1: Example of the change of ancestors. Panel A represent the initial tree, B is the new tree proposed after the movement. Initially, there are two ancestors (cases 1 and 2) in a group of 9 cases. 3 and 7 have different genotypes and cannot be part of the same tree, the genotypes of the other cases 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 and 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.

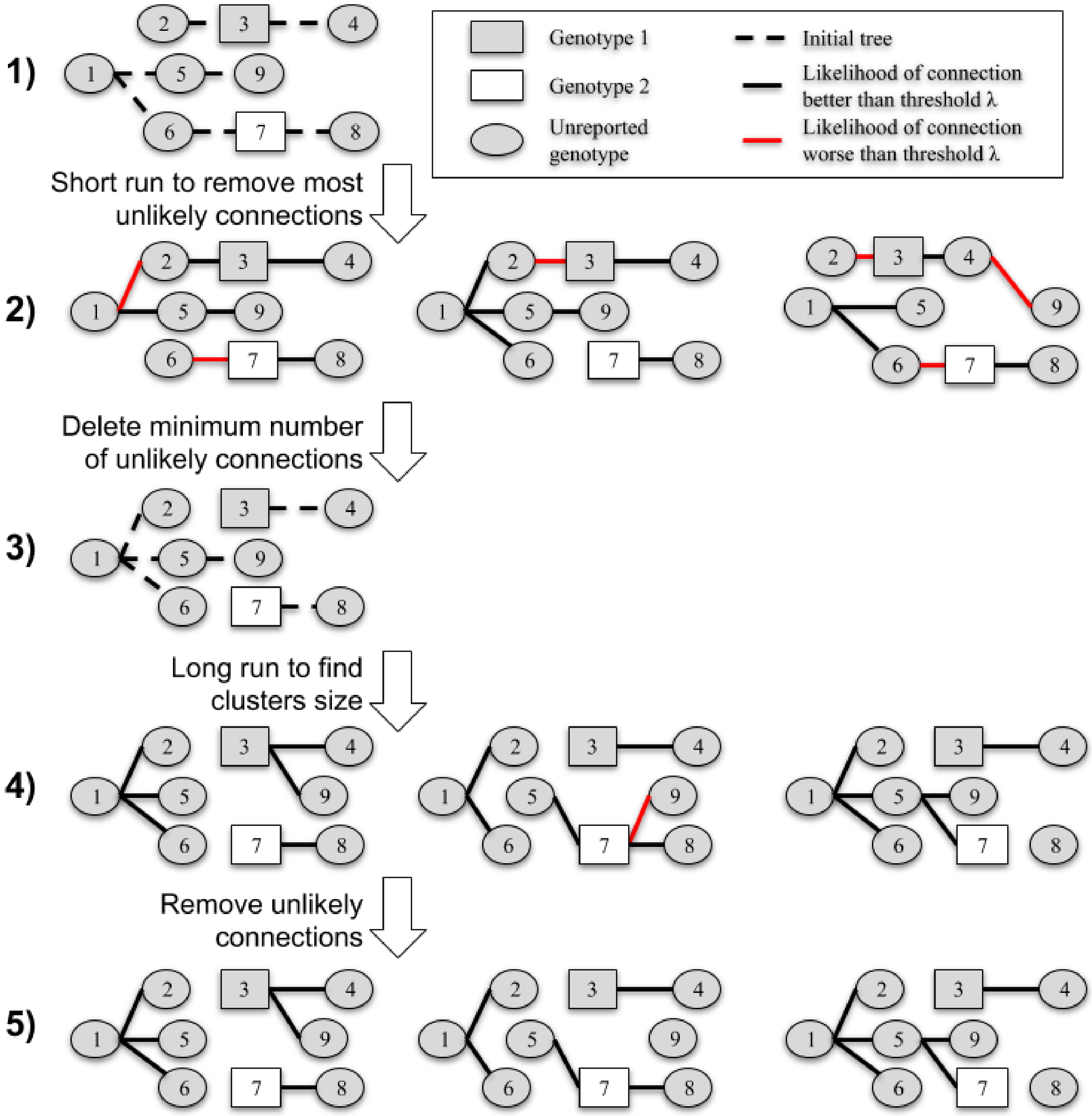

1)
Genotype 1
Genotype 2
Unreported genotype
Initial tree
Likelihood of connection better than threshold λ
Likelihood of connection worse than threshold λ
Short run to remove most unlikely connections
2)
Delete minimum number of unlikely connections
3)
Long run to find clusters size
4)
Remove unlikely connections
5)
1
2
3
4
5
6
7
8
9

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 line showing connection worse than the arbitrary threshold λ. Step 3: Initial tree for the final run, with 1 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.

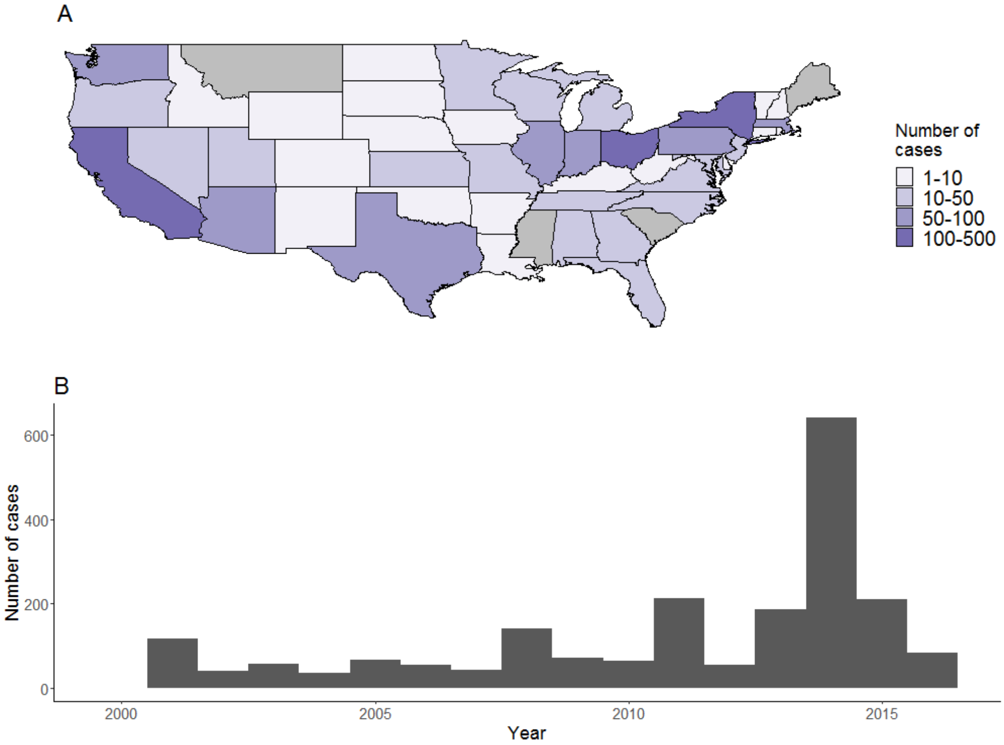


Figure 3: Panel A: number of cases per state and Panel B: Annual number of cases reported in the United States between 2001 and 2016. Alaska and Hawaii are not shown on Panel A.

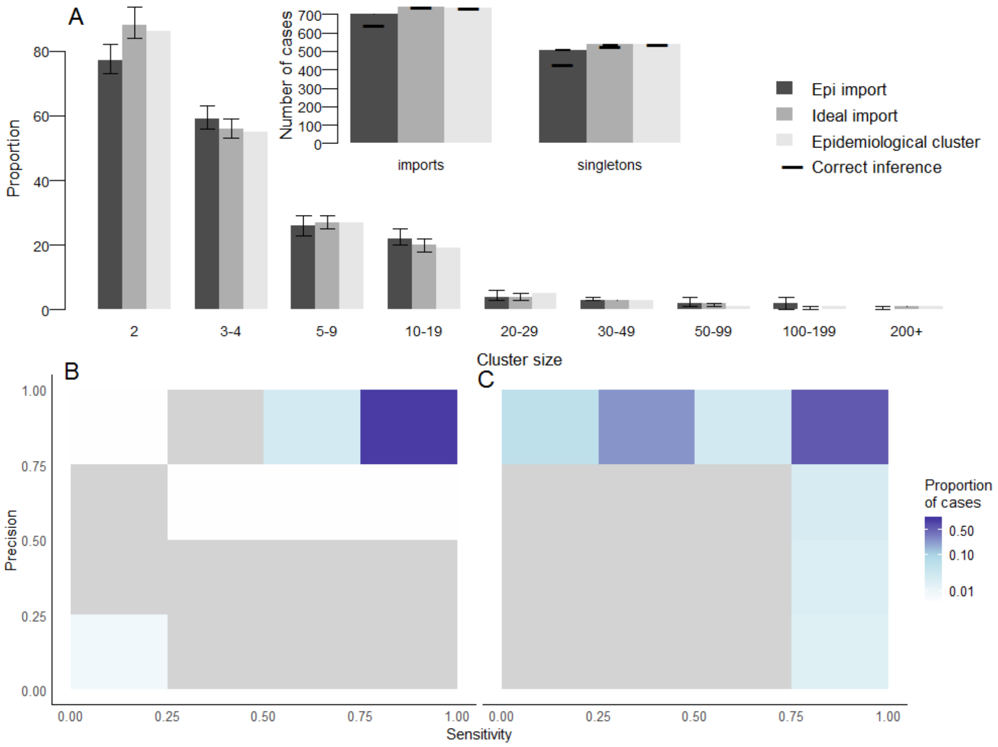


Figure 4: Description of transmission clusters inferred using prior knowledge on importation status of cases. Panel A: Cluster size distribution for the scenario 1 and 2 (grey and dark grey), compared to the reference clusters (lighgrey). Arrows represent the 95% credibility intervals of each estimate. Only clusters containing at least 2 cases are represented. Insert: Number of importations and number of isolated cases (singletons) in scenario 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. Panel 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. Panel C: Same for scenario 2.

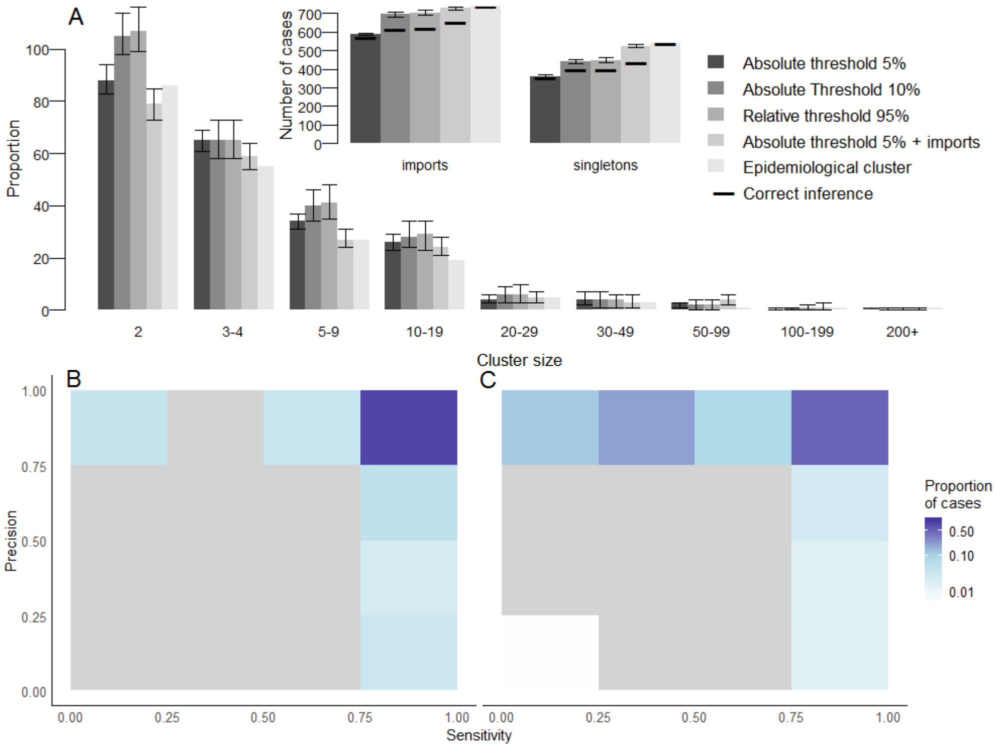
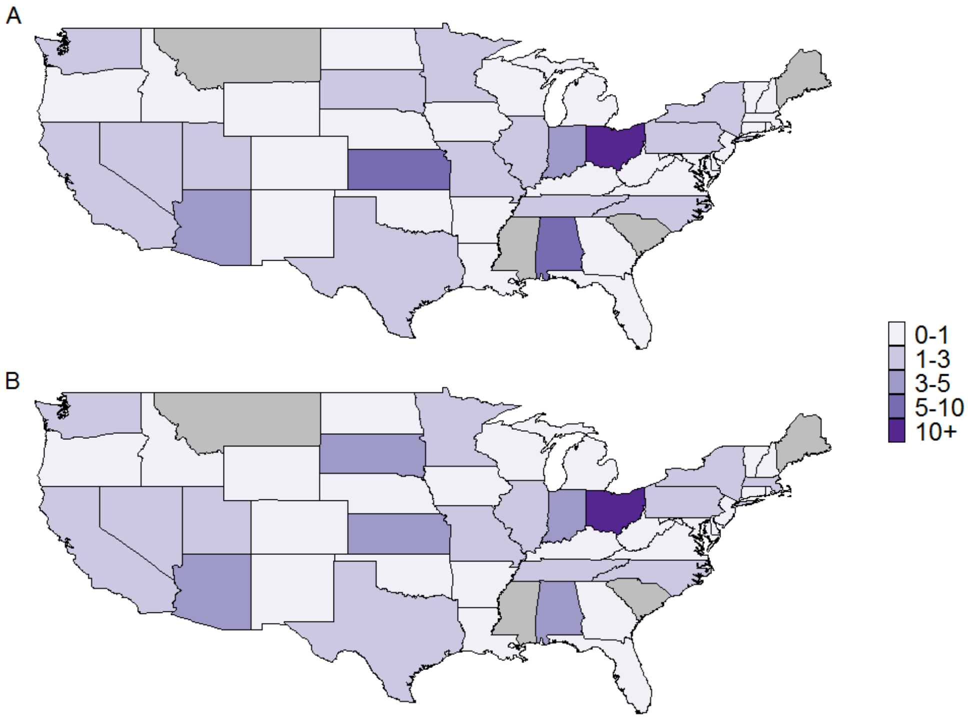


Figure 5: Description of transmission clusters generated with inferred importation status of cases. Panel A: Cluster size distribution for different value of threshold in the scenario 3 (sorted by shades of grey), compared to the reference clusters (lighgrey). Arrows represent the 95% credibility intervals of each estimate. Only clusters containing at least 2 cases are represented. Insert: Number of importations and number of isolated cases (singletons). For each scenario, the horizontal dark line represents the number of importation that are also importations in the reference clusters, same for singletons. Panel 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. Panel C: Same when importation status is taken from the contact tracing investigations and inferred using a 5% relative threshold.



460

461 Figure 6:Ratio of the number of importations over the number of subsequent cases in each state in A/ Scenario 1
462 (Ideal importations) and B/ Scenario 3 with epidemiological importations and $k = 0.05$. Grey states represent states
463 that did not report any case.

464 Tables

465 Table 1: Values of parameters used to cluster cases declared in the United States

Parameter	**Symbol**	**Distribution**
Incubation period	$f(t)$	Gamma, mean = 11.5, sd = 2.24
Generation time	$w(t)$	Normal, Mean = 11.7, sd = 2.0
Conditional report ratio	ρ	Prior: Beta distribution, Mean = 0.65, sd = 0.15
Spatial parameter 1	a	Prior: Uniform distribution
Spatial parameter 2	b	Prior: Uniform distribution
Spatial pre clustering	γ	Fixed: 100 km
Temporal pre clustering	δ	Fixed: 30 days
Importation threshold	λ	Absolute: • $5 * \log 0.05$ • $5 * \log 0.1$ Relative: • 5%

466

467

Reference

[1] Ferguson NM, Donnelly CA, Anderson RM. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. Nature 2001. https://doi.org/10.1038/35097116.

[2] Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal. Am J Epidemiol 2004;160:509–16.

[3] Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355–9. https://doi.org/10.1038/nature04153.

[4] Faye O, Boëlle P-Y, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect Dis 2015;15:320–6. https://doi.org/10.1016/S1473-3099(14)71075-8.

[5] Ypma RJF, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 2013;195:1055–62. https://doi.org/10.1534/genetics.113.154856.

[6] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B Biol Sci 2007;274:599–604. https://doi.org/10.1098/rspb.2006.3754.

[7] Cauchemez S, Ferguson NM. Methods to infer transmission risk factors in complex outbreak data. J R Soc Interface 2012;9:456–69. https://doi.org/10.1098/rsif.2011.0379.

[8] Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Comput Biol 2014;10. https://doi.org/10.1371/journal.pcbi.1003457.

[9] Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS Comput Biol 2019.

https://doi.org/10.1371/journal.pcbi.1006930.

[10] Haydon DT, Chase-Topping M, Shaw DJ, Matthews L, Friar JK, Wilesmith J, et al. The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. Proc R Soc B Biol Sci 2003. https://doi.org/10.1098/rspb.2002.2191.

[11] Cauchemez S, Boëlle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, et al. Real-time estimates in early detection of SARS. Emerg Infect Dis 2006.

[12] Heijne JCM, Rondy M, Verhoef L, Wallinga J, Kretzschmar M, Low N, et al. Quantifying transmission of norovirus during an outbreak. Epidemiology 2012. https://doi.org/10.1097/EDE.0b013e3182456ee6.

[13] Kendall M, Ayabina D, Colijn C. Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees 2016:1–22. https://doi.org/10.1214/17-STS637.

[14] Worby CJ, O'Neill PD, Kypraios T, Robotham J V., De Angelis D, Cartwright EJP, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. Ann Appl Stat 2016. https://doi.org/10.1214/15-AOAS898.

[15] Lau MSY, Marion G, Streftaris G, Gibson G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. PLoS Comput Biol 2015. https://doi.org/10.1371/journal.pcbi.1004633.

[16] Spada E, Sagliocca L, Sourdis J, Garbuglia AR, Poggi V, De Fusco C, et al. Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. J Clin Microbiol 2004. https://doi.org/10.1128/JCM.42.9.4230-4236.2004.

[17] Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, et al. A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. Proc R Soc B Biol Sci 2014. https://doi.org/10.1098/rspb.2013.3251.

[18] Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance

elucidates Ebola virus origin and transmission during the 2014 outbreak. Science (80-) 2014;345:1369--1372. https://doi.org/10.1126/science.1259657.

[19] Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. Nature 2015;524:97–101. https://doi.org/10.1038/nature14594.

[20] Ruan YJ, Wei CL, Ee LA, Vega VB, Thoreau H, Yun STS, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet 2003;361:1779–85. https://doi.org/10.1016/S0140-6736(03)13414-9.

[21] Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 2009;10:540–50. https://doi.org/10.1038/nrg2583.

[22] Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science (80-) 2004;303.

[23] Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLoS Pathog 2018. https://doi.org/10.1371/journal.ppat.1006885.

[24] Rota PA, Brown K, Mankertz A, Santibanez S, Shulga S, Muller CP, et al. Global distribution of measles genotypes and measles molecular epidemiology. J Infect Dis 2011;204. https://doi.org/10.1093/infdis/jir118.

[25] Hiebert J, Severini A. Measles molecular epidemiology : What does it tell us and why is it important? Canada Commun Dis Rep CCDR 2014;40.

[26] Brown KE, Rota PA, Goodson JL, Williams D, Abernathy E, Takeda M, et al. Genetic characterization of measles and rubella viruses detected through global measles and rubella elimination surveillance, 2016-2018. Morb Mortal Wkly Rep 2019;68:587–91. https://doi.org/10.15585/mmwr.mm6826a3.

[27] Gardy JL, Naus M, Amlani A, Chung W, Kim H, Tan M, et al. 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 2015;212:1574–8. https://doi.org/10.1093/infdis/jiv271.

[28] Penedos AR, Myers R, Hadef B, Aladin F, Brown KE. Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks 2015:1–16. https://doi.org/10.1371/journal.pone.0143081.

[29] World Health Organisation. Measles virus nomenclature Update: 2012. Wkly Epidemiol Rec 2012;87:73–80. https://doi.org/10.1016/j.actatropica.2012.04.013.

[30] Hagemann C, Streng A, Kraemer A, Liese JG. Heterogeneity in coverage for measles and varicella vaccination in toddlers - Analysis of factors influencing parental acceptance. BMC Public Health 2017;17. https://doi.org/10.1186/s12889-017-4725-6.

[31] Glasser JW, Feng Z, Omer SB, Smith PJ, Rodewald LE. 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 2016;16:599–605. https://doi.org/10.1016/S1473-3099(16)00004-9.

[32] Gastañaduy PA, Budd J, Fisher N, Redd SB, Fletcher J, Miller J, et al. A Measles Outbreak in an Underimmunized Amish Community in Ohio. N Engl J Med 2016;375:1343–54. https://doi.org/10.1056/NEJMoa1602295.

[33] Woudenberg T, Van Binnendijk RS, Sanders EAM, Wallinga J, De Melker HE, Ruijs WLM, et al. Large measles epidemic in the Netherlands, May 2013 to March 2014: Changing epidemiology. Eurosurveillance 2017;22:1–9. https://doi.org/10.2807/1560-7917.ES.2017.22.3.30443.

[34] Keenan A, Ghebrehewet S, Vivancos R, Seddon D, MacPherson P, Hungerford D. 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 2017;7. https://doi.org/10.1136/bmjopen-2016-014106.

[35] Kucharski AJ, Edmunds WJ. Characterizing the Transmission Potential of Zoonotic Infections from Minor Outbreaks. PLoS Comput Biol 2015;11:1–17. https://doi.org/10.1371/journal.pcbi.1004154.

[36] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 2008;5:0381–91. https://doi.org/10.1371/journal.pmed.0050074.

[37] Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLoS Comput Biol 2013;9:1–17. https://doi.org/10.1371/journal.pcbi.1002993.

[38] Blumberg S, Enanoria WTA, Lloyd-Smith JO, Lietman TM, Porco TC. Identifying postelimination trends for the introduction and transmissibility of measles in the United States. Am J Epidemiol 2014;179:1375–82. https://doi.org/10.1093/aje/kwu068.

[39] Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: A modular platform for outbreak reconstruction. BMC Bioinformatics 2018;19. https://doi.org/10.1186/s12859-018-2330-z.

[40] Lenormand M, Bassolas A, Ramasco JJ. Systematic comparison of trip distribution laws and models. J Transp Geogr 2016;51:158–69. https://doi.org/10.1016/j.jtrangeo.2015.12.008.

[41] Zipf GK. The P 1 P 2/D hypothesis: On the intercity movement of persons. Am Sociol Rev 1946;11:677–86. https://doi.org/10.2307/2657358.

[42] Barthélemy M. Spatial networks. Phys Rep 2011;499:1–79. https://doi.org/10.1016/j.physrep.2010.11.002.

[43] Xia Y, Bjørnstad ON, Grenfell BT. Measles Metapopulation Dynamics: A Gravity Model for Epidemiological Coupling and Dynamics. Am Nat 2004;164:267–81.

https://doi.org/10.1086/422341.

[44] Lenormand M, Huet S, Gargiulo F, Deffuant G. A Universal Model of Commuting Networks. PLoS One 2012;7. https://doi.org/10.1371/journal.pone.0045985.

[45] Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. Mach Learn 2003;50:5–43. https://doi.org/10.1023/A:1020281327116.

[46] Centers for Disease Control and Prevention (CDC). National Notifiable Disease Surveillance System: measles/rubeola 2013. https://wwwn.cdc.gov/nndss/conditions/measles/case-definition/2013/ (accessed October 23, 2019).

[47] Lessler J, Reich NG, Brookmeyer R, Perl TM, Nelson KE. Incubation periods of acute respiratory viral infections: a systematic review 2015;9:291–300. https://doi.org/10.1016/S1473-3099(09)70069-6.Incubation.

[48] Klinkenberg D, Nishiura H. The correlation between infectivity and incubation period of measles, estimated from households with two cases. J Theor Biol 2011;284:52–60. https://doi.org/10.1016/j.jtbi.2011.06.015.

[49] Fine PEM. The Interval between Successive Cases of an Infectious Disease. Am J Epidemiol 2003;158:1039–47. https://doi.org/10.1093/aje/kwg251.

[50] US Census Bureau. Centers of Population for the 2010 Census 2010. https://www.census.gov/geographies/reference-files/2010/geo/2010-centers-population.html (accessed August 22, 2019).

[51] Woudenberg T, Woonink F, Kerkhof J, Cox K, Ruijs WLM. The tip of the iceberg : incompleteness of measles reporting during a large outbreak in The Netherlands in 2013 – 2014. Epidemiol Infect 2018;146:716–22. https://doi.org/https://doi.org/10.1017/ S0950268818002698.

[52] Gastañaduy PA, Funk S, Paul P, Tatham L, Fisher N, Budd J, et al. Impact of public health responses during ameasles outbreak in an amish community in Ohio: Modeling the dynamics of

transmission. Am J Epidemiol 2018. https://doi.org/10.1093/aje/kwy082.

[53] Patel M, Lee AD, Clemmons NS, Redd SB, Poser S, Blog D, et al. National Update on Measles Cases and Outbreaks - United States, January 1-October 1, 2019. MMWR Morb Mortal Wkly Rep 2019;68:893–6. https://doi.org/10.15585/mmwr.mm6840e2.

[54] Zipprich J, Winter K, Hacker J, Xia D, Watt J, Harriman K. Measles outbreak--California, December 2014-February 2015. vol. 64. 2015. https://doi.org/10.1016/j.annemergmed.2015.04.002.

[55] Durrheim D. Measles elimination, immunity, serosurveys, and other immunity gap diagnostic tools. J Infect Dis 2018;218:341–3. https://doi.org/10.1093/infdis/jiy138.

[56] Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comput Biol 2017. https://doi.org/10.1371/journal.pcbi.1005697.