Inferring person-to-person networks of *Plasmodium falciparum* transmission: is routine surveillance data up to the task?

John H. Huber[1*], Michelle S. Hsiang[2,3,4], Nomcebo Dlamini[5], Maxwell Murphy[6], Sibonakaliso Vilakati[5], Nomcebo Nhlabathi[5], Anita Lerch[1], Rasmus Nielsen[7], Nyasatu Ntshalintshali[8], Bryan Greenhouse[6,9], T. Alex Perkins[1*]

[1]Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana, United States of America

[2]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America

[3]Malaria Elimination Initiative, Global Health Group, University of California, San Francisco, California, United States of America

[4]Department of Pediatrics, University of California, San Francisco, California, United States of America

[5]National Malaria Elimination Programme, Ministry of Health, Manzini, Eswatini

[6]Department of Medicine, University of California, San Francisco, California, United States of America

[7]Department of Integrative Biology and Statistics, University of California, Berkeley, California, United States of America

[8]Clinton Health Access Initiative, Eswatini Country Office, Mbabane, Eswatini

[9]Chan Zuckerberg Biohub, San Francisco, California, United States of America

*Corresponding Authors

Email: jhuber3@nd.edu (JHH); taperkins@nd.edu (TAP)

Short Title: Inferring person-to-person transmission networks

Abstract

Inference of person-to-person transmission networks using surveillance data is increasingly used to estimate spatiotemporal patterns of pathogen transmission. Several data types can be used to inform transmission network inferences, yet the sensitivity of those inferences to different data types is not routinely evaluated. We evaluated the influence of different combinations of spatial, temporal, and travel-history data on transmission network inferences for *Plasmodium falciparum* malaria. We found that these data types have limited utility for inferring transmission networks and may overestimate transmission. Only when outbreaks were temporally focal or travel histories were accurate was the algorithm able to accurately estimate the reproduction number under control, R_c. Applying this approach to data from Eswatini indicated that inferences of R_c and spatiotemporal patterns therein depend upon the choice of data types and assumptions about travel-history data. These results suggest that transmission network inferences made with routine malaria surveillance data should be interpreted with caution.

Introduction

Concomitant with improved epidemiological surveillance, there is growing interest to leverage the collected data to infer transmission networks for a wide range of pathogens and to use those inferences to inform public health efforts. Past studies have incorporated temporal data[1] and spatial data[2–5] to estimate pairwise probabilities of transmission between individual cases and to use those estimates to infer time-varying and spatially varying reproduction numbers, respectively. More recently, methods have been developed to incorporate this type of detailed, individual-level epidemiological data[6–8] to infer transmission networks for infectious diseases of humans, including severe acute respiratory syndrome[9] and tuberculosis[10], and of animals, such as rabies[11] and foot-and-mouth disease[12].

In addition to the diseases for which these methods have been applied to date, there is a growing need to apply similar methods to malaria in near-elimination settings. As incidence of malaria declines within a country, transmission becomes more heterogeneous in space and time[13]. Focal areas of high transmission, known as "hotspots," pose a serious risk of fueling resurgence if left untargeted, potentially reversing decades of progress towards elimination[14]. To this end, granular estimates of when and where transmission occurs are needed, as spatially aggregated estimates may obscure important heterogeneities of practical relevance to control efforts[15]. In addition to characterizing details of local transmission, measurement of progress towards malaria elimination hinges on correct classification of cases as imported or locally acquired[16,17], which is a byproduct of estimating transmission networks.

Previous work on malaria has made progress on the use of individual-level epidemiological data to infer transmission networks of *Plasmodium falciparum*, the parasite primarily responsible for human malaria in many regions of the world. Churcher *et al.*[18] used

temporal data to estimate the proportion of imported cases needed to confidently estimate the reproduction number under control, R_c, below one and thereby provide evidence of controlled, non-endemic malaria transmission. Reiner *et al.*[6] then built upon this work by incorporating spatial data and inferring an individual-level transmission network of *P. falciparum* in Eswatini. More recently, Routledge *et al.*[19,20] used related approaches to infer transmission networks and R_c of *P. vivax* in El Salvador and China.

As the adoption of these methods increases, in particular for malaria, care should be taken to assess how the epidemiological setting and the inclusion or exclusion of certain data types might affect the accuracy of transmission network inferences, as well as resultant inferences about epidemiological quantities including R_c and spatiotemporal variation therein. A recent study by Campbell *et al*.[21] noted that epidemiological data alone was generally insufficient to reconstruct transmission networks of other pathogens, ranging from *Mycobacterium tuberculosis* to SARS-CoV. Although *P. falciparum* malaria was not considered in that analysis, its long serial interval[22] calls into the question the utility of epidemiological data for this purpose, though this has been largely unaddressed in past studies. Furthermore, past transmission network inferences for malaria have relied on various types of epidemiological data, ranging from the timing of symptom onset[18–20] to more detailed spatiotemporal data[6]. Each study incorporated travel-history information into transmission network inferences and considered these data to be perfectly accurate, assuming that all cases that reported travel were imported. However, travel history may be an imperfect indicator of importation owing to errors in recall[17] and the fact that travel to an area of ongoing transmission alone does not guarantee that an individual was infected there[17,23]. *P. falciparum* transmission network inferences are likely to be sensitive to the choice of data types[24], and failure to evaluate the sensitivity of transmission network inferences

to choices about data types and different assumptions about possible errors in travel-history data could lead to apparently confident, though ultimately incorrect, assessments of *P. falciparum* transmission risk in near-elimination settings.

Here, we present a Bayesian method for inferring transmission networks based on temporal, spatial, and travel-history data for individual malaria cases. We use it to characterize the sensitivity of transmission network inferences to the inclusion of different data types and to different assumptions about the accuracy of travel histories. Our method builds upon previous work by leveraging individual-level epidemiological data to obtain posterior estimates of transmission networks and model parameters in a way that can accommodate different assumptions about errors in travel histories. After establishing a proof-of-concept of our inference method on simple test cases, we applied our method to real-world surveillance data from Eswatini and additional simulated data sets to understand how the inclusion or exclusion of different data types and different assumptions about travel-history error affect our ability to infer transmission networks and estimate transmission metrics, namely R_c.

Results

To establish proof-of-concept, we first applied our inference method on three simple test cases and evaluated how well our inferences recovered the true transmission networks. We then applied our method to surveillance data collected in Eswatini during 2013-2017. Our focus was less on understanding malaria epidemiology in Eswatini and more on understanding how epidemiological conclusions change with the inclusion or exclusion of different data types and different assumptions about travel histories. These inference settings used: (1) spatial and temporal data while estimating the accuracy of the travel history (default setting); (2) spatial and

temporal data while believing the travel history; (3) spatial and temporal data alone; (4) temporal data while estimating the accuracy of the travel history; and (5) temporal data while believing the travel history. To validate the inferences based on data from Eswatini, we simulated data generated using posterior parameter estimates obtained from the data from Eswatini and evaluated the ability of our inference method to recover the true transmission networks along with the underlying parameters on those simulated data. Finally, we performed a simulation sweep across different epidemiological settings to determine the range of conditions under which our inference method yielded reliable estimates of transmission. A full description of the analyses and additional results can be found in the Supplement.

Application to Eswatini surveillance data

We applied our method to surveillance data collected in Eswatini during 2013-2017. Under the default inference setting, we estimated the diffusion coefficient D, which quantifies the spatial spread of transmission, to be 4.42 km^2 day^{-1} (2.92 – 6.18 km^2 day^{-1}) (Fig 1A). This corresponded to a median inferred transmission distance of 13.0 km (0.0160 – 65.9 km), a median inferred serial interval of 47 days (-33 – 150 days) (Fig 2A & 2B), and median estimates of τ_s, the probability that an imported case reported travel, of 0.61 (0.44 – 0.78) compared to the prior distribution mean of 0.80 and τ_l, the probability that a locally acquired case reported travel, of 0.57 (0.53 – 0.61) compared to the prior distribution mean of 0.20 (Fig 1B & 1C). That the 95% credible interval for τ_s contained 0.50 indicated that our inference algorithm found limited use of travel-history data in discriminating between imported and locally acquired cases, because that implies that imported cases have equal probabilities of reporting or not reporting travel. The algorithm estimated the proportion of imported cases to be 0.052, corresponding to $R_c = 0.95$.

Mapping risk of importation and local transmission across Eswatini under the default inference setting, we estimated consistently low risk of importation throughout the country and transmission hotspots in the northeastern part of Eswatini, close to the border with Mozambique (Fig 3A & 3B).

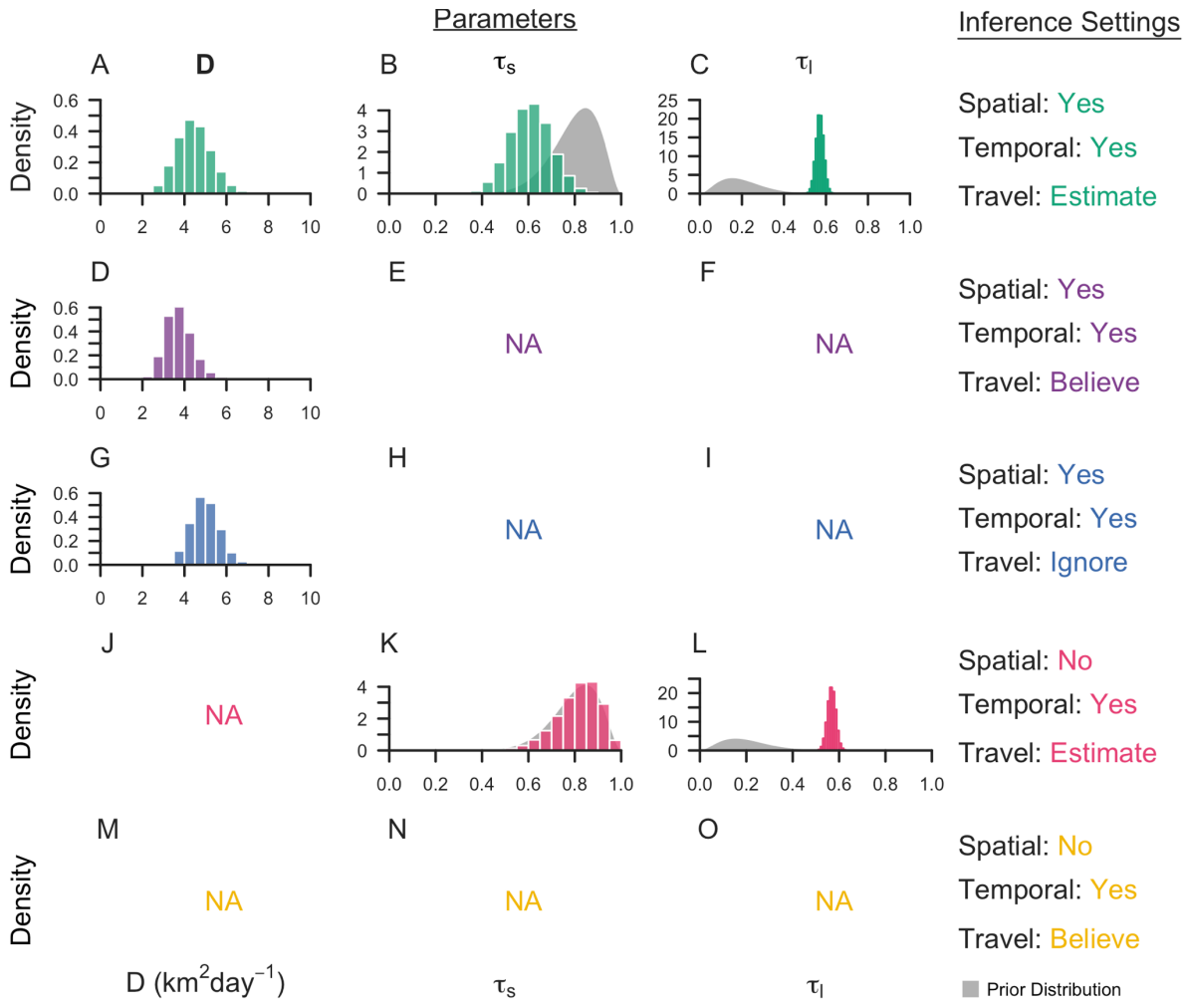


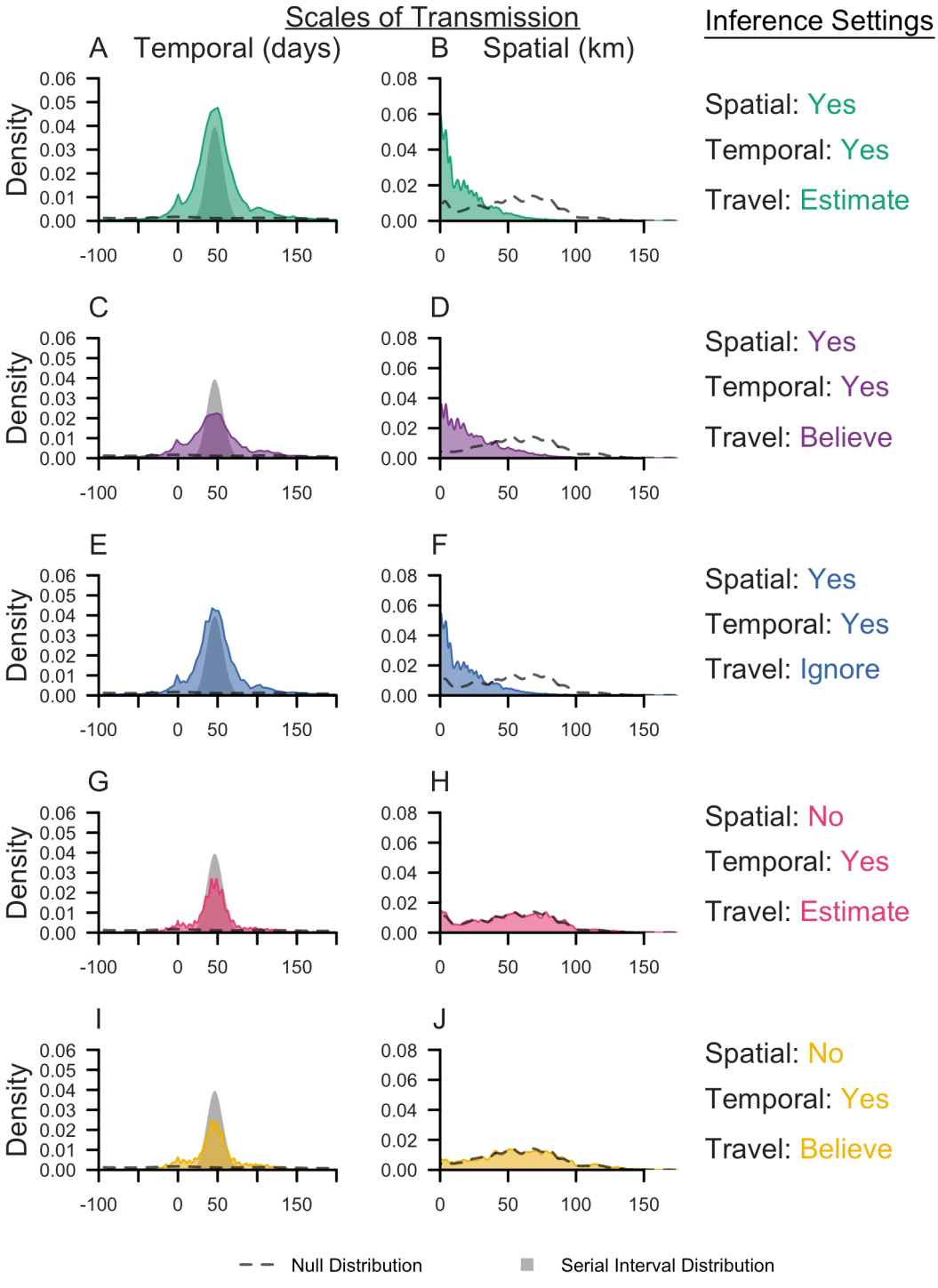
Fig 1. Marginal posterior distributions of parameters from Eswatini surveillance data.

Histograms represent the marginal posterior distribution of each parameter, color-coded by the inference settings used. D is the diffusion coefficient with units km^2day^{-1}, τ_s is the probability that an imported case reports travel, and τ_l is the probability that a locally acquired case reports

143 *travel. Gray shapes represent the prior distributions placed on each parameter. Inference*

144 *settings in which a given parameter was not estimated are indicated by NA.*

145



146

147 ***Fig 2. Spatial and temporal scales of transmission in Eswatini.*** *Kernel density plots of the*
148 *spatial (km) and temporal (days) scales of transmission are reported and color-coded for each*
149 *inference setting. Dashed lines indicate the corresponding null distribution, generated from all*
150 *random pairs of cases in the Eswatini surveillance data set. The null distribution was different if*
151 *we believed the travel history, because classification of cases on the basis on travel history*
152 *reduced the pairs of cases that could be randomly sampled. The grey shape is the serial interval*
153 *distribution used in the likelihood.*
154

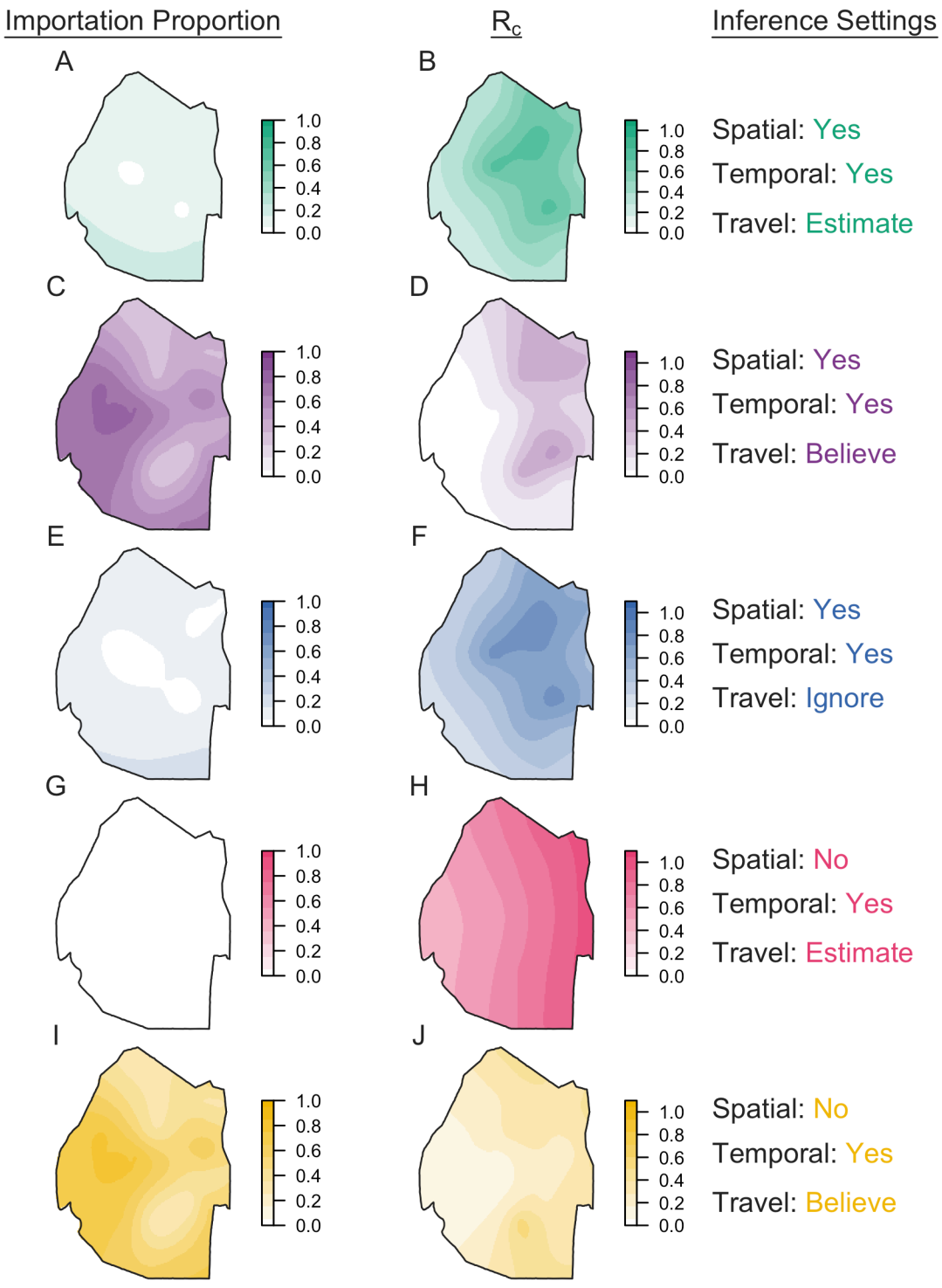


Fig 3. Spatial distribution of importation and transmission risk in Eswatini*. Maps of the proportion of cases that are imported and the reproduction number under control (R_c) were*

generated for each inference setting using a generalized additive model with a Gaussian process basis function setting using the mgcv package in R[25,26]*. In each plot, darker colors indicate greater importation or transmission risk.*

Parameter estimates and transmission network inferences differed under other inference settings. When we believed the travel history, we estimated a larger median transmission distance (Fig 2D). We attribute this increase in the spatial scale of transmission to clusters of cases with positive travel histories located near metropolitan areas. By forcing those cases to be imported, the algorithm tended to infer transmission across longer distances to explain the origins of the remainder of cases that did not report travel and were thereby inferred to be locally acquired. With respect to time, all five inference settings produced consistent serial interval estimates, though the inclusion of spatial data allowed for a wider range of transmission linkages in time (Fig 2A, 2C, & 2E). Finally, in the absence of spatial data, the model estimated higher predictive power of travel histories in identifying imported cases (τ_s: 0.83, [0.60, 0.95]), though the travel history was consistently found to be uninformative for identifying locally acquired cases (τ_l: 0.57, [0.53, 0.60]) (Fig 1K & 1L).

Classification of cases as imported or locally acquired, key information for control programs, was sensitive to the choice of inference setting. The proportion of cases classified as imported was most sensitive to different assumptions about the accuracy of the travel histories (Fig 3, left column; Fig 4). Believing the travel history yielded high estimates of importation in western Eswatini (Fig 3C & 3I), whereas estimating or ignoring the travel history yielded low, relatively homogeneous estimates of importation risk (Fig 3A, 3E, & 3G). For instance, using temporal data and estimating the accuracy of the travel history produced probabilities of

181 importation that ranged 0.0043 – 0.0050, suggesting that nearly all cases resulted from local
182 transmission (Figs 3G, 4D). Estimates of the spatial distribution of R_c depended most on the
183 choice of which data types we included (Fig 3, right column). Notably, inclusion of spatial and
184 temporal data produced a consistent spatial distribution of relative transmission risk, with
185 transmission hotspots in northeastern Eswatini (Fig 3B, 3D, & 3F). However, believing the travel
186 history reduced the magnitude of transmission that we inferred from a median R_c of 0.95 (Figs
187 3B, 4A) under default settings to 0.41 (Figs 3D, 4B). Omitting spatial data changed the spatial
188 distribution of transmission. Estimating the accuracy of the travel history yielded high
189 transmission estimates (median R_c: 1.00) in eastern Eswatini (Fig 3H), whereas believing the
190 travel history inferred hotspots of transmission (median R_c: 0.42) in southern Eswatini (Fig 3J).
191 We note that believing the travel history led to slightly different median estimates of R_c (0.41 vs.
192 0.42) depending upon whether spatial data was included, because the travel histories were
193 unknown for 36 cases included in the analysis. As part of the inference procedure, the algorithm
194 classified these cases as imported or locally acquired, and including spatial data caused a greater
195 number of cases to be inferred to be imported.
196

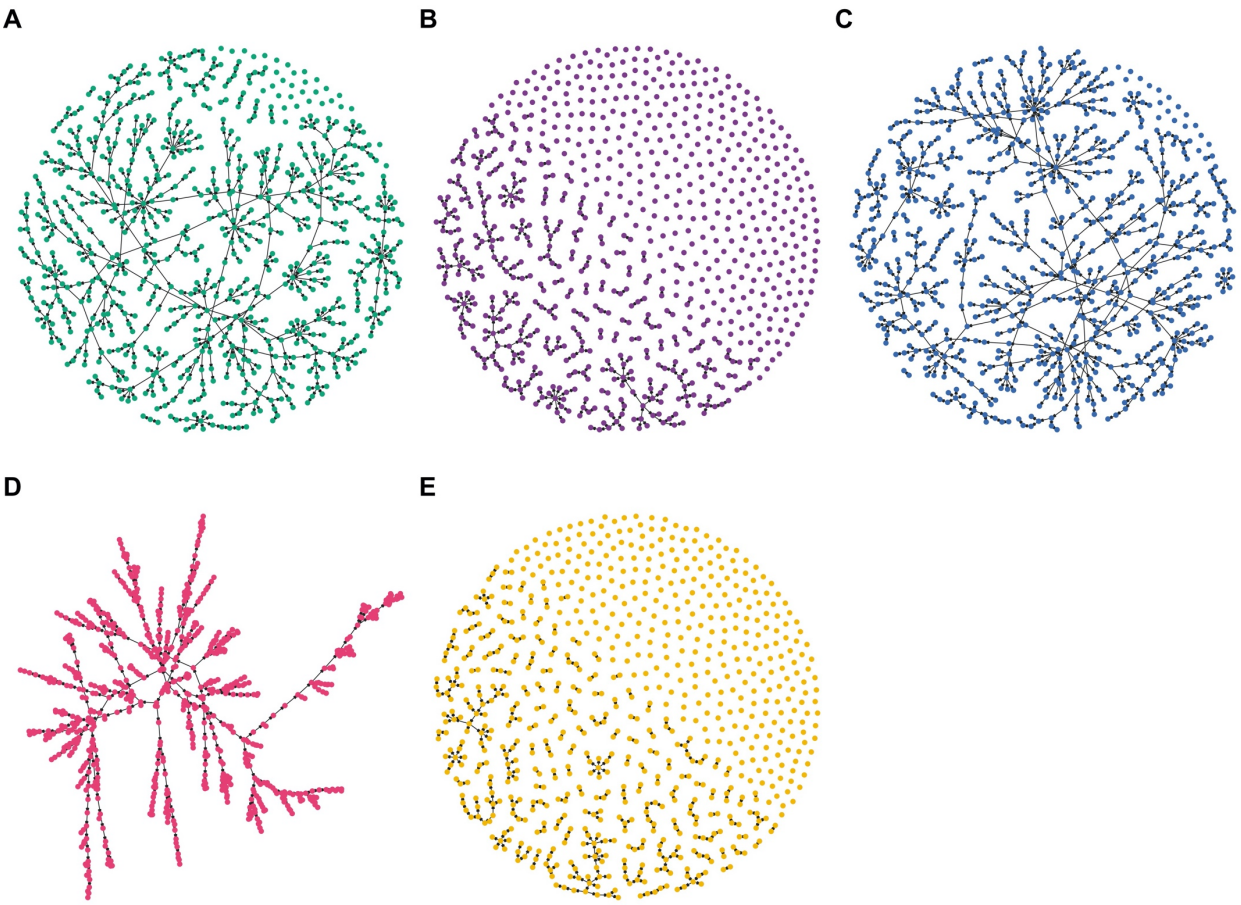


Fig 4. Maximum a posteriori transmission networks in Eswatini. *The maximum a posteriori transmission networks (i.e., the transmission network in the posterior distribution with the highest likelihood) is shown for each inference setting: (A) spatial and temporal data while estimating the accuracy of the travel history; (B) spatial and temporal data while believing the travel history; (C) spatial and temporal data alone; (D) temporal data while estimating the accuracy of the travel history; and (E) temporal data while believing the travel history. In each transmission network, circles represent nodes, and arrows represent directed edges.*

Validation of inferences from Eswatini

Reconciling the different inferences under different inference settings in Figs. 1-4 was challenging because the true, underlying network and parameters were unknown. Using median posterior estimates from the Eswatini data under each inference setting, we simulated data and assessed the ability of the inference method to recover the underlying parameters and transmission networks (Table 1). We found that the model was able to estimate τ_s and τ_l reasonably well, depending on the inference setting (Fig 5). One exception was that the algorithm slightly overestimated τ_s under the default inference setting. We attribute this to the low proportion (0.052) of imported infections in the simulated data set and the strong prior placed on this parameter. This tendency was not observed under the inference setting where spatial data was excluded, because the true value of τ_s closely matched the mean of the prior distribution in that case (see the Supplement for further discussion).

Table 1. Characteristics of simulated data generated using the branching process model.

Inference Setting			Network Size	# of Outbreaks	Prop. Imported	D	τ_s	τ_l
Space	**Time**	**Travel**						
Yes	Yes	Estimate	775	43	0.052	4.42	0.61	0.57
Yes	Yes	Believe	775	492	0.59	3.70	1	0
Yes	Yes	No	775	36	0.043	4.96	NA	NA
No	Yes	Estimate	775	1	0.0013	NA	0.83	0.57
No	Yes	Believe	775	489	0.58	NA	1	0

A description of the simulated data used in the inference exercises are reported for each of the five inference settings. The total number of nodes in the network, the number of distinct outbreaks, the proportion of cases that are imported, and the underlying parameters are provided.

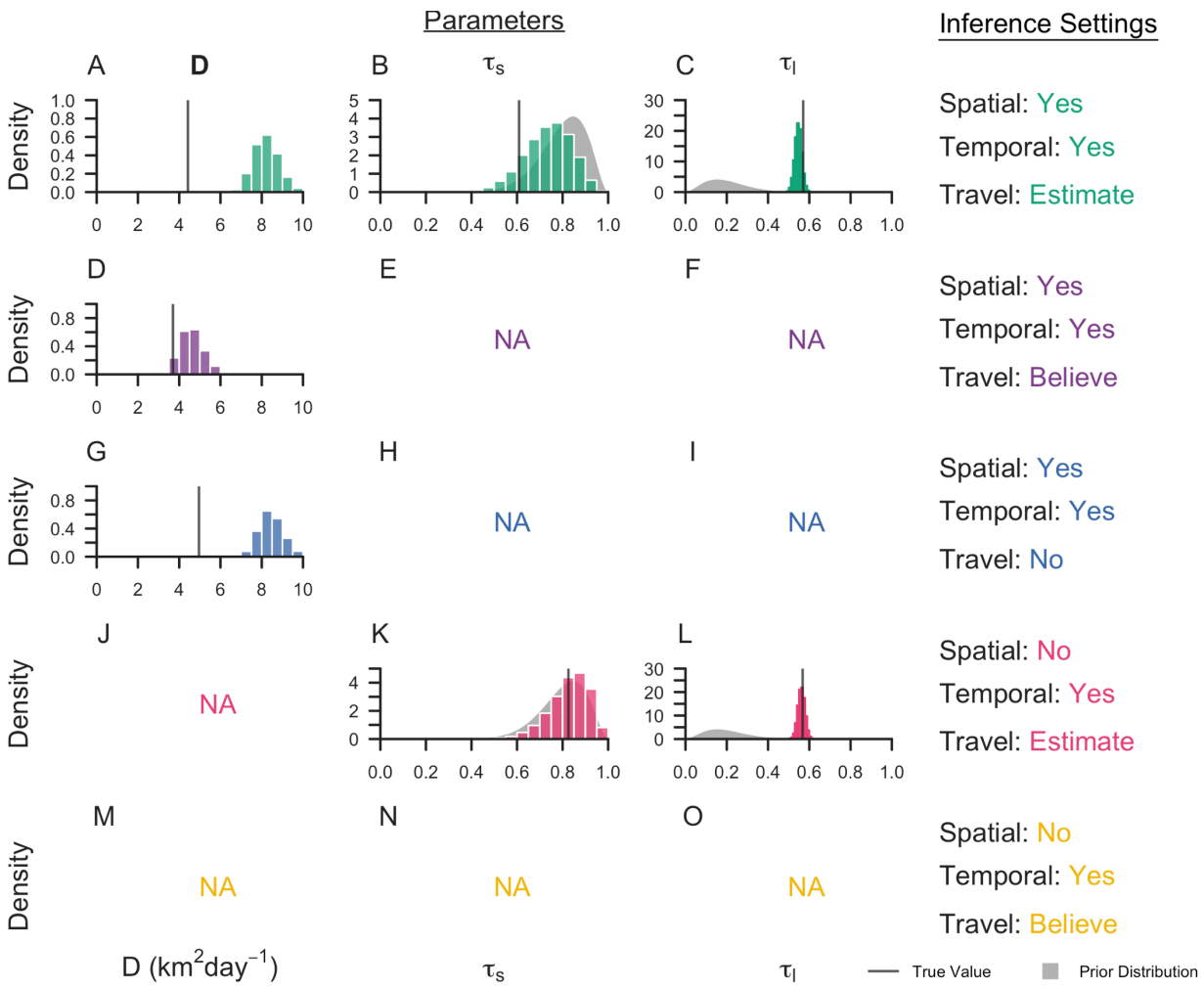


Fig 5. Marginal posterior distributions for parameters inferred from simulated data*. The marginal posterior distributions are reported for each inference setting from its respective simulated data set. Each line denotes the true value of the parameter, and the grey shapes represent the prior distributions of the parameters. Inference settings in which a given parameter was not estimated are indicated by NA.*

With the exception of believing the travel history, the model consistently overestimated the diffusion coefficient (Fig 5A, 5D, & 5G). We attribute the challenge of correctly estimating the diffusion coefficient to an inability to correctly estimate the underlying transmission network,

the extent of local transmission in the network, and a numerical insensitivity in the overall likelihood to changes in D. When we conditioned the likelihood of D on the true transmission network when R_c was high, the true values of D fell close to the range of maximum-likelihood estimates, suggesting that this parameter could be estimated correctly if the true network was identified (S3 Fig). The likelihood around the true value was very flat, however, making it easy for D to be estimated incorrectly. When R_c was low, we underestimated the diffusion coefficient, because the likelihood of imported cases increases as D decreases.

The overall accuracy of classifying cases as imported or locally acquired was close to one (Fig 6). Though seemingly promising, these high accuracies masked a tendency to overclassify cases as locally acquired, because many more cases were simulated to be locally acquired than imported. For example, under the default inference setting, the accuracy of correctly classifying imported cases was 0.023 (0.023 – 0.067). Similarly, the accuracies of identifying the parent of each transmission linkage were poor, despite simulating under the assumptions of the model, with accuracies ranging from 0.042 (0.022 – 0.065) when using temporal data and believing the travel history to 0.31 (0.28 – 0.34) when incorporating spatial and temporal data and estimating the accuracy of the travel history (Fig 6, circle points). This suggests that, as the number of cases increases within a fixed space-time window, the information content of routinely collected epidemiological data decreases and the method becomes incapable of correctly estimating the transmission network. Nevertheless, under some settings, the method was able to capture higher-order summaries of the network, such as case classification and R_c (Fig 6, square and diamond points).

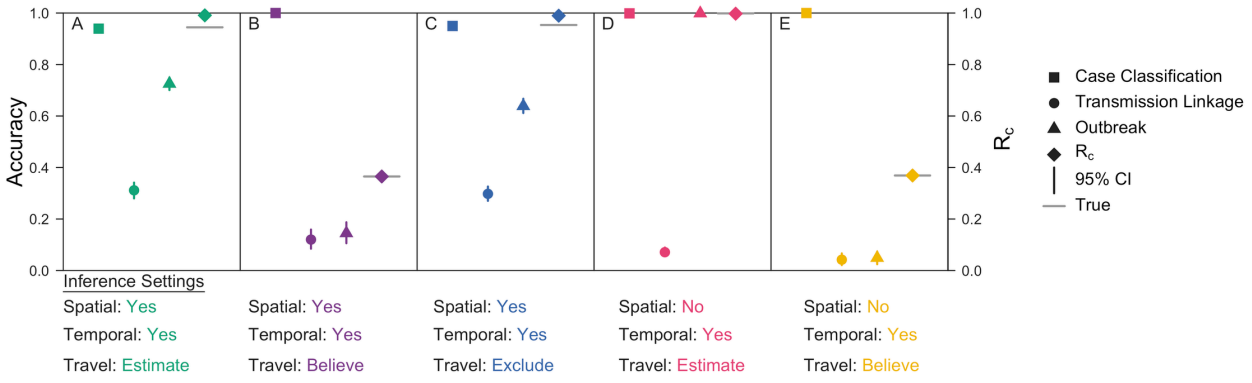


Fig 6. Inference accuracies for validation exercises. *Accuracy metrics are reported for each inference setting applied to its respective simulated data set. Case Classification, represented by squares, refers to the proportion of cases that are correctly classified as imported or locally acquired. Transmission Linkage, denoted by circles, is the proportion of locally acquired cases for which the true parent is correctly identified. Outbreak, represented by triangles, is the proportion of locally acquired cases for which the inferred parent belongs to the correct outbreak. Bars denote the 95% credible intervals, and the grey line is the true R_c value of the network.*

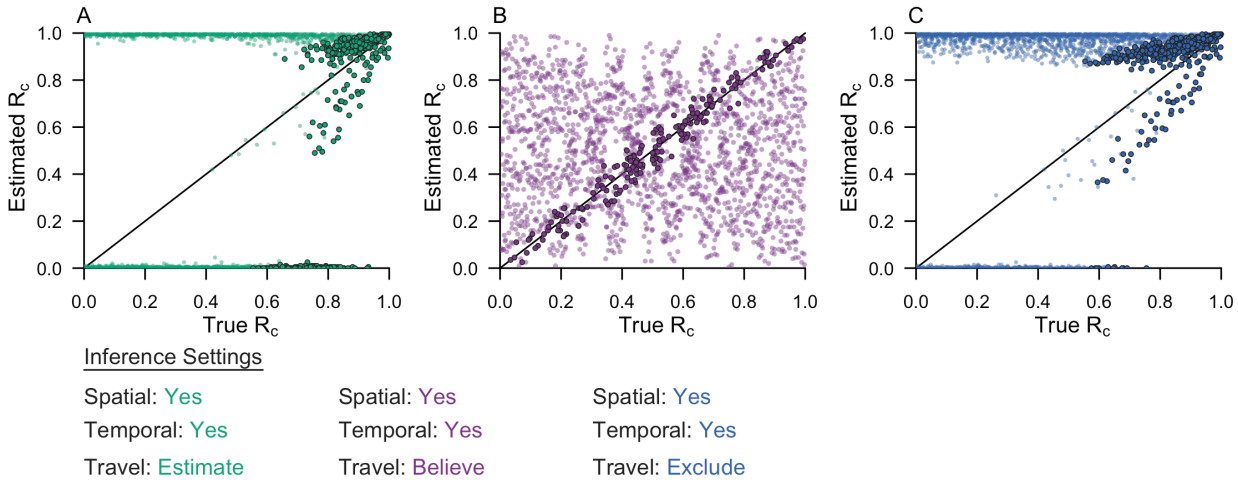
Simulation Sweep

Validation of our inference algorithm revealed that its performance varied across simulated data sets. When applied to a series of simple test cases in which the transmission networks were small and in an optimal spatiotemporal arrangement, the inference method was able to reconstruct the transmission network and correctly estimate R_c (S2 Fig). When applied to larger transmission networks in which outbreaks overlapped in space and time, performance of the inference method was poor (Fig 6). This indicated that the performance of our inference algorithm depends on the epidemiological setting to which it is applied. To address this observation, we generated 2,000 simulated data sets in which we varied the proportion of imported cases, the spatiotemporal

window over which imported cases were distributed, the diffusion coefficient, and the accuracies of the travel history (i.e., τ_s and τ_l) (S2 Table). We then applied our inference algorithm under three different inference settings and quantified the accuracy of reconstructing each transmission network. The three inference settings used: (1) spatial and temporal data while estimating the accuracy of the travel history (default setting); (2) spatial and temporal data while believing the travel history; and (3) spatial and temporal data alone (S1 Table).

We observed that the accuracy of reconstructing transmission networks depended upon both the inference setting used and the epidemiological features of the simulated data. When we used spatial and temporal data and estimated the accuracy of the travel history or excluded it, the accuracy of reconstructing transmission networks depended on the relative proportion and temporal distribution of imported cases (S10 and S12 Figs). As the temporal window over which imported cases are distributed increased, the accuracy of identifying the true parent and the true outbreak of each locally acquired case increased. With an increasing temporal window, outbreaks within the transmission network became relatively more focal in time, which made the likelihoods of alternative transmission linkages more readily distinguishable. More accurate estimates of R_c under these inference settings similarly depended on the temporal window over which imported cases were distributed (Fig 7A, 7C). When the mean temporal interval between imported infections was greater than two times the mean length of the serial interval (i.e., approximately 100 days), our estimates of R_c improved, though we generally overestimated it. Furthermore, as the proportion of imported cases increased and R_c decreased, the accuracy of identifying the correct outbreak of each locally acquired case decreased (S10 and S12 Figs). This pattern reflected the relationship between R_c and the size of individual outbreaks. As R_c

297 decreased, the size of individual outbreaks decreased, and, consequently, the probability that the
298 inferred parent of a locally acquired case belonged to the same outbreak decreased.
299



300

301 ***Fig 7. Comparison of R_c estimates across inference settings.*** *The inference algorithm was*
302 *applied to 2,000 simulated data sets. The estimated R_c is compared to the true R_c for each of the*
303 *inference settings: (A) spatial and temporal data while estimating the accuracy of the travel*
304 *history; (B) spatial and temporal data while believing the travel history; and (C) spatial and*
305 *temporal data alone. Each point represents a simulated data set. The darker, accented points are*
306 *simulated data sets with epidemiological features that improved performance. In (A) and (C), the*
307 *darker, accented points were simulated data sets where the mean temporal interval between*
308 *imported infections was greater than two times the mean serial interval. In (B), the darker,*
309 *accented points were simulated data sets where the proportion of cases reporting travel was*
310 *within 0.05 of the proportion of imported cases.*
311

312 By contrast, when we believed the travel history, the accuracy of reconstructing
313 transmission networks depended most strongly on the accuracies of the travel history. As the

probability of reporting travel increased, the accuracy of classifying imported cases increased, and the accuracy of classifying locally acquired cases decreased (S11 Fig). Under this inference setting, our estimate of R_c depended only on the proportion of cases that reported travel. When the proportion of cases that reported travel matched the proportion of cases that were imported, we correctly estimated R_c (Fig 7B).

Discussion

Our results show that, in most settings, routinely collected surveillance data offer limited value for reconstructing individual-level transmission networks of *P. falciparum* malaria and informing estimates of the reproduction number under control, R_c. Using simulated data similar to the Eswatini surveillance data that we analyzed, we observed that our inference algorithm correctly identified transmission linkages less than 35% of the time. We attribute this inaccuracy primarily to the inherently limited information content of spatiotemporal data on *P. falciparum*. Its characteristically long serial interval[22] means that an appreciable number of cases presenting within a short timeframe are difficult to link to each other based on their timing, even in a relatively facile test case in which the generative process assumed in the likelihood function matched that used to simulate the data. The inability to reconstruct transmission networks using routine surveillance data has been observed for other inference algorithms when applied to pathogens, such as *Mycobacterium tuberculosis* and *Klebsiella pneuomoniae*, with similarly long serial intervals, providing further evidence that the limitations noted in this study are generally inherent to the epidemiological data, rather than our method per se[21].

Under most simulated scenarios and assumptions about the accuracy of travel-history data, we overestimated the number of locally acquired cases, leading to overestimates of R_c.

Crucially, our simulation sweep demonstrated that routinely collected surveillance data was most informative of individual-level transmission networks and R_c when local outbreaks were highly focal in time. Otherwise, while we were able to reconstruct the true transmission network with modest accuracy, we tended to misclassify truly imported cases as locally acquired, thereby overestimating R_c. Taken together, these results suggest limited use of routinely collected surveillance data for informing fine-scale estimates of *P. falciparum* transmission. At broader spatial scales, however, routinely collected surveillance data may still have practical value, because the spatial distribution of cases can reveal epidemiological risk factors relevant for targeted control interventions[27,28].

Although we were able to reach some general conclusions about our inference algorithm, our inferences were highly sensitive to which data types we included and which assumptions we made about the accuracy of travel-history data. Applying our algorithm to surveillance data from Eswatini, we observed that inferred patterns of transmission depended on which data types we included. With the inclusion of spatial data, we captured a spatial pattern of transmission consistent with another analysis from Eswatini[29] with data from a different time period. Assumptions about the travel history appeared to have a strong influence on the overall magnitude of transmission that we inferred, due to the direct relationship between R_c and the proportion of imported cases[16]. As a result, believing the travel history, and thereby treating it as perfectly accurate as in previous approaches[6,18–20], could bias R_c estimates if there are errors in travel-history data. A study comparing community travel surveys to mobile-phone data in Kenya found that travel histories considerably underestimated the volume of travel, suggesting high rates of false negatives in community travel surveys[30]. Believing the travel history may, then, underestimate the number of imported cases and overestimate R_c. Accounting for inaccuracy in

travel-history data is therefore important, and studies pairing community travel surveys with mobile-phone data could be used to inform prior distributions on the likely accuracy of travel-history data[30,31].

The method that we used only considered a single spatial model to infer transmission linkages and assumed complete observation of cases, both of which are factors that could have affected our inferences based on the Eswatini surveillance data. The diffusion model that we used to represent spatial dispersion of parasites assumed that movement is isotropic in space and did not consider landscape features, such as heterogeneity in human population densities and environmental factors that may affect mosquito ecology. A study analyzing self-reported movement patterns in Mali, Burkina Faso, Zambia, and Tanzania found that gravity and radiation models of spatial dispersion fit the data well, though the appropriateness of each model depended on the type of traveler, the travel distance, and the population size of the destination considered[32]. Although a variety of spatial kernels could have been used in our analysis, we expect that the conclusions that we reached are robust to the choice of spatial kernel, because the spatial kernel used in the likelihood matched that used to simulate the data. Regarding the representation of *P. falciparum* infections in our data set from Eswatini, there are asymptomatic and mild infections that are unlikely to have been recorded in the surveillance system yet may comprise a substantial proportion of malaria infections within Eswatini [13]. Accordingly, it is possible that our assumption of complete observation of cases could have biased R_c estimates, likely downward due to the fact that missing cases will tend to make offspring numbers appear smaller than they actually are[33,34]. Even so, we expect that our conclusions about the sensitivity of transmission network inferences to the choice of data types and assumptions about travel-history data are robust to these limitations of our study. This further reinforces our conclusion of

the need for caution in attempting to reconstruct person-to-person transmission networks from routine surveillance data[35], because incomplete observation of cases would lead to greater inaccuracies in our transmission network inferences beyond what we noted in our study.

Given that some of the limitations of our approach may be inherent to the information content of these data types in this system, one potential avenue for improving inferences of fine-scale patterns of *P. falciparum* transmission could involve the integration of additional data streams. For example, mobile-phone data[36] and high-resolution friction surfaces[37] could more realistically characterize mobility patterns, whereas travel-history information that details the dates, duration, and location of each trip that has been used in programmatic contexts[27] could more accurately identify importation events. Additionally, the inclusion of pathogen genetic data, which has the potential to provide a more direct signal of parasite movement, could complement traditional epidemiological data[38]. Diverse genetic markers of *P. falciparum* have been characterized in near-elimination settings, such as Eswatini[39], and have been successfully used to identify imported cases in Bangladesh[40] and Namibia[31]. There is also scope for further methodological development, such as relaxing our assumption of complete observation of infections and incorporating an underlying mechanistic model of transmission (as in Lau *et al.*[8]; Guzzetta *et al.*[41]). Incorporating an underlying mechanistic model would relax our uninformative prior assumption on all possible transmission networks, ruling out transmission networks that are epidemiologically implausible and allowing us to account for spatial differences in transmission potential and the rate of importation due to different epidemiological and demographic factors. This approach would also permit us to estimate the serial interval distribution and seasonal variation therein directly from the data rather than borrow estimates from the literature[22,42,43]. To this end, we envision that leveraging the strengths of our method along with other,

406 complementary methods could strengthen inferences based on routinely collected
407 epidemiological data and open up new possibilities to make use of even more data types, such as
408 serological data, prevalence surveys, and pathogen genetic data[38,44,45].
409

410 ## Conclusions

411 We demonstrated the limitations of routinely collected surveillance data for the inference of
412 individual-level transmission networks of *P. falciparum*. We identified a tendency to
413 overestimate local transmission using routinely collected surveillance data, especially when
414 outbreaks overlapped in space and time. Using both real data from Eswatini and simulated data,
415 we identified strong sensitivities of our inferences to the epidemiological setting, the choice of
416 data types included, and assumptions about the accuracy of travel-history data. Our results
417 indicated that using spatial and temporal data and believing travel histories yielded the most
418 plausible estimates of transmission when applied to the Eswatini surveillance data. However, our
419 simulation sweep demonstrated that the accuracy of our inferences strongly depended on the
420 accuracy of the travel-history data when the travel-history data were assumed to be accurate.
421 These sensitivities to the choice of data types and assumptions about the accuracy of travel-
422 history data could have important programmatic implications if outputs of transmission network
423 inferences are operationalized. Although this study was specific to *P. falciparum*, the results of
424 our analyses indicate that future studies inferring transmission networks of *P. falciparum*, or any
425 pathogen, should carefully consider the epidemiological setting and the choice of data types and
426 assumptions that inform the model and should validate them using simulated data.

Methods

Bayesian framework for estimating transmission linkages

Our goal was to obtain probabilistic estimates of a transmission network N that defines transmission linkages among a set of known cases. The transmission network is defined as a directed, acyclic graph comprised of a set of directed edges represented as $N = \{N_{i,j}\}$ for all i, j. Each $N_{i,j}$ indicates that case i is hypothesized to contain parasites that are the most direct observed ancestors of the parasites contained in case j. In addition, at least one edge denoted $N_{u,j}$ must exist in the network, indicating that the parasites contained in case j have no ancestors among the parasites contained in any known local case and are instead contained in some unknown case u from some source population s, such that it is denoted u_s. To illustrate this terminology, an example transmission network is depicted in Fig 8.

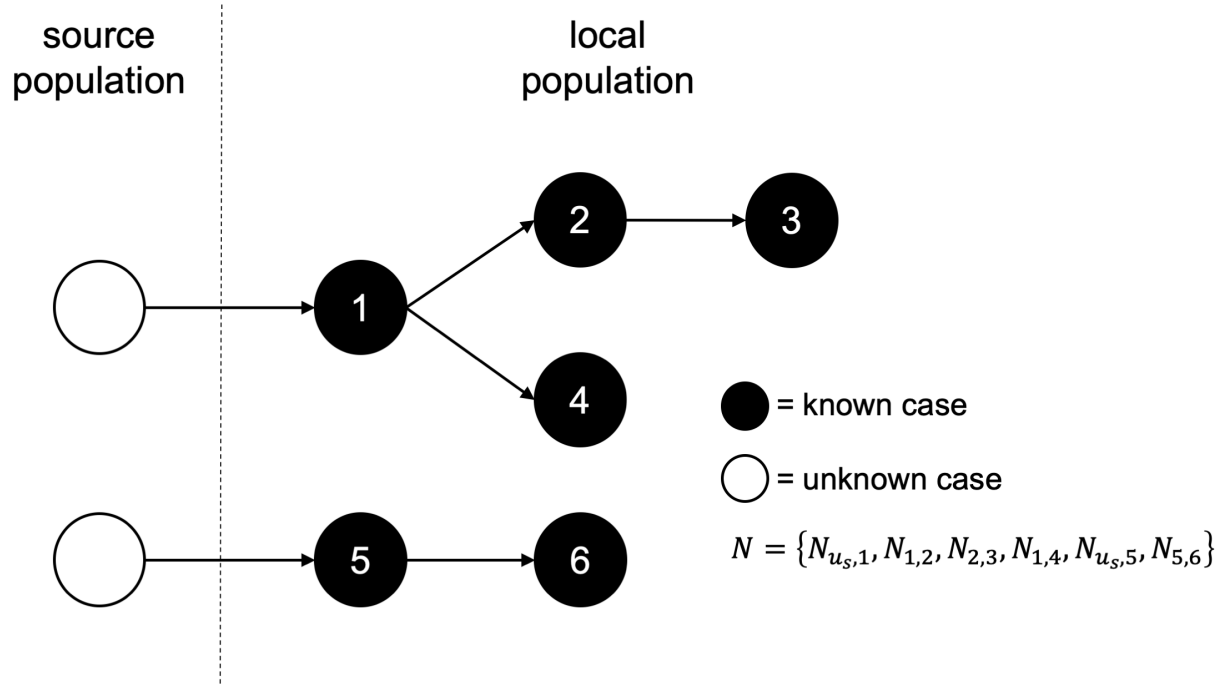


Fig 8. Schematic of a hypothetical transmission network. *A hypothetical transmission network is presented along with the corresponding notation. In the schematic, white circles denote unobserved cases, and black circle denote observed cases. Arrows represent transmission between two cases.*

To estimate N, we used spatial, temporal, and travel-history data about all cases, denoted as $\vec{X}_s$, X_t, and X_h, respectively. We did so within a Bayesian statistical framework, meaning that we sought to estimate the joint posterior probability density,

$$\Pr\left(N, \Theta \middle| X_t, \vec{X}_s, X_h\right) = \frac{\Pr\left(X_t, \vec{X}_s, X_h \middle| N, \Theta\right) \Pr\left(\mathrm{N}, \Theta\right)}{\Pr\left(X_t, \vec{X}_s, X_h\right)}, \quad (1)$$

of the transmission network defined by N and the model parameters Θ conditional on the data $\vec{X}_s$, X_t, and X_h. The first term in the numerator of eq. (1) is the likelihood of N and Θ conditional on the data. The second term in the numerator is the prior probability of N and Θ. The term in the denominator is the probability of the data, which is an intractable quantity to calculate directly given that it would require evaluation of an extremely high-dimensional integral over N and Θ. To address this, we used a Markov chain Monte Carlo algorithm to draw random samples of N and Θ from the posterior distribution specified in eq. (1).

The most critical piece of our inference framework is the likelihood, which we define as a function of each case j as

$$\mathcal{L}(N, \Theta | X_t, \vec{X}_s, X_h) = \prod_j \Pr\left(X_{t,j}, \vec{X}_{s,j}, X_{h,j} \middle| N_{.,j}, \Theta\right). \quad (2)$$

Below, we define the probability of the data associated with each known case *j* as a function of different assumptions that are possible about how case *j* is connected to the rest of the transmission network.

Scenario 1: Local transmission between known cases *i* and *j*

When case *i* contains parasites that are immediate ancestors of the parasites contained in case *j*, we represent its contribution to the likelihood as

$$\Pr\left(X_{t,j}, \vec{X}_{s,j}, X_{h,j} \middle| N_{i,j}, \Theta\right) = \Pr\left(X_{t,j} \middle| N_{i,j}, \Theta\right) \Pr\left(\vec{X}_{s,j} \middle| X_{t,j}, N_{i,j}, \Theta\right) \Pr\left(X_{h,j} \middle| N_{i,j}, \Theta\right), \quad (3)$$

which is the product of the probabilities of the temporal, spatial, and travel-history data given the network and model parameters. This formulation assumes that those data are generated independently for each individual, with the exception of a dependence of the spatial data on the temporal data.

Probability of the temporal data. To characterize the time elapsed between two cases resulting from local transmission, we used a model of the generation and serial intervals for *P. falciparum* malaria by Huber *et al*.[22]. The generation interval represents the time between infection of a primary and secondary case, whereas the serial interval represents the time between detection of those cases. Because the timing of infection per se (i.e., an infectious mosquito inoculating a

susceptible human) is typically unknown, we focused on the serial interval as the most apropos temporal quantity relating cases.

In deriving the probability of a given length of the serial interval, Huber *et al.*[22] convolved a random variable representing variability in the generation interval (*GI*) with a random variable representing variability in the time between infection with *P. falciparum* and detection by surveillance – i.e., the infection to detection period (*IDP*). That framework yielded

$$\Pr\left(SI_{i,j} = -a + b + c\right) = \sum_{a}\sum_{b}\sum_{c} \Pr(IDP_i = a)\Pr\left(GI_{i,j} = b\right)\Pr\left(IDP_j = c\right), \quad (4)$$

as the probability of a serial interval of length $SI_{i,j}$. We allowed for different models of the serial interval depending upon differences in the *GI* and *IDP* for different types of primary and secondary cases. For instance, we assumed that symptomatic cases present in a clinic some number of days after infection as informed by empirical data from Zanzibar[22]. For an asymptomatic infection, we assumed that detection occurred through active surveillance at a randomly drawn day among all days where its asexual parasitemia exceeds a detection threshold[22]. The choice of *IDP* for both the primary and secondary case informs the probability of two cases separated in time by $SI_{i,j} = X_{t,j} - X_{t,i}$ days.

Probability of the spatial data. Following Reiner *et al.*[6], we assumed that a simple two-dimensional Wiener diffusion process determines the location of secondary cases relative to the location of their associated primary case. It follows that, for a given diffusion coefficient *D* with units $km^2 day^{-1}$ and generation interval $GI_{i,j}$, the two-dimensional location $\vec{X}_{s,j}$ of the secondary case *j* is described by a bivariate normal distribution with probability density

$$f\left(\vec{X}_{s,j} \,\middle|\, \vec{X}_{s,i}, D, GI_{i,j}, N_{i,j}, \Theta\right) = \frac{1}{2\pi\sigma^2\left(GI_{i,j}\right)} e^{-\frac{\|\vec{X}_{s,j}-\vec{X}_{s,i}\|}{2\sigma^2\left(GI_{i,j}\right)}}, \quad (5)$$

where $\sigma^2\left(GI_{i,j}\right) = DGI_{i,j}$. This formulation assumes that each spatial dimension is independent, that the variance scales linearly with the generation interval, and that movement is isotropic across a continuous landscape.

One complication to eq. (5) is that the generation interval $GI_{i,j}$ is unobserved and, therefore, cannot take on a fixed value. Instead, we must use data about the serial interval $SI_{i,j}$ to inform our generative model for $\vec{X}_{s,j}$. To do so, we take advantage of the property of normal random variables that the sum of two or more random variables is itself a normal random variable[46]. This property allows us to recast eq. (5) as a function of *SI* rather than *GI* by computing the appropriate σ^2 as

$$\sigma^2(SI) = \int \sigma^2(GI)\,\Pr(GI|SI)\,dGI, \quad (6)$$

which is effectively a weighted sum of the spatial variances associated with a given *GI* proportional to the probability of each *GI* conditional on the observed *SI*. This results in

$$f\left(\vec{X}_{s,j} \,\middle|\, \vec{X}_{s,i}, D, SI_{i,j}, N_{i,j}, \Theta\right) = \frac{1}{2\pi\sigma^2\left(SI_{i,j}\right)} e^{-\frac{\|\vec{X}_{s,j}-\vec{X}_{s,i}\|}{2\sigma^2\left(SI_{i,j}\right)}}, \quad (7)$$

as the probability density of the spatial data that we assume.

In the event that case *i* has missing spatial data, we cannot compute the spatial likelihood of eq. (7). To address this, we define a latent unobserved quantity $\tilde{X}_{s,i}$, which represents the unknown location of case *i*. We then integrate over the uncertainty in $\tilde{X}_{s,i}$,

$$f\left(\vec{X}_{s,j}\middle|D, SI_{i,j}, N_{i,j}, \Theta\right) = \int f\left(\vec{X}_{s,j}\middle|\tilde{X}_{s,i}, D, N_{i,j}, \Theta\right) f\left(\tilde{X}_{s,i}\middle|\vec{X}_{s,j}, D\right) d\,\tilde{X}_{s,i}, \quad (8)$$

to compute the probability density of case *j* with known spatial location $\vec{X}_{s,j}$ arising from case *i* with unknown spatial location $\tilde{X}_{s,i}$. Equation (8) is computed as the product of the probability density of the location of a known case *j* conditional on an unknown location $\tilde{X}_{s,i}$ and the probability density of spatial separation $\vec{X}_{s,j} - \tilde{X}_{s,i}$ conditional on the diffusion coefficient *D* for all $\tilde{X}_{s,i}$. Because we assume that movement is isotropic, eq. (8) is a two-dimensional Gaussian integral, simplifying to

$$f\left(\vec{X}_{s,j}\middle|D, SI_{i,j}, N_{i,j}, \Theta\right) = \frac{1}{4\pi\sigma^2(SI_{i,j})}. \quad (9)$$

In the event that case *j* has missing spatial data and case *i* has known spatial data, the latent unobserved quantity becomes $\tilde{X}_{s,j}$. We then integrate over the uncertainty in $\tilde{X}_{s,j}$ and calculate $f\left(\vec{X}_{s,j}\middle|D, SI_{i,j}, N_{i,j}, \Theta\right)$ using eq. (8-9).

Probability of the travel-history data. Although we assume in this scenario that a person's infection was locally acquired, our model must still be capable of explaining the travel-history

data $X_{h,j}$. We define a probability τ_l that case j reported travel (i.e., $X_{h,j} = 1$) even though they were not infected during that period of travel, such that

$$\Pr\left(X_{h,j}\middle|N_{i,j},\Theta\right) = \begin{cases}\tau_l, & X_{h,j} = 1\\ 1-\tau_l, & X_{h,j} = 0\end{cases}. \quad (10)$$

In the event that case j has missing travel-history data, we cannot compute the travel-history likelihood of eq. (10). To address this, we defined a latent unobserved quantity $\tilde{X}_{h,j}$, which represents the unknown travel history of case j. We then sum across the uncertainty in $\tilde{X}_{h,j}$,

$$\Pr\left(X_{h,j} = \mathrm{NA}\middle|N_{i,j},\Theta\right) = \Pr\left(\tilde{X}_{h,j} = 1\right)\tau_l + \left(1 - \Pr\left(\tilde{X}_{h,j} = 1\right)\right)(1-\tau_l), \quad (11)$$

to compute the probability that case j was locally acquired given an unknown travel history. In eq. (11), $\Pr\left(\tilde{X}_{h,j} = 1\right)$ was computed as the proportion of cases with a positive travel history among all cases with known travel-history data.

Taken together with the probabilities of the temporal and spatial data described above, the product of these three probabilities constitutes the entirety of the contribution of a case j infected by a known local case i to the overall likelihood of N and Θ.

Scenario 2: Importation of local case *j* from source population *s*

In the event of $N_{u_s,j}$, we represent the contribution of such a case to the overall likelihood of N and Θ as the product of the probabilities of its temporal, spatial, and travel-history data under

similar assumptions as in Scenario 1. The key difference in this scenario is that there is no information about the unknown source case that gave rise to case *j*.

Probability of the temporal data. Because the person containing parasites that are the direct ancestors of those in case *j* is unobserved and does not have an $X_{t,i}$, we are unable to compute the probability of the temporal data as described in Scenario 1. It is important though to obtain a probability comparable to that from Scenario 1 as a reference point for determining whether it is more likely that a given case arose from some other known local case or from an unknown case u_s from source population *s*. To do so, we consider the variable $\tilde{X}_{t,u_s}$, which is a latent variable describing the timing of when u_s would have been detected, had it been detected.

Because u_s is not observed, we considered it to be asymptomatic and untreated. We then calculated the probability of the timing of a known case *j* arising from an unknown case u_s as

$$\Pr\left(X_{t,j}\middle|N_{u_s,j},\Theta\right)=\int \Pr\left(X_{t,j}\middle|\tilde{X}_{t,u_s},N_{u_s,j},\Theta\right)\Pr\left(\mathrm{SI}=X_{t,j}-\tilde{X}_{t,u_s}\right)d\tilde{X}_{t,u_s}, \quad (12)$$

by integrating over uncertainty in $\tilde{X}_{t,u_s}$. We represented this as the product of the probability of the timing of a known case *j* conditional on an unknown time of detection $\tilde{X}_{t,u_s}$ and the probability of the serial interval $X_{t,j}-\tilde{X}_{t,u_s}$ for all $\tilde{X}_{t,u_s}$. In equation (12), we did not distinguish between symptomatic and asymptomatic cases *j* because the calculation is identical; only the serial interval distributions differ.

Probability of the spatial data. Without an $\tilde{X}_{t,u_s}$ for the unobserved case u_s, we lacked information on the serial interval between it and case j. Consequently, we were unable to use the probability from eq. (7) in that particular form. Instead, we computed the spatial variance as a function of the diffusion coefficient alone, yielding

$$\sigma^2(D) = \int D\ GI\ \Pr(GI)\ dGI. \quad (13)$$

Equation (13) integrates across all possible generation intervals and simplifies to $D\mathbb{E}[GI]$, the product of the diffusion coefficient and the expectation of the generation interval.

We applied this spatial variance to the unobserved latent variable $\tilde{X}_{s,u_s}$, which represents the unknown location of the unobserved case u_s. We integrated over uncertainty in $\tilde{X}_{s,u_s}$ to compute the probability density,

$$f\left(X_{s,j}\middle|D, N_{u_s,j}, \Theta\right) = \int f\left(X_{s,j}\middle|\tilde{X}_{s,u_s}, D, N_{u_s,j}, \Theta\right) f\left(\tilde{X}_{s,u_s}\middle|X_{s,j}, D\right) d\,\tilde{X}_{s,u_s}, \quad (14)$$

of the location of a known case j arising from an unknown source case u_s with unknown location $\tilde{X}_{s,u_s}$. This is represented as the product of the probability density of the location of a known case j conditional on an unknown location $\tilde{X}_{s,u_s}$ and the probability density of spatial separation $X_{s,j} - \tilde{X}_{s,u_s}$ conditional on the diffusion coefficient D for all $\tilde{X}_{s,u_s}$. As in eq. (9), we treated eq. (14) as an evaluation of the Gaussian integral, evaluating to

$$f\left(X_{s,j}\middle|D, N_{u_s,j}, \Theta\right) = \frac{1}{4\pi D \mathbb{E}[GI]}. \quad (15)$$

In eq. (15), D is the diffusion coefficient and $\mathbb{E}[GI]$ is the expectation of the generation interval.

Probability of the travel-history data. We considered the travel history $X_{h,j}$ to be a binary variable with a value of 1 indicating reported travel to an area with known or assumed malaria transmission within a timeframe consistent with the person having become infected there. After defining the probability τ_s that $X_{h,j} = 1$ conditional on $N_{u_s,j}$, it follows that

$$\Pr\left(X_{h,j}\middle|N_{u_s,j}, \Theta\right) = \begin{cases} \tau_s, & X_{h,j} = 1 \\ 1 - \tau_s, & X_{h,j} = 0 \end{cases}, \quad (16)$$

which constitutes the contribution of the travel history of such a case to the overall likelihood of N and Θ. If the travel history of case j is unknown, an analogous calculation to eq. (11) is made using τ_s.

Bayesian inference

Markov Chain Monte Carlo Algorithm. To avoid evaluating the high-dimensional integral over N and Θ, we drew samples of N and Θ from their posterior distribution defined by eq. (1) using a Metropolis-Hastings Markov chain Monte Carlo (MCMC) method[47,48]. To begin the chain, N and Θ were initialized to $N^{(1)}$ and $\Theta^{(1)}$, and each subsequent step i in the chain was

denoted $N^{(i)}$ and $\Theta^{(i)}$. At each step, states N' and Θ' were proposed with $\Pr\left(\left(N^{(i)}, \Theta^{(i)}\right) \rightarrow (N', \Theta')\right)$. Proposed states were accepted with probability

$$\alpha_{\text{update}} = \min\left[1, \frac{\pi(N', \Theta')\Pr\left((N', \Theta') \rightarrow \left(N^{(i)}, \Theta^{(i)}\right)\right)}{\pi(N^{(i)}, \Theta^{(i)})\Pr\left((N^{(i)}, \Theta^{(i)}) \rightarrow (N', \Theta')\right)}\right], \quad (17)$$

where $\pi(N, \Theta)$ is the product of the likelihood $\Pr(\vec{X}_s, X_t, X_h | N, \Theta)$ of N and Θ conditional on the data and the assumed prior probability $\Pr(N, \Theta)$ of N and Θ. After a random draw R from a uniform distribution, the chain was updated according to

$$N^{(i+1)}, \Theta^{(i+1)} = \begin{cases} N', \Theta', & R \leq \alpha \\ N^{(i)}, \Theta^{(i)}, & R > \alpha \end{cases}. \quad (18)$$

To reduce the probability of the chain becoming stuck at a local maximum, we employed Metropolis-coupled Markov chain Monte Carlo (MC^3)[49]. Implementing MC^3 involved running multiple chains in parallel, with $\pi_c(N, \Theta)$ in chain c raised to the power β_c according to

$$\beta_c = 1 + \lambda(c - 1), \quad (19)$$

where $\lambda > 0$ is a temperature increment parameter that governs the degree to which each chain is "heated." As a result of setting $\beta_1 = 1$, $\pi_1(N, \Theta)$ is directly proportional to the joint posterior distribution and is referred to as the master or "cold" chain. This algorithm effectively flattens the likelihood in the heated chains by setting $\beta_c > 1$, allowing them to explore the parameter

space more freely and to encounter alternative high-density regions more readily than the cold chain would alone. At a pre-defined frequency, two randomly selected chains *i* and *j* were allowed to swap parameter sets according to a swap probability

$$\alpha_{\text{swap}} = \min\left[1, \frac{\pi\left(N^{(j)}, \Theta^{(j)}\right)^{\beta_i} \pi\left(N^{(i)}, \Theta^{(i)}\right)^{\beta_j}}{\pi(N^{(i)}, \Theta^{(i)})^{\beta_i} \pi(N^{(j)}, \Theta^{(j)})^{\beta_j}}\right], \quad (20)$$

where $\pi(N, \Theta)$ is the same as it was in eq. (17). A swap into the master chain only occurred if it was from one of the two randomly selected chains and $R \leq \alpha_{\text{swap}}$. We recorded a total of 100 million samples from the posterior distribution, discarding the first 50 million samples as burn-in and thinning the chain every 10,000 samples between each recorded sample.

Proposals. Proposals made by the MC^3 algorithm involved changes to the parameters (i.e., D, τ_s, and τ_l) and changes to the transmission network topology. Each proposal occurred with a fixed probability, where the sum of these proposal probabilities was equal to one.

Proposals to change parameters involved updating D, τ_s, or τ_l. To update the value of D, a new value was drawn from a normal distribution with mean set to the current value of the parameter and variance set to 2.5. Values of D proposed must be strictly nonnegative, so we rejected any proposed D that was less than zero and assigned $\alpha_{\text{update}} = 0$. Similarly, new values of τ_s and τ_l were chosen according to normal distributions with means set to their current parameter value and variance set to 0.25. Because τ_s and τ_l are probabilities, we rejected any proposed value that fell outside the range [0,1] and assigned $\alpha_{\text{update}} = 0$.

Changes proposed to the network topology involved the addition or removal of an ancestor from a randomly selected node. We assigned a uniform probability of proposing case a as an ancestor to a randomly selected case i, such that proposals to the network topology are uninformed by spatial and temporal data. Furthermore, we defined the proposal probability of removing case a as an ancestor to a randomly selected case i as

$$\Pr(\text{remove } a) = \bar{A}_i^{-1}, \quad (21)$$

where $\bar{A}_i$ represents the size of set A_i of all ancestors to case i. Proposed changes to the network are then accepted according to eq. (17).

Prior assumptions. We placed strong priors on τ_s and τ_l, because we assumed that travel histories were mostly, but not completely, accurate. We used a beta-distributed prior on τ_s, with parameters $\alpha_{\tau_s} = 12$ and $\beta_{\tau_s} = 3$, which resulted in a mean of 0.8 and a variance of 0.01 for this prior distribution. We also used a beta distributed prior on τ_l, with parameters $\alpha_{\tau_l} = 3$ and $\beta_{\tau_l} = 12$, which resulted in a mean of 0.2 and a variance of 0.01. We assumed a uniform prior on D over the interval $[10^{-3}, \infty)$ and an even prior across all possible network configurations, meaning that those prior probabilities canceled out in eqs. (17) and (20).

Assessing convergence. For D, τ_s, and τ_l, we assessed convergence using the Gelman-Rubin statistic[50], with values below 1.1 indicating convergence. For the transmission network N, we assessed convergence by calculating correlation coefficients of case-level probabilities across five chains from independent realizations of the MC^3 algorithm, for a total of 10 pairwise

comparisons across the five chains. The two case-level probabilities that we considered were the posterior probability that each case was infected by an unknown case u_s from a source population and the posterior probability that each case j was infected by each other case i. Higher values of these correlation coefficients provided stronger support for convergence.

Ethical Considerations

Ethical approval was obtained from the Eswatini Ministry of Health, the University of California, San Francisco, and the University of Notre Dame (IRB 19-06-5408). All data were analyzed anonymously.

Data Availability

The code and simulated data to reproduce the analyses can be found at https://github.com/johnhhuber/SpaceTime_Networks. The data collected from Eswatini contains sensitive household locations and are unable to be shared due to institutional review board restrictions.

Competing Interests

The authors have no competing interests to declare.

Acknowledgments

The authors thank the National Malaria Elimination Programme as well as Nontokozo Mngadi and Deepa Pindolia from the Clinton Health Access Initiative for their support in the collection of the surveillance data used in this study. We also thank Brooke Whittemore for her support in

data management. JHH acknowledges support from a National Science Foundation Graduate Research Fellowship and a Richard and Peggy Notebaert Premier Fellowship. BG and TAP received support from a grant from the Bill and Melinda Gates Foundation (OPP 1132226 to BG). MSH received support from NIAID (AI101012). The funders had no role in study design, analysis, decision to publish, or preparation of the manuscript.

Author Contributions

J.H.H, T.A.P, and B.G. conceived of the study. M.S.H., N.D., S.V., N.N., N.N, and B.G. curated the data. J.H.H., M.S.H., M.M., A.L., B.G., and T.A.P performed the formal analysis. Funding was acquired by B.G. and M.S.H. Investigation was by J.H.H, B.G., and T.A.P. Methodology was developed by J.H.H., M.M., A.L., R.N, B.G, and T.A.P. Project administration was by J.H.H. and T.A.P. J.H.H worked with the software. T.A.P and B.G. supervised the project. J.H.H, B.G., and T.A.P wrote the original draft of the manuscript. All authors contributed to the final version of the manuscript.

Materials and Correspondence

Please direct all requests to jhuber3@nd.edu (JHH) and taperkins@nd.edu (TAP).

References

1. Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516 (2004).
2. White, L. F., Archer, B. & Pagano, M. Estimating the reproductive number in the presence of spatial heterogeneity of transmission patterns. *Int J Health Geogr* **12**, 35 (2013).

3. Métras, R. *et al.* Transmission Potential of Rift Valley Fever Virus over the Course of the 2010 Epidemic in South Africa. *Emerg. Infect. Dis.* **19**, 916–924 (2013).
4. Backer, J. A. & Wallinga, J. Spatiotemporal Analysis of the 2014 Ebola Epidemic in West Africa. *PLoS Comput Biol* **12**, e1005210 (2016).
5. Salje, H., Cummings, D. A. T. & Lessler, J. Estimating infectious disease transmission distances using the overall distribution of cases. *Epidemics* **17**, 10–18 (2016).
6. Reiner, R. C. *et al.* Mapping residual transmission for malaria elimination. *eLife* **4**, (2015).
7. Lau, M. S. Y. *et al.* Spatial and temporal dynamics of superspreading events in the 2014-2015 West Africa Ebola epidemic. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2337–2342 (2017).
8. Lau, M. S. Y. *et al.* A mechanistic spatio-temporal framework for modelling individual-to-individual transmission—With an application to the 2014-2015 West Africa Ebola outbreak. *PLoS Comput Biol* **13**, e1005798 (2017).
9. Jombart, T. *et al.* Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Comput Biol* **10**, e1003457 (2014).
10. Didelot, X., Gardy, J. & Colijn, C. Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. *Molecular Biology and Evolution* **31**, 1869–1879 (2014).
11. Mollentze, N. *et al.* A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. Biol. Sci.* **281**, 20133251 (2014).
12. Ypma, R. J. F., van Ballegooijen, W. M. & Wallinga, J. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. *Genetics* **195**, 1055–1062 (2013).

13. Sturrock, H. J. W. *et al.* Targeting Asymptomatic Malaria Infections: Active Surveillance in Control and Elimination. *PLoS Med* **10**, e1001467 (2013).

14. Bousema, T. *et al.* Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination. *PLoS Med* **9**, e1001165 (2012).

15. Bejon, P. *et al.* A micro-epidemiological analysis of febrile malaria in Coastal Kenya showing hotspots within hotspots. *eLife* **3**, e02130 (2014).

16. Cohen, J. M., Moonen, B., Snow, R. W. & Smith, D. L. How absolute is zero? An evaluation of historical and current definitions of malaria elimination. *Malar J* **9**, 213 (2010).

17. Cohen, J. M. *et al.* Mapping multiple components of malaria risk for improved targeting of elimination interventions. *Malar J* **16**, 459 (2017).

18. Churcher, T. S. *et al.* Measuring the path toward malaria elimination. *Science* **344**, 1230–1232 (2014).

19. Routledge, I. *et al.* Estimating spatiotemporally varying malaria reproduction numbers in a near elimination setting. *Nat Commun* **9**, 2476 (2018).

20. Routledge, I. *et al.* Tracking progress towards malaria elimination in China: Individual-level estimates of transmission and its spatiotemporal variation using a diffusion network approach. *PLoS Comput Biol* **16**, e1007707 (2020).

21. Campbell, F., Strang, C., Ferguson, N., Cori, A. & Jombart, T. When are pathogen genome sequences informative of transmission events? *PLoS Pathog* **14**, e1006885 (2018).

22. Huber, J. H., Johnston, G. L., Greenhouse, B., Smith, D. L. & Perkins, T. A. Quantitative, model-based estimates of variability in the generation and serial intervals of Plasmodium falciparum malaria. *Malaria Journal* **15**, (2016).

787 23. Marshall, J. M., Bennett, A., Kiware, S. S. & Sturrock, H. J. W. The Hitchhiking Parasite:
788 Why Human Movement Matters to Malaria Transmission and What We Can Do About It.
789 *Trends in Parasitology* **32**, 752–755 (2016).
790 24. Campbell, F., Cori, A., Ferguson, N. & Jombart, T. Bayesian inference of transmission
791 chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput Biol*
792 **15**, e1006930 (2019).
793 25. Wood, S. N. *Generalized additive models: an introduction with R*. (Chapman & Hall/CRC,
794 2006).
795 26. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R
796 Foundation for Statistical Computing, 2017).
797 27. Cohen, J. M. *et al.* Rapid case-based mapping of seasonal malaria transmission risk for
798 strategic elimination planning in Swaziland. *Malar J* **12**, 61 (2013).
799 28. Hsiang, M. S. *et al.* Effectiveness of reactive focal mass drug administration and reactive
800 focal vector control to reduce malaria transmission in the low malaria-endemic setting of
801 Namibia: a cluster-randomised controlled, open-label, two-by-two factorial design trial. *The*
802 *Lancet* **395**, 1361–1373 (2020).
803 29. Sturrock, H. J. *et al.* Fine-scale malaria risk mapping from routine aggregated case data.
804 *Malar J* **13**, 421 (2014).
805 30. Wesolowski, A. *et al.* Quantifying travel behavior for infectious disease research: a
806 comparison of data from surveys and mobile phones. *Sci Rep* **4**, 5678 (2015).
807 31. Tessema, S. *et al.* Using parasite genetic and human mobility data to infer local and cross-
808 border malaria connectivity in Southern Africa. *eLife* **8**, e43510 (2019).

32. Marshall, J. M. *et al.* Mathematical models of human mobility of relevance to malaria transmission in Africa. *Sci Rep* **8**, 7713 (2018).

33. Blumberg, S. & Lloyd-Smith, J. O. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. *PLoS Comput Biol* **9**, e1002993 (2013).

34. Blumberg, S. & Lloyd-Smith, J. O. Comparing methods for estimating R0 from the size distribution of subcritical transmission chains. *Epidemics* **5**, 131–145 (2013).

35. Robert, A., Kucharski, A. J., Gastañaduy, P. A., Paul, P. & Funk, S. Probabilistic reconstruction of measles transmission clusters from routinely collected surveillance data. *J. R. Soc. Interface.* **17**, 20200084 (2020).

36. Wesolowski, A. *et al.* Quantifying the Impact of Human Mobility on Malaria. *Science* **338**, 267–270 (2012).

37. Weiss, D. J. *et al.* A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336 (2018).

38. Wesolowski, A. *et al.* Mapping malaria by combining parasite genomic and epidemiologic data. *BMC Med* **16**, 190 (2018).

39. Roh, M. E. *et al.* High Genetic Diversity of Plasmodium falciparum in the Low-Transmission Setting of the Kingdom of Eswatini. *The Journal of Infectious Diseases* **220**, 1346–1354 (2019).

40. Chang, H.-H. *et al.* Mapping imported malaria in Bangladesh using parasite genetic and human mobility data. *eLife* **8**, e43481 (2019).

41. Guzzetta, G., Marques-Toledo, C. A., Rosà, R., Teixeira, M. & Merler, S. Quantifying the spatial spread of dengue in a non-endemic Brazilian metropolis via transmission chain reconstruction. *Nature Communications* **9**, (2018).

42. Salje, H. *et al.* How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study. *Proc Natl Acad Sci USA* **113**, 13420–13425 (2016).

43. Guzzetta, G. *et al.* Spatial modes for transmission of chikungunya virus during a large chikungunya outbreak in Italy: a modeling analysis. *BMC Med* **18**, 226 (2020).

44. Weiss, D. J. *et al.* Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17: a spatial and temporal modelling study. *The Lancet* **394**, 322–331 (2019).

45. Greenhouse, B., Smith, D. L., Rodríguez-Barraquer, I., Mueller, I. & Drakeley, C. J. Taking Sharper Pictures of Malaria with CAMERAs: Combined Antibodies to Measure Exposure Recency Assays. *The American Journal of Tropical Medicine and Hygiene* **99**, 1120–1127 (2018).

46. Lemons, D. S. & Langevin, P. *An introduction to stochastic processes in physics: containing 'On the theory of Brownian motion' by Paul Langevin, translated by Anthony Gythiel*. (Johns Hopkins University Press, 2002).

47. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21**, 1087–1092 (1953).

48. Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97 (1970).

49. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407–415 (2004).

855 50. Gelman, A. & Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences.

856 *Statistical Science* **7**, 457–472 (1992).

857