Inferring the epidemiologic sources of infection from cross-sectionally sampled pathogen sequence data

Xiaoyue Xi

Department of Mathematics, Imperial College London, London SW72AZ, United Kingdom.

Simon EF Spencer

Department of Statistics, University of Warwick, Coventry CV47AL, United Kingdom.

M Kate Grabowski

Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA; Rakai Health Sciences Program, Entebbe, Uganda.

Oliver Ratmann

Department of Mathematics, Imperial College London, London SW72AZ, United Kingdom.

E-mail: oliver.ratmann@imperial.ac.uk

on behalf of Rakai Health Sciences Program and PANGEA-HIV

Summary.

1. Introduction

1.1. The main burden of HIV transmission is in sub-Saharan Africa

HIV remains one of the largest public health threats, especially in sub-Saharan Africa where approximately 66% of all new cases worldwide occur (UNAIDS, 2018). In most countries, such as Uganda or Zambia, the epidemic spreads among sexually active men and women in general, reaching an HIV prevalence of 40% in particular areas such as fishing communities on Lake Victoria (Chang et al., 2016).

A large number of HIV interventions are now available to stop viral spread effectively, such as voluntary medical male circumcision (VMMC) to reduce the risk of HIV acquisition in men, or immediate provision of antiretroviral therapy (ART) to suppress the virus in infected individuals and thereby stop onward transmission (Grabowski et al.,

2017; Kagaayi et al., 2019). Since the widespread adoption of such prevention measures, rates of incident cases have overall dropped considerably, although they remain too high for HIV elimination (UNAIDS, 2014b).

1.2. Inferring the sources, sinks and hubs of transmission flows to aid the design of HIV prevention interventions

In this context, there is increasing focus on identifying groups of individuals that are at high risk of acquiring HIV and at high risk of spreading the virus, and then to target tailored control interventions to these groups (UNAIDS, 2014a). Conceptually, the first step in this strategy is to break down the epidemic into source, sink and hub populations, according to the transmission flows that occur between them (figure 1). Sources are population groups that disproportionally pass on infection, sinks are groups that disproportionally acquire infection, and hubs are both sources and sinks (Abeler-Dörner et al., 2019). The population groups can be defined in various ways.

For instance, Dwyer-Lindgren et al. (2019) provided sub-national estimates of HIV prevalence across Africa, adding to data showing that the epidemic is highly heterogeneous across Africa, with small areas of very high prevalence (hotspots) that are surrounded by neighbouring areas with substantially lower prevalence. The WHO and others are recommending to target interventions to hotspots to maximise cost-effectiveness of interventions and reduce infection burden in hotspots (UNAIDS, 2014a; The Office of the US Global AIDS Coordinator, 2014). Although often assumed, it is not clear if hotspots are also sources of epidemic spread to the neighbouring lower-prevalence communities (Ratmann et al., 2020). If they are, spatial targeting could have a substantial knock-on effect in reducing infection in the broader population (Aral et al., 2015). In this application, interest centres on dividing the full population into individuals living in high-prevalence areas (h) and low-prevalence areas (l), and estimate the transmission flows within and between them while accounting for transmission flow from/to external locations (e),

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_{hh} & \pi_{hl} \\ \pi_{lh} & \pi_{ll} \end{pmatrix}, \tag{1}$$

where π_{ab} denotes the proportion of transmission flows from group a to group b subject

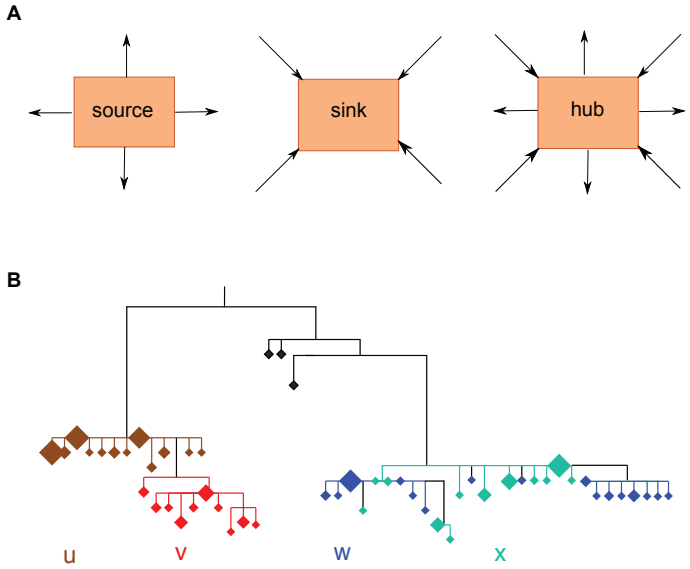

Fig. 1. Analysis aims and sketch of deep-sequence phylogenetic data to address these aims. (**A**) The overall aim of phylogenetic source attribution analyses is to infer how pathogens are passed on between population groups. Conceptually populations can be divided into source populations that predominantly transmit disease, sink populations that predominantly receive infection, and hub populations that both disproportionally transmit and receive infections. **(B)** Viral deep-sequencing generates many sequence samples per host, which can be used to establish phylogenetic orderings between individuals, and thereby estimate the direction of pathogen spread among sampled individuals. The figure sketches a deep-sequence phylogeny of pathogen sequences from individuals u, v, w, and x. Each tip (diamonds) represents a unique sequence, and the size of the tip copy number. Black tips correspond to out-of-sample reference sequences. Phylogenetic lineages are attributed to individuals (colours) using ancestral state reconstruction. Black lineages cannot be attributed to individuals. The subgraph of the tree associated with individual u is ancestral to that of v, suggesting that infection spread from u to v potentially via unsampled intermediates. Individual w has five subgraphs, some of which are ancestral to those of x and some of which are descendent from those of x, indicating a complex ordering from which the direction of infection spread cannot be inferred. Subgraphs of v and w are not phylogenetically adjacent (disconnected), suggesting that one did not infect the other. With such information from a population-based sample of infected individuals, it is possible to quantify population-level transmission flows, sources, sinks, and hubs.

to $\sum_{ab} \pi_{ab} = 1$.

Another prominent application of this framework concerns the interruption of infection cycles between men and women of different ages. De Oliveira et al. (2017) proposed the scenario that young women aged <25 years are predominantly infected by older men aged 25-40 years, and later spread the virus to similarly aged men in their late twenties and early thirties. If true, combination prevention interventions that seek to reduce HIV acquisition in young women from older men, and seek to reduce transmission from

older men and women could break the transmission cycle (De Oliveira et al., 2017). In this application, the full population is divided by sex-specific age bands (we consider 1-year age bands between 15 to 49 years), and the main objective is to estimate the transmission flow vector between men and women across age groups,

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}^{mf} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\pi}^{fm} \end{pmatrix}, \quad \boldsymbol{\pi}^{mf} = \begin{pmatrix} \pi_{11}^{mf} & \cdots & \pi_{1K}^{mf} \\ \vdots & \ddots & \vdots \\ \pi_{K1}^{mf} & \cdots & \pi_{KK}^{mf} \end{pmatrix} \quad \boldsymbol{\pi}^{fm} = \begin{pmatrix} \pi_{11}^{fm} & \cdots & \pi_{1K}^{fm} \\ \vdots & \ddots & \vdots \\ \pi_{K1}^{fm} & \cdots & \pi_{KK}^{fm} \end{pmatrix}, \quad (2)$$

where π_{ab}^{mf} denotes the proportion of transmissions from men in age band a to women in age band b, and similarly for π_{ab}^{fm}. The flow matrix (2) has $2K^2$ non-zero entries to estimate, which for 1-year age bands amounts to 2450 variables. Important summary statistics are the vector of sources of infection in group b individuals ($\boldsymbol{\delta}_b$), for example in young women aged <25 years; the vector of recipients of infection from group a individuals ($\boldsymbol{\omega}^a$), for example from men aged 25-40 years; and flow ratios from a to b (γ_{ab}), for example the ratio of transmissions from high-prevalence to low-prevalence areas compared to transmissions from low-prevalence to high-prevalence areas. Respectively these quantities are defined by

$$\boldsymbol{\delta}^b = (\delta_a^b)_{a=1,\ldots,A}, \quad \delta_a^b = \pi_{ab} / \sum_c \pi_{cb}; \quad (3a)$$

$$\boldsymbol{\omega}^a = (\omega_b^a)_{b=1,\ldots,A}, \quad \omega_b^a = \pi_{ab} / \sum_c \pi_{ac}; \quad (3b)$$

$$\gamma_{ab} = \pi_{ab} / \pi_{ba}. \quad (3c)$$

1.3. Inference from pathogen sequence data

Here, we are concerned in estimating the quantities (1– 3) from a population-based sample of pathogen sequences, obtained from a sample of infected individuals of a cross-sectionally surveyed study population. The statistical model that we describe is applicable whenever the sampling frame is relatively short compared to the infectious disease dynamics, so that transmission flows can be considered constant in time.

Pathogen sequences are the result of an evolutionary process that occurs in infected individuals of an epidemic. Thus, for fast-evolving pathogens like HIV, the phylogenetic

relationship of pathogen sequences can be used in reverse to estimate transmission dynamics. These methods have, for example, been used to estimate the origins of HIV (Faria et al., 2014), to quantify the contribution of different disease phases to onward spread (Volz et al., 2013; Ratmann et al., 2016), to improve outbreak detection (Poon et al., 2016), and they have been applied to a large range of other infectious diseases (Uhlemann et al., 2014; Didelot et al., 2017; Dellicour et al., 2018).

Traditionally, HIV sequences are obtained through Sanger sequencing, which returns for each sample one consensus nucleotide sequence that captures the entire viral diversity in the sample from one individual. The genetic distance between two consensus sequences can be used to estimate if the corresponding two individuals are epidemiologically closely related (Hué et al., 2005), however the data are insufficient to estimate the direction of transmission between any two sampled individuals (Leitner and Romero-Severson, 2018). For this reason most available methods propose to infer the flow parameters (1– 3) indirectly from statistics of the entire phylogeny, usually the coalescent times (i.e. the times when two lineages coalesce into one, backwards in time) and the disease states of infected individuals at time of sampling (such as location or age in the two applications discussed above). In the mugration model (Lemey et al., 2009), the states of viral lineages at any time are described with a continuous-time Markov chain (CTMC) that is independent of the evolutionary process. Flow estimates can be obtained from the posterior distribution of the transition rates of the CTMC model via MCMC sampling, as well as posterior estimates of the phylogeny and the states of its lineages, which are latent variables in the model (Lemey et al., 2009). The MultiTypeTree model (Vaughan et al., 2014) removes the independence assumption that the evolutionary history of the genealogy is independent of population structure, however the induced model complexity often renders sampling the latent phylogeny and state history computationally infeasible. This limitation is addressed with the structured coalescent of Volz et al. (2009), which integrates over the state histories of the phylogeny and describes the marginal probabilities of each viral lineage to be in a particular state at a particular time. The changes in the state probabilities along lineages and through coalescent events are derived under an assumed ordinary differential equations (ODE) model of disease spread. The flow

parameters are obtained as by-products of the estimated latent states and parameters of the compartment model, and in general vary in time over the phylogenetic history. Müller et al. (2017) and others (Popinga et al., 2015) have clarified the assumptions that underlie the marginalisation of state histories into lineage state probabilities, which lead to more accurate maximum likelihood and Bayesian estimation routines (Müller et al., 2018; Volz and Siveroni, 2018). With this marginalisation trick, a greater range of flow models and data sets can be analysed, though computational run-times often remain on the order of several weeks for data from hundreds of individuals and moderately complex flow models (Stadler and Bonhoeffer, 2013; Ratmann et al., 2017; Rasmussen et al., 2018).

An alternative strategy for estimating the flow parameters (1– 3) involves phylogenetic analysis of multiple distinct pathogen sequences per infected host, because such data make possible to attribute viral lineages to individuals and infer their ancestral relationships, which can provide direct evidence into the direction of transmission between two individuals (Romero-Severson et al., 2016; Leitner and Romero-Severson, 2018). This strategy is becoming broadly applicable to fast-evolving pathogens such as HIV, because standard deep sequencing protocols (Gall et al., 2012; Bonsall et al., 2018; Zhang et al., 2020) generate thousands to millions of distinct pathogen sequence fragments. In particular the PANGEA consortium is providing access to deep sequences from $> 20,000$ HIV-infected individuals from across sub-Saharan Africa, including dense population-based samples from locations where the burden of HIV is among the highest in Africa (Abeler-Dörner et al., 2019). Prior work focused on software development (Wymant et al., 2017; Skums et al., 2018), validation of the bioinformatics protocol for inferring the direction of transmission (Ratmann et al., 2018; Rose et al., 2019; Todesco et al., 2019; Zhang et al., 2020), and reconstruction of partially observed transmission networks at the population level in the case of HIV (Ratmann et al., 2018).

The starting point of this paper is the output of a typical deep-sequence phylogenetic analysis (Wymant et al., 2017), which includes for each ordered pair of sampled individuals a viral phylogenetic measure in $[0, 1]$ that transmission likely occurred from the first to the second individual, possibly via unsampled intermediate individuals (phylo-

genetic direction scores). Here we show that such deep-sequence data offer distinctive advantages for estimating population-level transmission flows, when compared to viral phylogenetic approaches based on standard Sanger sequencing. First, the data enable us to present the estimation problem in terms of a class of hierarchical Bayesian Poisson models that can flexibly describe a range of epidemiological questions including transmission in space (1), or by age and sex (2). This class of models is in principle similar to previously developed, high-dimensional Bayesian models used for estimating contact rates from non-Gaussian response data (van de Kassteele et al., 2017). Second, the models can account for multi-level sampling heterogeneity, which are typically present in population-based disease occurrence data. We leverage Bayesian data augmentation to adjust for sampling heterogeneity (Givens et al., 1997), and exploit the fact that the additional latent variables can be integrated out in our framework, so that computational inference remains inexpensive. Third, while typical phylodynamic approaches are limited to estimating transmission flows between coarse population strata, for example by age brackets $15-24$ years and $25-40$ years (De Oliveira et al., 2017; Le Vu et al., 2019), we can employ Gaussian-process-based regularisation techniques to capture fine detail in transmission flows by annual age increments. To illustrate, we present these innovations on combined epidemiologic and HIV deep sequence data from the Rakai Community Cohort Study (RCCS) of the Rakai Health Science program, situated in south-eastern Uganda (Grabowski et al., 2017; Chang et al., 2016). Between August 10 2011 to January 30 2015, virus from 2652 HIV-infected individuals could be deep-sequenced, and 293 pairs of individuals with phylogenetically strong support for the direction of transmission were identified. We demonstrate that the new type of phylogenetic data and our statistical model enable characterisation of cross-sectional transmission flows at unprecendented detail while remaining computationally scalable.

2. Methodology

2.1. Notation and Definitions.

In this section we present the notation that is used to estimate transmission flows in a population $\mathcal{P}$ during some observation period $\mathcal{T} = [t_1, t_2]$. We define by $i = 1, \ldots, N$ the unique identifier of an individual in $\mathcal{P}$ during $\mathcal{T}$. Individual-level characteristics such

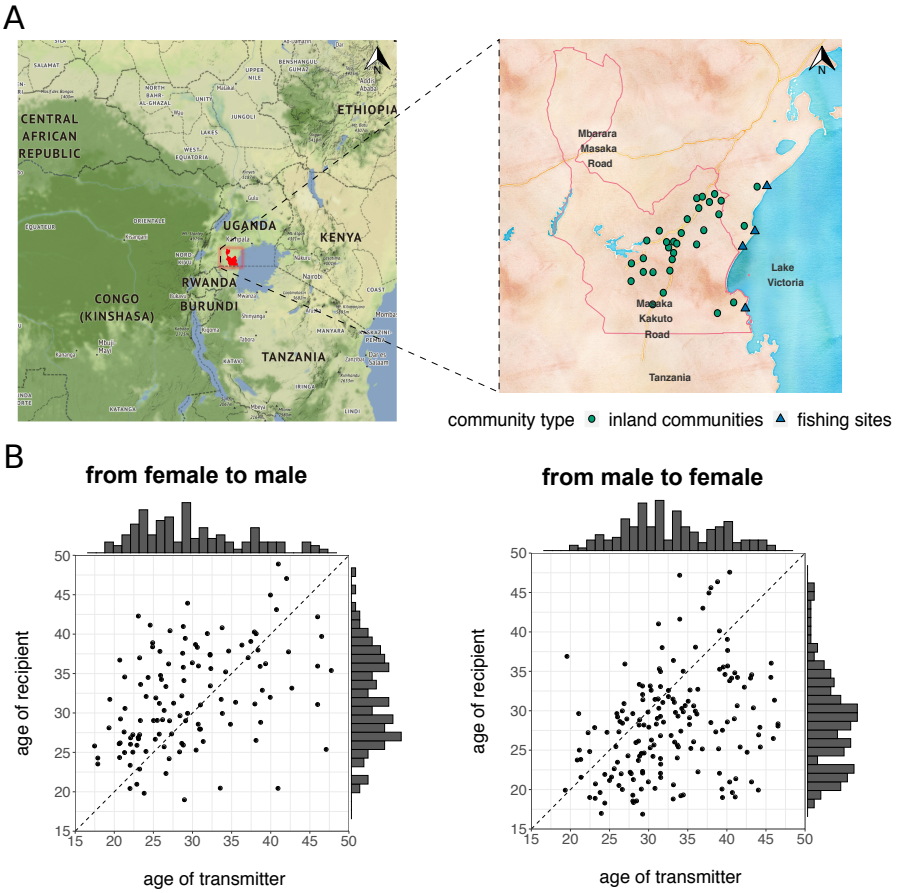

Fig. 2. Location of the Rakai Community Cohort Study, and data. (**A**) Location of Rakai District (red) in south-eastern Uganda at the shores of Lake Victoria. HIV surveillance data were obtained from 2 survey rounds in 36 inland communities of the Rakai Community Cohort Study (green circles) and three survey rounds in the main 4 fishing communities within 3km of Lake Victoria (green triangles) between August 10, 2011 and January 30, 2015. (**B**) The study did a household census, and all individuals aged 15-49 years capable to provide informed consent and resident for at least 1 month with the intention to stay were invited to participate. Viral deep-sequencing was performed on plasma blood samples from HIV infected participants who reported no ART use. Deep-sequence phylogenetic analysis returned phylogenetic transmission scores between individuals, and 293 pairs had strong support of phylogenetic linkage and transmission direction. 173 pairs were male-to-female and 120 were female-to-male. The figures show the phylogenetically likely source-recipient pairs by age of the source and recipient at the midpoint of the observation period.

as age or location of residence are described with p covariates, which we collect for all individuals in the $N \times p$ matrix $\mathbf{X}$. The sampling status of individual i is a binary variable, and individuals are either sampled ($s_i = 1$) or not ($s_i = 0$). We denote the sampling status vector for all individuals in $\mathcal{P}$ by $\boldsymbol{s} = (s_i)$. The number of sampled

individuals is N^s, which corresponds in our exposition to the 2652 individuals for whom a viral deep sequence is available for analysis. The output of the phylogenetic deep sequence analysis can thus be described in a $N^s \times N^s$ direction score matrix $\mathbf{W}$ that describes the evidence for transmission from i to j with the weight $w_{ij} \in [0, 1]$. The direction score matrix is in general not symmetric, and diagonal entries are zero.

We estimate transmission flows between population strata, defined for example by the cartesian product of gender and age (rounded to years), and denote the strata by a and the set of strata by $\mathcal{A}$, which is of dimension $A > 0$. Thus the primary object of interest is the $A \times A$ flow matrix $\boldsymbol{\pi}$, which is in general not symmetric, has positive diagonal entries, and is subject to $\sum_{a,b \in A} \pi_{ab} = 1$. The matrix may contain structural zeros, for example in the case of HIV female to female transmission is typically not observed [], and we denote the number of structurally non-zero entries by L, which satisfies $L \leq A^2$.

In general the flow matrix is time-dependent due to changes in population composition and varying transmission rates [Anderson]. For instance in a compartment model of susceptible (S), infected (I) and treated (T) men and women of high (h) and low risk (l) of onward transmission, the ODE equations pertaining to the male (m) high risk population are

$$\begin{aligned} \dot{S}_{mh} &= -\lambda(t) S_{mh}(t) + \mu - \mu S_{mh}(t) \\ \dot{I}_{mh} &= \lambda(t) I_{mh}(t) - \gamma(t) I_{mh}(t) - \mu I_{mh}(t) \\ \dot{T}_{mh} &= \gamma(t) I_{mh}(t) - \mu T_{mh}(t), \end{aligned} \tag{4}$$

where the force of infection is $\lambda(t) = \beta_{fh}(t) I_{fh}(t)/N_{fh}(t) + \beta_{fl}(t) I_{fl}(t)/N_{fl}(t)$, the birth/death rate μ is constant, and the viral suppression rate γ and transmission rates β_{fh}, β_{fl} are time-dependent. The actual, unobserved number of transmissions from high risk women to high risk men in $\mathcal{T} = [t_1, t_2]$ are

$$z_{fh,mh}([t_1, t_2]) = \int_{t_1}^{t_2} \beta_{fh}(t) I_{fh}(t) S_{mh}(t)/N_{fh}(t) dt, \tag{5}$$

and the corresponding proportion of transmissions is

$$\pi_{fh,mh}([t_1, t_2]) = \frac{z_{fh,mh}([t_1, t_2])}{Z([t_1, t_2])}, \tag{6}$$

where Z is the sum of transmission events in $\mathcal{T} = [t_1, t_2]$. Here, we focus on estimating transmission flows in a given observation window from data collected in the same period.

Such estimates provide a snapshot of transmission dynamics. For ease of notation we now drop the dependence of our data and estimates on $\mathcal{T} = [t_1, t_2]$.

As a first approach to estimating the flow matrix, consider the observed flow counts

$$n_{ab} = \sum_{i \in a, j \in b} \mathbb{1}\{s_i = 1\} \mathbb{1}\{s_j = 1\} \mathbb{1}\{w_{ij} > \zeta\} \in \mathbb{N}_0^+, \tag{7}$$

where $\zeta \in (0, 1)$ is a threshold that can be used to select phylogenetically highly supported source-recipient pairs. The counts can be arranged into the $A \times A$ count matrix $\mathbf{n}$, and sum to $N^f = \sum_{a,b} n_{ab}$. In previous studies, ζ was set to 0.5 or 0.6, and N^f was between 100 to 500 (Hall et al., 2019; Ratmann et al., 2020). The naïve flow estimator is defined by

$$\hat{\pi}_{ab} = \frac{n_{ab}}{\sum_{c,d} n_{cd}}. \tag{8}$$

If we suppose that each population group a is independently sampled at random with probability ξ_a, and the actual flows from group a to group b are z_{ab}, then $\mathbb{E}(\hat{\pi}_{ab}) = (z_{ab}\xi_a\xi_b)/(\sum_{c,d} z_{cd}\xi_c\xi_d)$. This clarifies that the naïve flow estimator (8) is only unbiased when the population groups were homogeneously sampled, i. e. ξ_a is the same for all a, which is rarely the case (Ratmann et al., 2020).

2.2. Inferring transmission flows from heterogeneously sampled deep-sequence data

Considering the actual, unobserved number of transmissions between all population groups, $\mathbf{z} = (z_{ab})$, and total $z^+ = \sum_{a,b} z_{ab}$, the complete data likelihood that arises under mathematical models of the form (4) in a fixed observation period is the multinomial

$$p(\mathbf{z}|z^+, \boldsymbol{\pi}) = \prod_{a,b} (\pi_{ab})^{z_{ab}}. \tag{9}$$

This model ignores potential second-order correlations between transmission events, for example that a female infected by an older male may be more likely to transmit to men of older age. For computational efficiency, we consider the related Poisson model

$$p(\mathbf{z}|\boldsymbol{\lambda}) = \prod_{a,b} (\lambda_{ab})^{z_{ab}} \exp(-\lambda_{ab}) = \left[\prod_{a,b} (\pi_{ab})^{z_{ab}}\right] \left[\eta^{z^+} \exp(-\eta)\right], \tag{10}$$

where λ_{ab} can be interpreted as the transmission intensities from group a to group b, $\eta = \sum_{c,d} \lambda_{cd}$, and π_{ab} are recovered via $\pi_{ab} = \lambda_{ab}/\eta$.

The actual transmission flows $\mathbf{z}$ are not observed. We assume that individuals are sampled at random within strata (SARWS) in a suitably chosen stratification of the population. SAWRS implies in particular that sampling is independent of transmission status (i.e. being a transmitter or not), and the likelihood of the observed counts conditional on the complete data is $p(\mathbf{n}|\mathbf{z}, \boldsymbol{\xi}) = \prod_{a,b} \text{Binomial}(n_{ab}; z_{ab}, \xi_a \xi_b)$, where ξ_a is the sampling probability in group a. In this class of models the latent transmission counts z_{ab} can be conveniently integrated out, yielding for the observed flow counts the Poisson model

$$p(\mathbf{n}|\boldsymbol{\lambda}, \boldsymbol{\xi}) = \prod_{a,b} (\lambda_{ab} \xi_a \xi_b)^{n_{ab}} \exp(-\lambda_{ab} \xi_a \xi_b). \tag{11}$$

The sampling-adjusted maximum-likelihood estimates of λ_{ab} and π_{ab} under (11) can be derived under the SARWS assumption. The number of sampled individuals in a, $N_a^s = \sum_{i \in a} \mathbb{1}\{s_i = 1\}$, is a Binomial sample of the number of all individuals in a, N_a, which leads to

$$\hat{\pi}_{ab} = \frac{n_{ab}}{\hat{\xi}_a \hat{\xi}_b} \Big/ \Big[\sum_{c,d} \frac{n_{cd}}{\hat{\xi}_c \hat{\xi}_d} \Big] \tag{12}$$

where $\hat{\xi}_a = N_a^s / N_a$ (Supplementary Text Sxx).

2.3. Regularisation

Bayesian regularisation techniques play a central role in obtaining robust and suitably smoothed flow estimates. Considering population sampling, we exploit additional information on the sampling vector $\boldsymbol{s}$. We assume that transmission is independent of sampling, allowing us to decompose the joint posterior distribution into

$$\begin{aligned} p(\boldsymbol{\lambda}, \boldsymbol{\xi}|\mathbf{n}, \boldsymbol{s}, \mathbf{X}) &\propto p(\mathbf{n}|\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{s}, \mathbf{X}) p(\boldsymbol{\lambda}|\boldsymbol{\xi}, \boldsymbol{s}, \mathbf{X}) p(\boldsymbol{\xi}|\boldsymbol{s}, \mathbf{X}) \\ &= p(\mathbf{n}|\boldsymbol{\lambda}, \boldsymbol{\xi}) p(\boldsymbol{\lambda}|\boldsymbol{\xi}) p(\boldsymbol{\xi}|\boldsymbol{s}, \mathbf{X}). \end{aligned} \tag{13}$$

A possible limitation of (12) is that the counts N_a, N_a^s can be small when the population is finely stratified. It is thus often advantageous to model individual-level sampling probabilities in terms of a linear combination of predictors. Using a logistic regression approach, we obtain

$$p(\xi_a|\boldsymbol{s}, \boldsymbol{X}) = \int \text{logit}^{-1}(\boldsymbol{x_a}\boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{s}, \boldsymbol{X}) d\boldsymbol{\beta}, \tag{14}$$

where $\boldsymbol{x_a}$ is the row vector of population characteristics that specify group a, $\boldsymbol{\beta}$ are the regression coefficients, and $p(\beta|\boldsymbol{s}, \boldsymbol{X})$ is the posterior density of the regression coefficients, estimated from the sampling status vector $\boldsymbol{s}$ of all individuals in the study population.

Considering the prior density on the transmission intensities $\boldsymbol{\lambda}$, in some applications the population strata are unordered such as when estimating transmission flows within and between high- and low-prevalence areas (1). In this case we propose using

$$\lambda_{ab}|\xi_a, \xi_b \sim \text{Gamma}(\alpha_{ab}, \beta), \quad \alpha_{ab} = 0.8/L, \beta = 0.8/Z^p(\xi_a, \xi_b), \tag{15}$$

where L is the number of structurally non-zero entries in $\boldsymbol{\pi}$ and Z^p is the number of expected transmission events, $Z^p = \sum_{a,b:n_{ab}\neq 0} \frac{n_{ab}}{\xi_a \xi_b} + \sum_{a,b:n_{ab}=0} \frac{1-\xi_a\xi_b}{\xi_a\xi_b}$. This choice is motivated by the fact that (15) induces on $\boldsymbol{\pi}$ an objective Dirichlet prior density with parameters $\alpha_{ab} = 0.8/L$ (Berger et al., 2015). However when the population groups can be ordered, such as the 1-year age bands in (2), the structure of the flow model (13) enables using regularising prior densities that penalise against large deviations in transmission intensities between similar source and recipient populations. For (2), we opted for two-dimensional Gaussian-process priors on the non-zero entries of $\boldsymbol{\lambda}$,

$$\begin{aligned}
&\log \boldsymbol{\lambda} = \mu + \mu^{mf} + \boldsymbol{f}, \quad \boldsymbol{f} \sim \mathcal{GP}(0, k), \\
&k\big((a_1, b_1), (a_2, b_2)\big) = \begin{cases} \sigma^2_{mf} \exp\Big(- \Big[\frac{(a_2-a_1)^2}{2\ell^2_{mf,a}} + \frac{(b_2-b_1)^2}{2\ell^2_{mf,b}} \Big] \Big) & \text{if } (a_1, b_1), (a_2, b_2) \in mf \\ \sigma^2_{fm} \exp\Big(- \Big[\frac{(a_2-a_1)^2}{2\ell^2_{fm,a}} + \frac{(b_2-b_1)^2}{2\ell^2_{fm,b}} \Big] \Big) & \text{if } (a_1, b_1), (a_2, b_2) \in fm \\ 0 & \text{otherwise} \end{cases} \\
&\sigma^2_{mf}, \sigma^2_{fm} \sim \text{Half-Normal}(0, 10) \\
&\ell_{d,i} \sim \text{Inv-Gamma}(\alpha_{d,i}, \beta_{d,i}),
\end{aligned} \tag{16}$$

where μ is the baseline log transmission intensity, μ^{mf} is a scalar on the elements of $\boldsymbol{\lambda}$ in the male-female direction, and k is a gender-specific squared exponential kernel with variance parameters $\sigma^2_{mf}, \sigma^2_{fm}$ and length scales $\ell_{mf,a}$, $\ell_{mf,b}$, $\ell_{fm,a}$, $\ell_{fm,b}$ (Rasmussen, 2003). The prior distributions on the length scale parameters were set so that their 95% credibility range matched the empirical 95% quantile range in Figure 2.

2.4. Numerical inference

The transmission flows and parameters for these highly structured Bayesian models can be efficiently estimated with the Hamiltonian Monte Carlo sampler of the Stan computing language (Carpenter et al., 2017). The implementation is available at `https://github.com/BDI-pathogens/phyloscanner/tree/master/phyloflows` and uses a spectral basis function approximation for the Gaussian process prior (16). For comparison a tailored Metropolis-within-Gibbs algorithm is also available. See Supplementary Text S2 for further details.

3. Applications

3.1. Bias in source attribution when sampling differences are ignored

We first assessed the accuracy in estimating transmission flows from biased flow data in a series of simulation experiments, that are fully reported in Supplementary Text S3. The first experiment was a minimal example involving flows between two population groups, which for simplicity we refer to as individuals in rural areas (group a) and individuals in large communities (group b). Transmission chains were simulated under the ODE model (4) in, and the simulated flow matrix $\boldsymbol{\pi}_r^T$ was recorded in $r = 1, \ldots, 100$ replicate simulations. The sampling probability in rural areas was $\xi_a = 60\%$, and sampling probabilities in large communities decreased from $\xi_b = 60\%$ to 35%, resulting in sampling differences of (0%, 5%, 10%, 15%, 20%, 25%). First, we estimated the marginal median posterior transmission flows $\widehat{\pi}_{r,ab}$ from (13) under the assumption of no sampling differences, which we implemented by setting $\xi_a = \xi_b = 0.5$. Second, we estimated the marginal median posterior transmission flows with information on sampling differences included, as in (**??**). Figure 3A compares the accuracy in transmission flow estimates when sampling differences are ignored (light grey bars) compared to when they are not ignored (dark grey bars) in terms of the worst case error (WCE) $\varepsilon_r = \max_{a,b} |\widehat{\pi}_{r,ab} - \pi^T_{r,ab}|$. For a 10% sampling difference between the two population groups, the median WCE was $6.5\%(3.5\% - 10.2\%)$ without adjusting for sampling differences, and $2.1\%(0.5\% - 5.1\%)$ after adjusting for sampling differences. More complex simulation experiments yielded similar results (Supplementary Text S3).

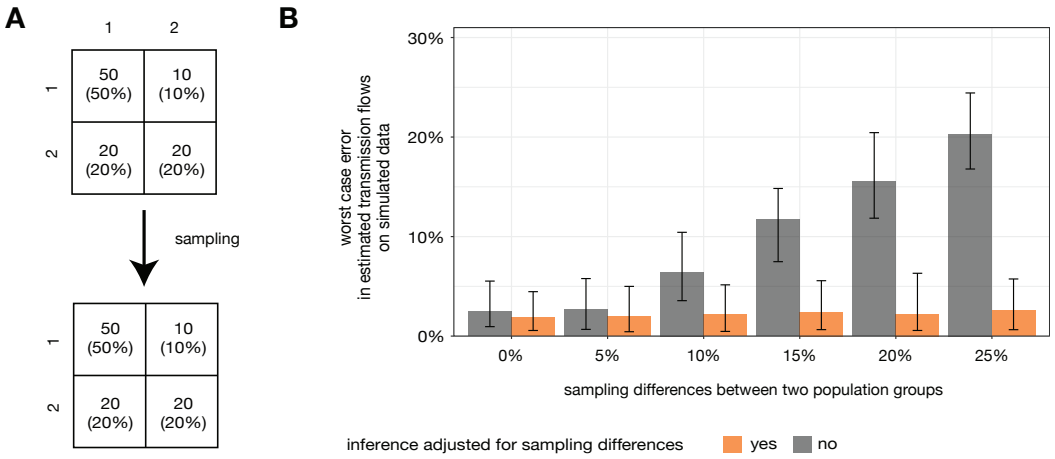

Fig. 3. Comparing errors in estimates of transmission flow. We assessed bias in estimating transmission flows with model (13) before and after accounting for sampling differences on simulations. (**A**) Design of simulation study. Transmission flows were simulated between two population groups according to an ODE model of disease spread. The number and proportion of actual transmissions was recorded (top 2 by 2 table, with source group in rows and recipient group in columns). The probability of sampling individuals in the first group was 60%, and decreased in the second group from 60% to 35%. Transmission events between two sampled events were recorded as observed (bottom 2 by 2 table). (**B**) The worst case error between the simulated transmission flows and median posterior estimates was calculated on 100 simulations in each sampling scenario, and is shown on the y-axis against increasing sampling differences on the x-axis. With increasing differences in sampling differences, we found increasing bias in flow estimates when sampling differences were not adjusted for. When sampling differences were adjusted for, the median worst case error in flow estimates remained on average below 5%.

3.2. Population-based deep-sequence data from Rakai, Uganda

We illustrate application of the Poisson flow model (13) on a population-based sample of HIV deep sequences from the RCCS in south-eastern Uganda at the shores of Lake Victoria (Ratmann et al., 2018, 2020). Between August 10 2011 to January 30 2015, 29116 individuals aged 15-49 years were eligible to participate in one of the 36 inland communities surveyed, and 8526 individuals in 4 fishing communities in population censuses immediately preceding the survey (Figure 2A). 19799 (68.0%) participated in inland communities, and 6083 (71.3%) in fishing communities. 11404 (96.9%) of non-participants were absent for school or work. Participation increased with age for both men and women among residents, was higher and more uniform among recent in-migrants by age, and was similar in fishing and inland communities (Supplementary Text S1). 2703 (13.6%) individuals were HIV-infected in inland communities on their last visit, and 2439 (40.1%) in fishing sites.

HIV sequencing was only feasible under the deep-sequencing protocol that we used when individuals had detectable virus (Gall et al., 2012), but viral load measurements were not available. We selected infected individuals who did not report ART use at their first visit for sequencing. In inland communities, 1803 (66.7%) infected participants did not report ART use, of whom 1138 (63.1%) could be deep-sequenced. In fishing communities, 2059 (84.4%) infected participants did not report ART use, and 1514 (73.6%) could be deep-sequenced (Ratmann et al., 2018). Deep-sequencing rates decreased with age for both men and women, were higher among men than women, and higher in fishing sites (Supplementary Text S1).

Using (7), there were 293 heterosexual pairs with phylogenetic support for linkage and direction of transmission above the threshold $\zeta = 0.6$ (source-recipient pairs), of whom 173 (59.0%) were male-to-female, and 120 (41.0%) were female-to-male. The estimated infection times of the recipients were between October 2009 and January 2015, which defined the observation period $\mathcal{T}$. Figure 2B illustrates the reconstructed source-recipient pairs by age of both individuals at the midpoint of the study period. We estimated that our population-based sample contained approximately 800 transmission pairs (Supplementary Text Sxx), suggesting that not all transmission pairs could be

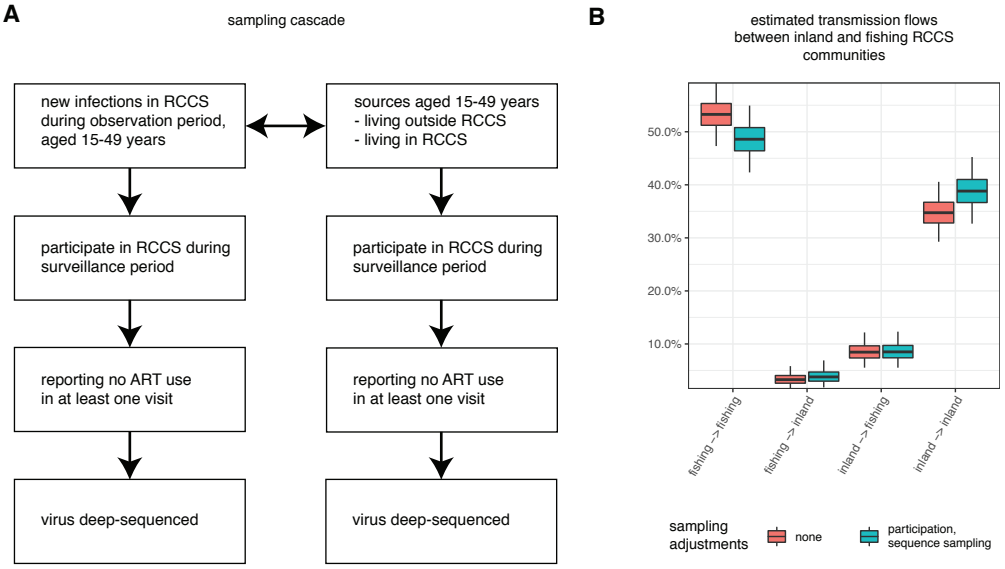

Fig. 4. Sampling cascade of transmission events, and impact on flow estimates. (**A**) The sampling cascade formalises the steps involved in sampling sources and recipients of transmission events that occurred in the observation period October 2009 to January 2015. Recipients were defined as individuals aged 15-49 years who acquired infection in one of the RCCS communities during the observation period. Sources were defined as individuals aged 15-49 years who transmitted to one of the recipients. Each arrow corresponds to a sampling step of the source and recipient populations, which are described in terms of individual-level predictive variables; see further Supplementary Text S1. (**B**) Estimated transmission flow estimates between inland and fishing communities of the RCCS. We first estimated transmission flows without sampling adjustments (red), using model (13) and setting the sampling probabilities ξ_a to the overall empirical average. We next iteratively drew sampling probabilities associated with each step of the sampling cascade for both sources and recipients, and used the resulting overall sampling probabilities to adjust transmission flow estimates with model (13). The resulting flow estimates (blue) indicated higher transmission flows within the inland communities of the RCCS and lower transmission flows within the fishing communities, which primarily reflected higher participation rates and higher sequence sampling rates in fishing communities.

identified through phylogenetic analysis. Using the thresholds $\zeta = 0.55$ and 0.5, 356 and 417 source-recipient pairs were found.

To interpret these observations, we formalised the individual steps in the sampling cascade of infection events to individuals aged 15-49 years between October 2009 and January 2015 in the RCCS and their sources (Figure 4A). The sources and recipients had to participate in at least one survey round between August 2011 and January 2015, report no ART use, and have virus sequenced successfully. For each step, we estimated based on detailed cohort data population-level sampling probabilities independently for sources and recipients under the assumption that sampling was independent of infection

and transmission status (i.e. being a source or recipient, Supplementary Text S1).

3.3. Transmission flows between areas with high and low disease prevalence

We used the source-recipient data of Figure 2B to address problem (1) and estimate transmission HIV flows within and between high- and low-prevalence RCCS communities. The high-prevalence communities comprised the four fishing communities, and the low-prevalence communities included the remaining 36 inland communities. Migration is common in the study area (Grabowski et al., 2017). To account for sampling heterogeneity, the high- and low-prevalence communities were sub-stratified by gender, migration status, and three age bands ($15-24$, $25-34$, $35-49$), which resulted in 24 sampling groups and 576 flow variables. For each step of the sampling cascade, Monte Carlo draws from the posterior distribution of the conditional sampling distributions were obtained for each of the 24 population groups, multiplied, and used to adjust the flow estimates. Numerical inference of the joint posterior density (13) took 3.7 hours on a 2.4 Ghz processor (Supplementary Text S4). The flow estimates by the 24 sampling groups were then aggregated into 4 flow estimates between and within the RCCS fishing and inland communities. Figure 4B compares the marginal posterior distributions for each of the 4 transmission flows obtained with and without adjusting for sampling differences of the population. Participation and sequencing rates were overall higher in fishing communities compared to inland communities, which led to significantly different flow estimates after adjusting for sampling heterogeneity. The estimated flow ratio (inland$\rightarrow$fishing / fishing$\rightarrow$inland) was 2.58 (1.24 - 5.91) without sampling adjustments and 2.25 (1.04-5.23) with sampling adjustments. Therefore both analyses supported, contrary to what has commonly been assumed (Uganda AIDS Commission, 2014), the finding that the high-prevalence fishing communities were net sinks of local infection flows (Ratmann et al., 2020).

3.4. Transmission flows between age groups

We next turned to estimating transmission flows by age from the source-recipient data shown in Figure 5, our initial problem (2). Here, we divided the data by gender and 1-year age bands $(15, \ldots, 49)$, and accounted for sampling heterogeneity using the same

strata, which resulted in 70 sampling groups and 2450 flow variables. To obtain smooth estimates, we specified a GP prior on the log transmission intensities as in (15), and determined the prior distributions on the GP hyper-parameters from the data shown in Figure 2C. Specifically, the possible ranges of length-scale are $1-25$ (ρ_1), $1-15$ (ρ_2 when the source of infection is women) and $1-20$ (ρ_2 when the source of infection is men). Numerical inference of the joint posterior density (13) took 51 hours on a 2.4 Ghz processor (Supplementary Text S4).

Figure 5ACE shows the inferred age-distribution of male sources to female infections. For example, the red "wave" shows the mean posterior density estimate that women of age 16 are infected by men of particular ages (x-axis), i.e. $\mathbb{E}(\gamma_{ab}|\boldsymbol{n},\boldsymbol{s},\boldsymbol{X})$ for $b=(\mathrm{f},16)$ in our notation of (3a). Corresponding coefficients of variation are shown in Figure 5E . For young women aged <25 years, the mean posterior age-distribution of their sources plateaued at age 35. Our data thus contain evidence that some women aged <25 years were infected by men who are at least 10 years older, often referred to as sugar-daddies. However our non-parametric approach also indicates that transmission from sugar-daddies was uncommon. Considering women aged 20 years, an estimated XX% were infected from similarly aged men <24 years, XX% from slightly older men $24-29$ years, and XX% from considerably older men $30-49$ years. Further, as infected women age, our data contradict the idea that these women would then continue to infect their peers. Figure 5BDF shows the mean posterior age-distribution of female sources to male infections. Corresponding coefficients of variation are shown in Figure 5F. For men aged >35 years, the mean posterior age-distribution of their female source is not centred among their peers. In fact, it is very broad, and much broader than for men aged <25 years.

4. Discussion

In this study we introduce a hierarchical Bayesian Poisson model for estimating time homogeneous disease flows in human populations from pathogen deep-sequence data. The model (13) allows inference of transmission flows between arbitrarily defined strata of human populations, which enables addressing a range of epidemiological questions on

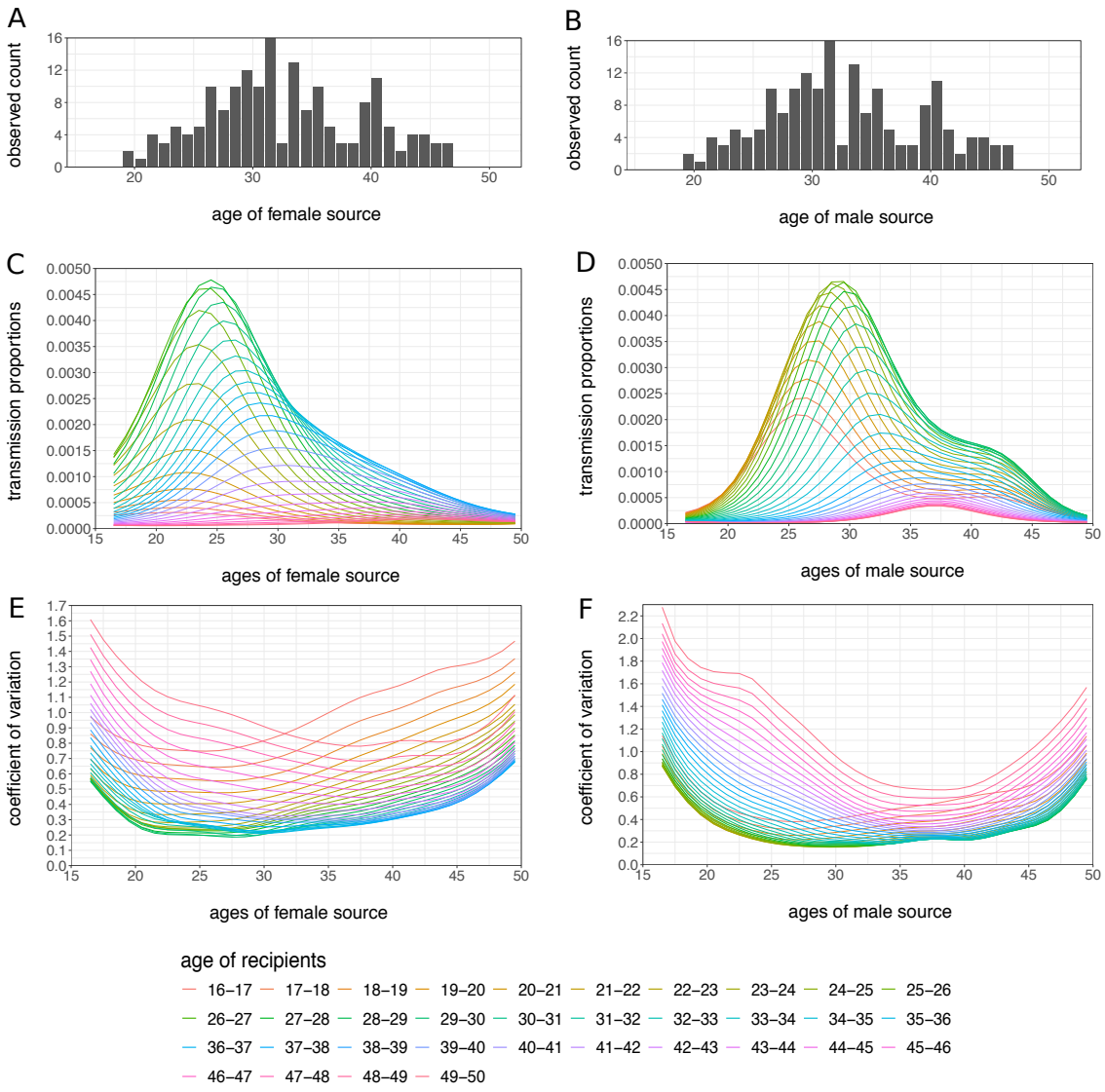

Fig. 5. transmission flows between age groups. Panel A/B shows the observed transmission counts from female/male one-year increment age groups. Gaussian process regression (16) was applied to estimate transmission flows between age groups. Panel C/D shows the proportion of transmissions from women/men in an age group (in horizontal axis), given the recipients' age groups (in color). Panel E/F plots the associated coefficients of variation of the estimated flows.

pathogen spread between geographic areas (CITE), by age categories (CITE), or other discretely-valued sociodemographic characteristics (CITE). The modelling framework is flexible and easy to extend. Here, we showed how Gaussian-process based regularisation techniques (CITE) can be incorporated to obtain smoothed flow estimates at high resolution. We also showed how data sampling bias can be adjusted for in this framework under the assumption that data are missing at random within strata (CITE). To fit the hierarchical Bayesian Poisson flow model, we provide an R interface to the NUTS Hamiltonian Monte Carlo algorithm of the Stan computing language (CITE). Means, medians, standard deviations and credibility intervals of the flow estimates and epidemiologically important, related quantities (3) are straightforward to obtain from numerical samples of the posterior distribution of the flow parameters.

Existing phylodynamic estimation approaches (CITE) are tailored for pathogen consensus sequences. The approach described here is tailored for pathogen deep-sequence data that captures in contrast to consensus sequences the genetic diversity of pathogens within infected hosts. The higher resolution of these data make possible to estimate the direction of pathogen spread between pairs of individuals (CITE). Although it is not possible to determine transmission direction with certainty between two individuals (CITE), deep-sequence data provide an alternative starting point for estimating population-level patterns of pathogen spread. The main advantages are first, that population-level spread can be directly estimated from individual source-recipient relationships and associated individual covariates. This allows modelling of and accounting for individual-level variables that shape disease transmission in remarkable detail, such as the sources of HIV infection by 1-year age bands. Second, little computational effort is needed to fit the Poisson flow model XX to deep-sequence data, because it falls within the class of Bayesian hierarchical models for binary data, for which efficient fitting procedures exist (CITE). This makes it computationally feasible to investigate complex models of disease spead such as the gender-and-age specific flow model XX that we considered here. It is also possible to automate numerical inference using for instance Stan (CITE) as we did here, which makes available well-tested and versatile numerical techniques for phylodynamic inference. In comparison, when using pathogen consensus sequences, existing inference

techniques typically require solving complex systems of ordinary differential equations (CITE BEAST PHYDYN), which limits modelling of disease transmission computationally to a relatively small number of covariates (CITE). Alternative phylogenetic clustering techniques are computationally faster, but it remains difficult to interpret the inferred associations between cluster membership and covariates associated with sampled pathogen consensus sequences (CITE Poon, Le Vu).

The method we propose has limitations. First, deep-sequencing protocols generate short sequence fragments, usually of 200 to 300 base pairs in length after trimming adaptors and low quality ends, and merging paired end fragments. This implies that pathogens need to evolve at a rate above $1/200$ mutations per site within the time scales of interest in human hosts, because otherwise reconstructed deep-sequence phylogenies do not contain the pattern of ancestral subgraphs that is characteristic of pathogen spread in one direction. Such high evolutionary rates are typical for viral pathogens that infect and evolve in humans over long periods of time, such as hepatitis C or HIV (CITE). We expect that the methods developed here will become applicable to a broad range of viral and bacterial infectious diseases as existing deep-sequencing methods that generate substantially longer pathogen sequence fragments become cheaper (CITE), or alternative approaches are being developed. Second, our method requires that deep-sequence data from a large population-based sample of infected individuals ($> 30\%$) are available. Such data are emerging, for example from the HPTN072 Popart HIV prevention trial (CITE), several HIV surveillance cohorts in Africa (CITE), the molecular epidemiologic HIV surveillance program in Germany (CITE), or TODO HPC EXAMPLE. Third, our inferences are based on source-recipient pairs with strong evidence for the direction of transmission, which is a subset of all the data available. We cannot exclude that this selection step introduces bias into the estimates obtained with the flow model XX, and recommend performing sensitivity analysis on the selection threshold in XX. Fourth, no continuous variables. Fifth, time-homogeneous.

Acknowledgements

References

Abeler-Dörner, L., Grabowski, M. K., Rambaut, A., Pillay, D., Fraser, C. et al. (2019) Pangea-hiv 2: Phylogenetics and networks for generalised epidemics in africa. *Current Opinion in HIV and AIDS*, **14**, 173–180.

Aral, S. O., Torrone, E. and Bernstein, K. (2015) Geographical targeting to improve progression through the sexually transmitted infection/hiv treatment continua in different populations. *Current Opinion in HIV and AIDS*, **10**, 477.

Berger, J. O., Bernardo, J. M. and Sun, D. (2015) Overall objective priors. *Bayesian Analysis*, **10**, 189–221.

Bonsall, D., Golubchik, T., de Cesare, M., Limbada, M., Kosloff, B., MacIntyre-Cockett, G., Hall, M., Wymant, C., Ansari, M. A., Abeler-Dörner, L. et al. (2018) A comprehensive genomics solution for hiv surveillance and clinical monitoring in a global health setting. *bioRxiv*, 397083.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of statistical software*, **76**.

Chang, L. W., Grabowski, M. K., Ssekubugu, R., Nalugoda, F., Kigozi, G., Nantume, B., Lessler, J., Moore, S. M., Quinn, T. C., Reynolds, S. J. et al. (2016) Heterogeneity of the HIV epidemic: an observational epidemiologic study of agrarian, trading, and fishing communities in Rakai, Uganda. *The lancet. HIV*, **3**, e388.

De Oliveira, T., Kharsany, A. B., Gräf, T., Cawood, C., Khanyile, D., Grobler, A., Puren, A., Madurai, S., Baxter, C., Karim, Q. A. et al. (2017) Transmission networks and risk of hiv infection in kwazulu-natal, south africa: a community-wide phylogenetic study. *The lancet HIV*, **4**, e41–e50.

Dellicour, S., Baele, G., Dudas, G., Faria, N. R., Pybus, O. G., Suchard, M. A., Rambaut, A. and Lemey, P. (2018) Phylodynamic assessment of intervention strategies for the west african ebola virus outbreak. *Nature communications*, **9**, 1–9.

Didelot, X., Fraser, C., Gardy, J. and Colijn, C. (2017) Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution*, **34**, 997–1007.

Dwyer-Lindgren, L., Cork, M. A., Sligar, A., Steuben, K. M., Wilson, K. F., Provost, N. R., Mayala, B. K., VanderHeide, J. D., Collison, M. L., Hall, J. B. et al. (2019) Mapping hiv prevalence in sub-saharan africa between 2000 and 2017. *Nature*, **570**, 189.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J. et al. (2014) The early spread and epidemic ignition of hiv-1 in human populations. *science*, **346**, 56–61.

Gall, A., Ferns, B., Morris, C., Watson, S., Cotten, M., Robinson, M., Berry, N., Pillay, D. and Kellam, P. (2012) Universal amplification, next-generation sequencing, and assembly of hiv-1 genomes. *Journal of clinical microbiology*, **50**, 3838–3844.

Givens, G. H., Smith, D. and Tweedie, R. (1997) Publication bias in meta-analysis: a bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*, 221–240.

Grabowski, M. K., Serwadda, D. M., Gray, R. H., Nakigozi, G., Kigozi, G., Kagaayi, J., Ssekubugu, R., Nalugoda, F., Lessler, J., Lutalo, T. et al. (2017) HIV prevention efforts and incidence of HIV in uganda. *New England Journal of Medicine*, **377**, 2154–2166.

Hall, M. D., Holden, M. T., Srisomang, P., Mahavanakul, W., Wuthiekanun, V., Limmathurotsakul, D., Fountain, K., Parkhill, J., Nickerson, E. K., Peacock, S. J. et al. (2019) Improved characterisation of mrsa transmission using within-host bacterial sequence diversity. *eLife*, **8**.

Hué, S., Pillay, D., Clewley, J. P. and Pybus, O. G. (2005) Genetic analysis reveals the complex structure of hiv-1 transmission within defined risk groups. *Proceedings of the National Academy of Sciences*, **102**, 4425–4429.

Kagaayi, J., Chang, L. W., Ssempijja, V., Grabowski, M. K., Ssekubugu, R., Nakigozi, G., Kigozi, G., Serwadda, D. M., Gray, R. H., Nalugoda, F. et al. (2019) Impact of combination hiv interventions on hiv incidence in hyperendemic fishing communities in uganda: a prospective cohort study. *The Lancet HIV*, **6**, e680–e687.

van de Kassteele, J., van Eijkeren, J., Wallinga, J. et al. (2017) Efficient estimation of age-specific social contact rates between men and women. *The Annals of Applied Statistics*, **11**, 320–339.

Le Vu, S., Ratmann, O., Delpech, V., Brown, A. E., Gill, O. N., Tostevin, A., Dunn, D., Fraser, C., Volz, E. M. and Database, U. H. D. R. (2019) Hiv-1 transmission patterns in men who have sex with men: Insights from genetic source attribution analysis. *AIDS research and human retroviruses*, **35**, 805–813.

Leitner, T. and Romero-Severson, E. (2018) Phylogenetic patterns recover known hiv epidemiological relationships and reveal common transmission of multiple variants. *Nature microbiology*, **3**, 983–988.

Lemey, P., Rambaut, A., Drummond, A. J. and Suchard, M. A. (2009) Bayesian phylogeography finds its roots. *PLoS computational biology*, **5**.

Müller, N. F., Rasmussen, D. and Stadler, T. (2018) Mascot: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, **34**, 3843–3848.

Müller, N. F., Rasmussen, D. A. and Stadler, T. (2017) The structured coalescent and its approximations. *Molecular biology and evolution*, **34**, 2970–2981.

Poon, A. F., Gustafson, R., Daly, P., Zerr, L., Demlow, S. E., Wong, J., Woods, C. K., Hogg, R. S., Krajden, M., Moore, D. et al. (2016) Near real-time monitoring of hiv transmission hotspots from routine hiv genotyping: an implementation case study. *The lancet HIV*, **3**, e231–e238.

Popinga, A., Vaughan, T., Stadler, T. and Drummond, A. J. (2015) Inferring epidemiological dynamics with bayesian coalescent inference: the merits of deterministic and stochastic models. *Genetics*, **199**, 595–607.

Rasmussen, C. E. (2003) Gaussian processes in machine learning. In *Summer School on Machine Learning*, 63–71. Springer.

Rasmussen, D. A., Wilkinson, E., Vandormael, A., Tanser, F., Pillay, D., Stadler, T. and De Oliveira, T. (2018) Tracking external introductions of hiv using phylodynamics reveals a major source of infections in rural kwazulu-natal, south africa. *Virus evolution*, **4**, vey037.

Ratmann, O., Grabowski, M. K., Hall, M., Golubchik, T., Wymant, C., Hoppe, A., Brown, A. L., de Oliveira, T., Gall, A., Kellam, P., Pillay, D., Quinn, T., Wawer, M., Laeyendecker, O., Serwadda, D., Gray, R. and Fraser, C. (2018) Inferring HIV-1 transmission networks and sources of ongoing viral spread in Africa with next-generation sequencing. In preparation.

Ratmann, O., Hodcroft, E. B., Pickles, M., Cori, A., Hall, M., Lycett, S., Colijn, C., Dearlove, B., Didelot, X., Frost, S. et al. (2017) Phylogenetic tools for generalized hiv-1 epidemics: findings from the pangea-hiv methods comparison. *Molecular biology and evolution*, **34**, 185–203.

Ratmann, O., Kagaayi, J., Hall, M., Golubchick, T., Kigozi, G., Xi, X., Wymant, C., Nakigozi, G., Abeler-Dörner, L., Bonsall, D. et al. (2020) Quantifying hiv transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in rakai, uganda. *The Lancet HIV*.

Ratmann, O., Van Sighem, A., Bezemer, D., Gavryushkina, A., Jurriaans, S., Wensing, A., De Wolf, F., Reiss, P., Fraser, C. et al. (2016) Sources of hiv infection among men having sex with men and implications for prevention. *Science translational medicine*, **8**, 320ra2–320ra2.

Riutort-Mayol, G., Vehtari, A., Bürkner, P.-C. and Riis, A. M. (2020) .

Romero-Severson, E. O., Bulla, I. and Leitner, T. (2016) Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, **113**, 2690–2695.

Rose, R., Hall, M., Redd, A. D., Lamers, S., Barbier, A. E., Porcella, S. F., Hudelson, S. E., Piwowar-Manning, E., McCauley, M., Gamble, T. et al. (2019) Phylogenetic methods inconsistently predict the direction of hiv transmission among heterosexual pairs in the hptn 052 cohort. *The Journal of infectious diseases*, **220**, 1406–1413.

Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D. et al. (2018) Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, **34**, 163–170.

Solin, A. and Särkkä, S. (2014) Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, 1–28.

Stadler, T. and Bonhoeffer, S. (2013) Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**, 20120198.

The Office of the US Global AIDS Coordinator (2014) Pepfar 3.0 controlling the epidemic: delivering on the promise of an aids-free generation. URL: `https://aidsfree.usaid.gov/resources/prevention-update/editions/december-2014/pepfar-30-controlling-epidemic-delivering`.

Todesco, E., Wirden, M., Calin, R., Simon, A., Sayon, S., Barin, F., Katlama, C., Calvez, V., Marcelin, A.-G. and Hué, S. (2019) Caution is needed in interpreting hiv transmission chains by ultradeep sequencing. *Aids*, **33**, 691–699.

Uganda AIDS Commission (2014) National strategic plan for hiv aids, 2014. URL: `http://www.nationalplanningcycles.org/sites/default/files/country_docs/Uganda/national_strategic_plan_for_hiv_aids_2011_2015.pdf`.

Uhlemann, A.-C., Dordel, J., Knox, J. R., Raven, K. E., Parkhill, J., Holden, M. T., Peacock, S. J. and Lowy, F. D. (2014) Molecular tracing of the emergence, diversification, and transmission of s. aureus sequence type 8 in a new york community. *Proceedings of the National Academy of Sciences*, **111**, 6738–6743.

561 UNAIDS (2014a) 90-90-90: an ambitious treatment target to help end the aids epidemic.
562 geneva: Unaids; 2014.

563 — (2014b) Fast-track: ending the aids epidemic by 2030. URL: `https://www.unaids.`
564 `org/en/resources/documents/2014/JC2686_WAD2014report`.

565 — (2018) Unaids data 2018. URL: `https://www.unaids.org/sites/default/files/`
566 `media_asset/unaids-data-2018_en.pdf`.

567 Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D. and Drummond, A. J. (2014)
568 Efficient bayesian inference under the structured coalescent. *Bioinformatics*, **30**, 2272–
569 2279.

570 Volz, E. M., Ionides, E., Romero-Severson, E. O., Brandt, M.-G., Mokotoff, E. and
571 Koopman, J. S. (2013) Hiv-1 transmission during early infection in men who have sex
572 with men: a phylodynamic analysis. *PLoS medicine*, **10**.

573 Volz, E. M., Pond, S. L. K., Ward, M. J., Brown, A. J. L. and Frost, S. D. (2009)
574 Phylodynamics of infectious disease epidemics. *Genetics*, **183**, 1421–1430.

575 Volz, E. M. and Siveroni, I. (2018) Bayesian phylodynamic inference with complex mod-
576 els. *PLoS computational biology*, **14**, e1006546.

577 Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., Gall, A.,
578 Cornelissen, M., Fraser, C., STOP-HCV Consortium, T. M. P. C. and Collaboration,
579 T. B. (2017) PHYLOSCANNER: inferring transmission from within-and between-host
580 pathogen genetic diversity. *Molecular biology and evolution*, **35**, 719–733.

581 Zhang, Y., Wymant, C., Laeyendecker, O., Grabowski, M. K., Hall, M., Hudelson, S.,
582 Piwowar-Manning, E., McCauley, M., Gamble, T., Hosseinipour, M. C. et al. (2020)
583 Evaluation of phylogenetic methods for inferring the direction of hiv transmission:
584 Hptn 052. *Clinical Infectious Diseases*.

Supplementary material

S1. Regression models for sampling

Transmission networks in our study are based on three survey rounds in Rakai Community Cohort Study between August 10, 2011 and January 30, 2015. Despite the enormous efforts devoted to enumerating all the household, unsampled individuals still remain in the survey. Specifically, potential sampling bias could be reduced by modelling approach. Informative priors of sampling proportions for each sampling category were built through Bayesian logistic regression, and prior samples were then propagated into the transmission flow estimation. This supplementary text explains how we modelled the sampling procedure.

All the reconstructed transmission events occurred between October 1, 2009 and January 30, 2015 where the median time of infected was estimated by [CITE]. Thus, we considered transmission events between eligible individuals in the observational period during the period between October 1, 2009 and January 30. However, 293 reconstructed pairs were between successfully sequenced participants who are not on the treatment. We start by providing an overview of the sampling cascade for eligible recipients.

The newly infected eligible individuals between October 1, 2009 and January 30, 2015 are recipients of interests. A fraction of eligible recipients participated. Each RCCS survey round is preceded by a population census, which enabled us to obtain posterior estimates of participation probabilities of the form (14) with a Bayesian logistic regression model implemented in Stan (Supplementary Text S1.2). Ideally, the probability of an eligible newly infected individuals participating is required. However, there is no knowledge of the infection status of individuals who were absent. Thus, we assumed participation was independent of being newly HIV infected for simplicity, and the probability of eligible individuals participating is an approximation to the targeted probability. Next, participating sink cases were sequencing-eligible only if they reported to be ART-naive at the first visit. Finally, a proportion of ART-naive newly infected participants were deep sequenced successfully. The final two steps were combined together, and the aim is to estimate the proportion of newly infected participants being sequenced

(Supplementary Text S1.3). Though the time of infection were estimated for sequenced individuals only instead of the newly infected participants, whether an infection is in the infection window could be predicted in Supplementary Text S1.1.

We then considered the source cases of new infections. Of all source cases infecting individuals in the cohort, approximately 30% were previously estimated to reside outside the surveyed area (Grabowski PLoS Med) and thus ineligible in RCCS. Of those in the surveyed area, we assumed that all sources were eligible to participate. Next, a fraction of eligible sources participated. Similarly, we assumed infectious individuals participated in the same manner as other eligible individuals, and estimated the posterior distribution of the probability of eligible individuals participating (Supplementary Text S1.2). Next, participating source cases were able to be sequenced if they were ART-naive at the first visit. Assume ART-naive patients are infectious individuals, because all the infected people on the treatment at the first visit were predicted to be infected before 2009.10, which indicates all the participating source cases were sequencing-eligible. Then, a fraction of ART naive individuals who participated were sequenced for virus samples satisfying minimum quality criteria. These prompt us to model the probability of possible infectious participants being sequenced (Supplementary Text S1.4).

In applying the sequence-success probabilities to the sampling cascade, we assumed that sequencing success was independent of individuals being a source or not. We further modelled the sampling probabilities of each transmission pair as the product of the sampling probabilities of the source and those of the recipient. Subject to the sampling probability of transmission events, a fraction of transmission events between eligible individuals in the observational window occurring during the infection window were observed.

S1.1. Estimation of infection in 2009.10-2015.1

To obtain whether an infected participant acquired infection between 2009.10 and 2015.1, sequenced individuals were used as the training set due to their known infection time from the estimation by XXX. In addition, the training set included individuals who were tested to be HIV-positive before 2009.10. 948 individuals who were infected before 2009.10 and 1691 newly infected individuals were used to predict newly infection status

of all other 2386 infected participants through Bayesian logistic regression,

$$\begin{aligned} N_i^n &\sim \text{Bernoulli}(1, \xi_i^n), \forall a \\ \text{logit}(\boldsymbol{\xi^n}) &= \boldsymbol{\beta X} \\ \beta_0 &\sim \mathcal{N}(0, 100) \\ \beta_j &\sim \mathcal{N}(0, 10), j \neq 0 \end{aligned} \tag{S1}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{X}$ refers to coefficients and design matrix, N_i^n is the indicator of the ith person were newly infected, and ξ_i^n is the probability of the ith person being newly infected. Specifically,

$$\begin{aligned} \boldsymbol{\beta} &= (\beta_0, \beta_m, \beta_i, \beta_a, \beta_f, \beta_d, \beta_{fp}, \beta_{art}) \in \mathbb{R}^8 \\ \boldsymbol{X}_{i\cdot} &= (1, x_{i,m}, x_{i,i}, x_{i,a}, x_{i,f}, x_{i,d}, x_{i,fp}, x_{i,art}) \in \mathbb{R}^8 \end{aligned}$$

where $x_{i,m}$ is the indicator of individual i being male, $x_{i,i}$ is the indicator of individual i being immigrant, $x_{i,f}$ is the indicator of individual i being in fishing sites, $x_{i,a}$ is the age of individual i, $x_{i,d}$ is the visit date of individual i in the study, $x_{i,fp}$ is the date of individual i first found to be HIV-positive and $x_{i,art}$ is the indicator of individual i being on ART. β_0 is the intercept, β_m, β_i, β_d, β_a, β_d, β_{fp} and β_{art} are regression coefficients for corresponding input variables. 10-fold cross validation were implemented to assess the model performance: accuracy of prediction on the test set is 96.96% (95.61%-98.47%), F1 score for the new infection class is 97.71% (96.59% - 98.90%) and 95.79% (93.82 % - 97.71%). Details of other competing models are in Table XXX.

S1.2. Probability of an eligible individual participating

To obtain the probability of an eligible recently infected (or infectious) individuals participating, the probability of an eligible individuals participating in the survey was used instead due to unknown infection status for absent individuals. However, the approximation assumes participation is independent of newly infection or infectious status. The following figures show the differences of participation rates between XXXXXX. Other competing models are in Table XXX.

INSERT A FIGURE SHOWING THE PARTICIPATION DIFF

Table S1. Model comparison

Quantities	Model	Accuracy	F1 score (new infection)	F1 score (old infection)
infection between 2009.10 and 2015.1	model 1 (no interaction)			
	model 2 (age × sex interaction)			
	model 3 (age × sex × immigrant interaction)			
participation fractions	model 1 (no interaction)			
	model 2 (age × sex interaction)			
	model 3 (age × sex interaction, dispersion)			
sequencing fractions (source)	model 1 (no interaction)			
	model 2 (age × sex interaction)			
sequencing fractions (recipient)	model 1 (no interaction)			
	model 2 (age × sex interaction)			

The number of individuals in a who were eligible N_a^e and participated N_a^p, and predictors $\boldsymbol{X}$ enables us to model the proportion of participation $\boldsymbol{\xi^p} = \{\xi_a^p\}_a$through a Bayesian logistic model. Specifically,

$$\begin{aligned} N_a^p &\sim \mathrm{BB}(N_a^n, \xi_a^p, \gamma), \forall a \\ \mathrm{logit}(\boldsymbol{\xi^p}) &= \boldsymbol{\beta X} \\ \beta_0 &\sim \mathcal{N}(0, 100) \\ \beta_j &\sim \mathcal{N}(0, 10), j \neq 0 \\ \gamma &\sim \exp(1) \end{aligned} \tag{S2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{X}$ refers to coefficients and design matrix, and and BB is beta-binomial distribution with dispersion parameter γ. Specifically,

$$\boldsymbol{\beta} = (\beta_0, \beta_f, \beta_{m1i}, \beta_{m2i}, \beta_{m3i}, \beta_{f1i}, \beta_{f2i}, \beta_{f3i}, \beta_{m1r}, \beta_{m2r}, \beta_{m3r}, \beta_{f1r}, \beta_{f2r}) \in \mathbb{R}^{11}$$

$$\boldsymbol{X}_{a\cdot} = (1, x_{a,f}, x_{a,m1i}, x_{a,m2i}, x_{a,m3i}, x_{a,f1i}, x_{a,f2i}, x_{a,f3i}, x_{a,m1r}, x_{a,m2r}, x_{a,m3r}, x_{a,f1r}, x_{a,f2r}) \in \mathbb{R}^{11}$$

where $x_{a,f}$ is the indicator of group a being in fishing sites, and $x_{a,xyz}$ is 1 if individuals in group a are gender x, in age group y and with migration status z where x takes m for male and f for male, y takes $1(15-24), 2(25-34), 3(35+)$, and z is i for immigrants and r for residents. β_0 is the intercept, $\beta_{a,f}$ and $\beta_{a,xyz}$ are regression coefficients for corresponding

input variables. The number of participants in 97.73% (90.91% - 100%) groups were correctly predicted in the 10-fold cross validation. Effective sample sizes were between 4192.857 and 20089.852, and Rhat were between 0.9999312 and 1.0007630. Figure XXXX shows the density plots of the posterior distribution of regression coefficients.

S1.3. Probability of a newly infected individual being sequenced

A proportion of newly infected participants were not on treatment and eligible to be sequenced, of whom a proportion were taken sequence samples to the acceptable standard. These two proportions could be combined into a single quantity, the probability of a newly infected participant being sequenced successfully $\boldsymbol{\xi^{sr}}$, which may vary between subpopulations. Figure XXXX shows INSERT A FIGURE SHOWING THE DIFF

The number of newly infected participants $\boldsymbol{N^{pr}}$ predicted in Supplementary Text S1.1, those who were sequenced $\boldsymbol{N^{sr}}$ and attributes of strata were used to estimate $\boldsymbol{\xi^{sr}}$. Specifically,

$$\begin{aligned} N_a^{sr} &\sim \text{Bin}(N_a^{pr}, \xi_a^{sr}), \forall a \\ \text{logit}(\boldsymbol{\xi^{sr}}) &= \boldsymbol{\beta X} \\ \beta_0 &\sim \mathcal{N}(0, 100) \\ \beta_j &\sim \mathcal{N}(0, 10), j \neq 0 \end{aligned} \tag{S3}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{X}$ refers to coefficients and design matrix. Specifically,

$$\begin{aligned} \boldsymbol{\beta} &= (\beta_0, \beta_f, \beta_i, \beta_m, \beta_1, \beta_2) \in \mathbb{R}^6 \\ \boldsymbol{X}_{a\cdot} &= (1, x_{a,f}, x_{a,i}, x_{a,m}, x_{a,1}, x_{a,2}) \in \mathbb{R}^6 \end{aligned}$$

where $x_{a,f}$, $x_{a,i}$, $x_{a,m}$, $x_{a,1}$, $x_{a,2}$ are taking value one if all the people in group a are in fishing sites, immigrants, males, aged 15-24, aged 25-34. β_0 is the intercept, $\beta_f, \beta_i, \beta_m, \beta_1, \beta_2$ are regression coefficients for the input variables. The number of sequenced in 95.24% (85.16% - 100%) groups were correctly predicted in the 10-fold cross validation. Effective sample sizes were between 11833.35 and 31646.49, and Rhat were between 0.9999662 and 1.0000076. Figure XXXX shows the density plots of the posterior distribution of regression coefficients.

S1.4. Probability of an infectious individual being sequenced

Assume all the infectious participants were not on treatment at their first visit and eligible to be sequenced, and a proportion of those were sequenced. Taking both into account, we modelled the probability of an ART-naive participant being sequenced successfully $\boldsymbol{\xi^{st}}$. Figure XXXX shows the varying $\boldsymbol{\xi^{st}}$ between strata. INSERT A FIGURE SHOWING THE DIFF

The number of ART-naive participants $\boldsymbol{N^{pt}}$, those who were sequenced $\boldsymbol{N^{st}}$ and attributes of strata were used to obtain the posterior distribution of $\boldsymbol{\xi^{st}}$. Specifically,

$$
\begin{aligned}
N_a^{st} &\sim \text{Bin}(N_a^{pt}, \xi_a^{st}), \forall a \\
\text{logit}(\boldsymbol{\xi^{st}}) &= \boldsymbol{\beta X} \\
\beta_0 &\sim \mathcal{N}(0, 100) \\
\beta_j &\sim \mathcal{N}(0, 10), j \neq 0
\end{aligned}
\tag{S4}
$$

where $\boldsymbol{\beta}$ and $\boldsymbol{X}$ refers to coefficients and design matrix. Specifically,

$$
\begin{aligned}
\boldsymbol{\beta} &= (\beta_0, \beta_f, \beta_i, \beta_m, \beta_1, \beta_2) \in \mathbb{R}^6 \\
\boldsymbol{X}_{a\cdot} &= (1, x_{a,f}, x_{a,i}, x_{a,m}, x_{a,1}, x_{a,2}) \in \mathbb{R}^6
\end{aligned}
$$

where $x_{a,f}$, $x_{a,i}$, $x_{a,m}$, $x_{a,1}$, $x_{a,2}$ are taking value one if all the people in group a are in fishing sites, immigrants, males, aged 15-24, aged 25-34. β_0 is the intercept, $\beta_f, \beta_i, \beta_m, \beta_1, \beta_2$ are regression coefficients for the input variables. The number of sequenced in 100% (95.05% - 100%) groups were correctly predicted in the 10-fold cross validation. Effective sample sizes were between 13289.51 and 32010.51, and Rhat were between 0.9999662 and 1.0000500. Figure XXXX shows the density plots of the posterior distribution of regression coefficients.

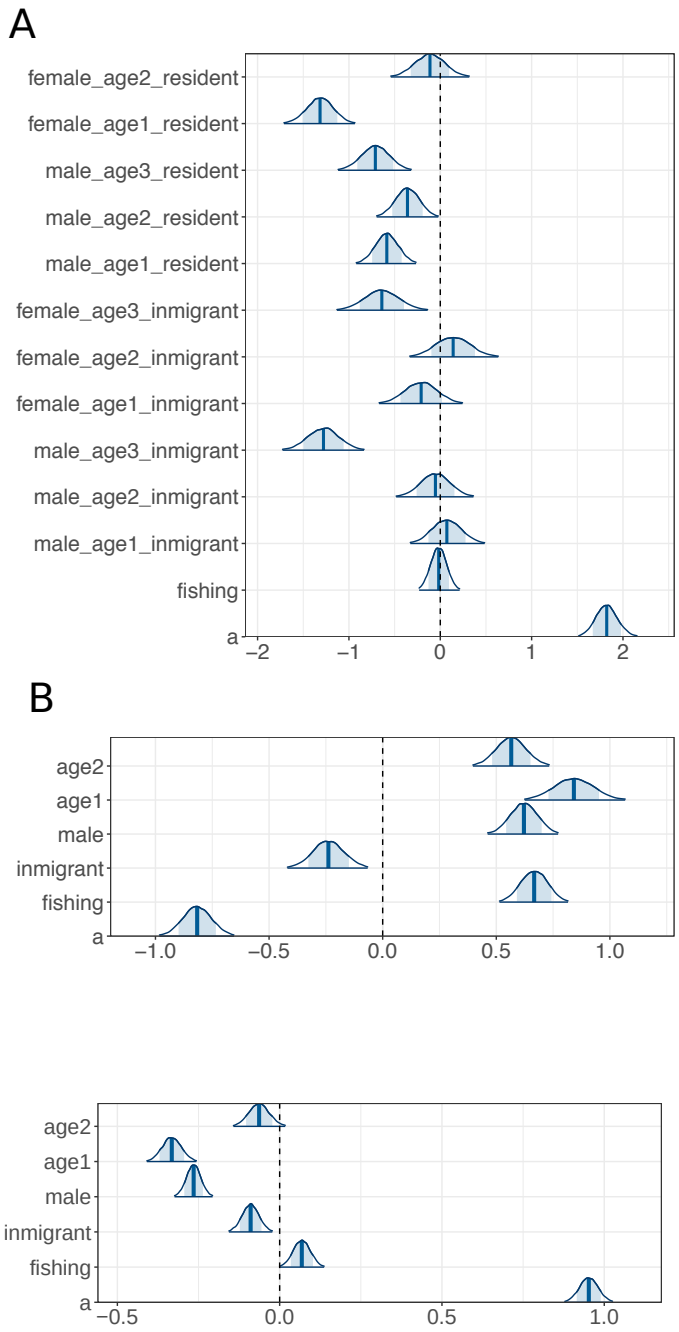

Fig. S1. Densities of the posterior distribution of regression coefficients. Panel A, B, C show the density plots of posterior distributions of parameters in (S2) and (??), (??)

S2. Algorithm

S2.1. Markov Chain Monte Carlo algorithm

To estimate the posterior density (13) through Markov Chain Monte Carlo (MCMC), we considered the full conditional distributions

$$p(\xi_a|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \mathbf{s}, \mathbf{n}, \mathbf{X}) \text{ for all } a, \tag{S5}$$

$$p(\lambda_{ab}|\boldsymbol{\xi}, \boldsymbol{\lambda}_{-ab}, \mathbf{s}, \mathbf{n}, \mathbf{X}), \tag{S6}$$

where ξ_a is the sampling probability of the ath population group, λ_{ab} denotes the average transmission counts from population group a to population group b, $\boldsymbol{\xi}_{-a} = \boldsymbol{\xi} \setminus \xi_a$ and $\boldsymbol{\lambda}_{-ab} = \boldsymbol{\lambda} \setminus \lambda_{ab}$. The density (S5) is not available in closed form, and we performed Metropolis-within-Gibbs updates for each ξ_a. The density (S6) is of Gamma form, and so a Gibbs step can be used to update $\boldsymbol{\pi}$. The resulting MCMC algorithm iterates through Metropolis-within-Gibbs updates for each ξ_a, followed by a series of Gibbs updates for each λ_{ab}; see Supplementary Text S2.1.1 for full details.

This Supplementary Text describes implementation details of the Metropolis-within-Gibbs algorithm to estimate the posterior distribution (13),

$$p(\boldsymbol{\lambda}, \boldsymbol{\xi}|\mathbf{n}, \mathbf{s}, \mathbf{X}) \propto \Pi_{a,b} p(n_{ab}|\lambda_{ab}, \xi_a \xi_b) \Pi_a p(\xi_a|\mathbf{X}_a, \mathbf{s}_a) \Pi_{a,b} p(\lambda_{ab}|\boldsymbol{\xi}).$$

with unknown parameters

- $\boldsymbol{\xi}$ vector of sampling proportions for each population group a, $a = 1, \ldots, A$
- $\boldsymbol{\lambda}$ vector of average actual transmission counts from group a to group b, with length L corresponding to the number of observed counts among all possible pairwise combinations of counts a, b between population strata
- $\boldsymbol{\pi}$ vector of transmission flows from group a to group b, of length L.

The algorithm is implemented in the `phyloflows` R package, `https://github.com/BDI-pathogens/phyloscanner/tree/master/phyloflows`. Calculations in this paper were performed with version 1.2.0.

S2.1.1. Overall structure of algorithm

Owing to the large number of model parameters in the presence of deep sampling heterogeneity, the algorithm exploits the factorisation of the posterior density (13) into the full conditionals (S5-S6) to update parameters in blocks.

S2.1.2. Gibbs step

The full conditional distribution $p(\lambda_{ab}|\boldsymbol{\xi}, \boldsymbol{\lambda}_{-ab}, \mathbf{s}, \mathbf{n}, \mathbf{X})$ is

$$\begin{aligned} p(\lambda_{ab}|\boldsymbol{\xi}, \boldsymbol{\lambda}_{-ab}, \mathbf{s}, \mathbf{n}, \mathbf{X}) &\propto p(\lambda_{ab}) p(n_{ab}|\lambda_{ab}, \xi_a, \xi_b) \\ &\propto \exp(-\lambda_{cd}\xi_c\xi_d)(\lambda_{cd}\xi_c\xi_d)^{n_{cd}} \lambda_{cd}^{\alpha_{cd}-1} \exp(-\beta\lambda_{cd}) \\ &\propto \exp(-\lambda_{cd}(\xi_c\xi_d + \beta))\lambda_{cd}^{n_{cd}+\alpha_{cd}-1} \\ &\sim \text{Gamma}(n_{cd} + \alpha_{cd}, \xi_c\xi_d + \beta) \end{aligned} \tag{S7}$$

so that each λ_{ab} can be updated in a single Gibbs step by sampling from the right hand side.

S2.1.3. Metropolis-Hastings within Gibbs steps

Next, the algorithm updates in turn the sampling proportions of the source populations. The full conditional distribution of ξ_a follows from (13),

$$\begin{aligned} &p(\xi_a|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \mathbf{n}, \mathbf{s}, \mathbf{X}) \\ &\quad\propto \Pi_{c=a \text{or} d=a} p(n_{cd}|\lambda_{cd}, \xi_c\xi_d) p(\xi_a|\boldsymbol{X}) \\ &\quad\propto \Pi_{c=a \text{or} d=a} \text{Poi}(n_{cd}; \lambda_{cd}\xi_c\xi_d) \text{Gamma}(\lambda_{cd}; \xi_c\xi_d) p(\xi_a|\boldsymbol{X}). \end{aligned} \tag{S8}$$

The full conditional (S8) does not have closed form and so the algorithm performs a Metropolis-Hastings update. We sought to avoid tuning parameters as much as possible, and for this reason adopted the independence proposal density

$$q(\xi_a'|\xi_a) = q(\xi_a') \tag{S9}$$

where ξ_a' is proposed from the sampling prior $p(\xi_a')$ and

$$\boldsymbol{\xi}' = \begin{cases} \xi_c & c \neq a \\ \xi_c' & c = a \end{cases}$$

The resulting Metropolis Hasting ratio is

$$
\begin{aligned}
r &= \frac{p(\xi_a'|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \mathbf{n}, \mathbf{s}, \mathbf{X})}{p(\xi_a|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \mathbf{n}, \mathbf{s}, \mathbf{X})} \\
&\quad \times \frac{q(\xi_a)}{q(\xi_a')} \\
&= \frac{\Pi_b p(n_{ab}|\lambda_{ab}, \xi_a'\xi_b)\Pi_{b\neq a} p(n_{ba}|\lambda_{ba}, \xi_b\xi_a')\Pi_b p(\lambda_{ab}|\boldsymbol{\xi}')\Pi_{b\neq a} p(\lambda_{ba}|\boldsymbol{\xi}')p(\xi_a')}{\Pi_b p(n_{ab}|\lambda_{ab}, \xi_a\xi_b)\Pi_{b\neq a} p(n_{ba}|\lambda_{ba}, \xi_b\xi_a)\Pi_b p(\lambda_{ab}|\boldsymbol{\xi})\Pi_{b\neq a} p(\lambda_{ba}|\boldsymbol{\xi})p(\xi_a)} \\
&\quad \times \frac{p(\xi_a)}{p(\xi_a')} \\
&= \frac{\Pi_b p(n_{ab}|\lambda_{ab}, \xi_a'\xi_b)\Pi_{b\neq a} p(n_{ba}|\lambda_{ba}, \xi_b\xi_a')\Pi_b p(\lambda_{ab}|\boldsymbol{\xi}')\Pi_{b\neq a} p(\lambda_{ba}|\boldsymbol{\xi}')}{\Pi_b p(n_{ab}|\lambda_{ab}, \xi_a\xi_b)\Pi_{b\neq a} p(n_{ba}|\lambda_{ba}, \xi_b\xi_a)\Pi_b p(\lambda_{ab}|\boldsymbol{\xi})\Pi_{b\neq a} p(\lambda_{ba}|\boldsymbol{\xi})}
\end{aligned}
\tag{S10}
$$

Proposed moves are accepted with probability $\min(1, r)$.

The sampling fraction in subpopulation a, ξ_a could be separated into that for transmitters and recipients, ξ_a^t and ξ_a^r if different sampling cascades were applied to source and recipients. Gibbs steps remain same and updates for ξ_a will be divided into updating ξ_a^t and ξ_a^r sequentially.

S2.1.4. Discussion

(a) The observed phylogenetic data do not provide information on the sampling probabilities and consequently the posterior distributions of the group sampling probabilities will be close to the corresponding prior distributions. For this reason we do not think that a random walk proposal would confer advantages over the independence proposal (S9).

S2.2. Hamiltonian Monte Carlo

S2.2.1. Independent prior

The posterior distribution (13) under the prior (15) could also be alternatively inferred through Hamiltonian Monte Carlo in Stan. However, it is impossible to incorporate prior samples in Stan. Instead, we fitted beta distributions to sampling probabilities, and provided parameters of beta distributions.

```
data {
  int<lower=1> N;  \\ number of observations
  int Y[N];  \\ response variable
  // sampling fractions
```

```
  int<lower=1> N_xi; \\ number of sampling categories
  matrix[N_xi,2] shape; \\ parameters of sampling rates
  int xi_id_src[N]; \\ sampling category indices for source
  int xi_id_rec[N]; \\ sampling category indices for recipients
  real alpha;  \\ the parameter of gamma prior
}
parameters {
  vector<lower=0>[N] lambda; \\ poisson rate
  vector<lower=0,upper=1>[N_xi] xi;   \\ sampling fractions
}
transformed parameters {
  vector[N] lxi_pair; \\ log sampling fractions
  vector[N] betav; \\ the parameter of gamma prior
  real beta;
  for (i in 1:N){
   lxi_pair[i] = log(xi[xi_id_src[i]]) + log(xi[xi_id_rec[i]]);
   if (Y[i]==0){
     betav[i] = 1/(xi[xi_id_src[i]]*xi[xi_id_rec[i]])-1;}
   else{
     betav[i] = Y[i]/(xi[xi_id_src[i]]*xi[xi_id_rec[i]]);}
  }
  beta = sum(betav);
}
model {
  for (i in 1:N_xi){
    target += beta_lpdf(xi[i]|shape[i,1],shape[i,2]);
  }
  for (i in 1:N){
    target += gamma_lpdf(lambda[i]|alpha,beta);
  }
  // likelihood including all constants
  target += poisson_log_lpmf(Y | lxi_pair + log(lambda));
}
```

S2.2.2. Gaussian process prior

The aim to estimate the flow vector under the prior (16),

$$p(\boldsymbol{\lambda}, \boldsymbol{\xi}, \mathbf{f}, \mu, \alpha, \boldsymbol{\rho} | \mathbf{n}, \mathbf{s}, \mathbf{X}) \tag{S11}$$

with unknown parameters

$\boldsymbol{\xi}$ vector of sampling proportions for each population group a, $a = 1, \ldots, A$

$\boldsymbol{\lambda}$ vector of average actual transmission counts from group a to group b, with length L corresponding to the number of observed counts among all possible pairwise combinations of counts a, b between population strata

$\boldsymbol{\pi}$ vector of transmission flows from group a to group b, of length L.

$\mathbf{f}$ vector of Gaussian process realisations over the deviation of log average transmission counts from the baseline, of length L.

μ a scalar representing the baseline of log average transmission counts for all groups.

α a scalar representing the marginal standard deviation of Gaussian process.

$\boldsymbol{\rho}$ a vector of length-scales of Gaussian process.

θ set of Gaussian process hyperparameters, $\theta = \{\alpha, \boldsymbol{\rho}\}$

The Gaussian process regression is computationally expensive to implement in Stan for a considerable number of data points while hyperparameters were estimated at the same time. Therefore, we used basis function approximation approach proposed by Solin and Särkkä (2014) and implemented by Riutort-Mayol et al. (2020).

Specifically, we considered the Gaussian process regression under squared exponential kernel in Equation (16). Bochner's theorem demonstrates a stationary covariance function is a Fourier transform of a positive finite measure which is called spectral density if the density exist. In addition, a covariance operator defined for each stationary covariance function is translation invariant, and thus expressed as a series of Laplace operators, deduced from the polynomial expansion of spectral density. Laplace operators could represented as an eigenfunction expension with Dirichlet boundary conditions. These leads to the theorem that a stationary covariance function can be represented as

$$k(\boldsymbol{x}, \boldsymbol{x}') \approx \sum_{j=1}^{\infty} S_\theta \left(\sqrt{\boldsymbol{\lambda}_j} \right) \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{x}'), \tag{S12}$$

where S_θ is the spectral density of the stationary covariance function k, θ is the set of hyperparameters of k, $\{\boldsymbol{\lambda}_j\}_{j=1}^{\infty}$ and $\{\phi_j(\boldsymbol{x})\}_{j=1}^{\infty}$ are respectively the jth eigenvalues and eigenfunctions of the Laplacian operator (Solin and Särkkä, 2014).

The spectral density S_θ for a squared exponential kernel k with hyperparameters $\theta = (\alpha, \boldsymbol{\rho})$ where α represents marginal standard deviation and $\boldsymbol{\rho}$ denotes length-scales

has the form

$$S_\theta(\boldsymbol{\omega}) = \alpha^2\sqrt{2\pi}^2 \prod_{i=1}^{2} \rho_i \exp\left(-\frac{1}{2}\sum_{i=1}^{2}\rho_i^2\omega_i^2\right), \tag{S13}$$

By contrast, the form of the jth eigenvalues and eigenfunctions of the Laplace operator is independent of covariance functions,

$$\begin{aligned} \boldsymbol{\lambda}_j &= \left\{\lambda_{\boldsymbol{I}_{jd}}\right\}_{d=1}^{2} = \left\{\left(\tfrac{\pi \boldsymbol{I}_{jd}}{2L_d}\right)^2\right\}_{d=1}^{2} \\ \phi_j(\boldsymbol{x}) &= \prod_{d=1}^{2}\phi_{\boldsymbol{I}_{jd}}(x_d) = \prod_{d=1}^{2}\sqrt{\frac{1}{L_d}}\sin\left(\sqrt{\lambda_{\boldsymbol{I}_{jd}}}(x_d + L_d)\right) \end{aligned} \tag{S14}$$

where $[-L_d, L_d]$ is the domain for the dth dimension and $\boldsymbol{I}$ denotes the matrix containing indices.

$$\boldsymbol{I} = \begin{pmatrix} 1 & 1 & \cdots & 2 & 2 & \cdots & 3 & 3 & \cdots & \cdots \\ 1 & 2 & \cdots & 1 & 2 & \cdots & 1 & 2 & \cdots & \cdots \end{pmatrix}^T$$

The fact that eigenvalues of the Laplace operator rise with j and spectral densities decrease faster to zero with j for a bounded covariance function suggests Equation (S12) could be truncated to the first m terms,

$$k(\boldsymbol{x}, \boldsymbol{x}') \approx \sum_{j=1}^{m} S_\theta\left(\sqrt{\boldsymbol{\lambda}_j}\right)\phi_j(\boldsymbol{x})\phi_j(\boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T\Delta\boldsymbol{\phi}(\boldsymbol{x}'), \tag{S15}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_j(\boldsymbol{x}))_{j=1}^m \in \mathbb{R}^m$ is the column vector of eigenfunctions and $\Delta \in \mathbb{R}^{m\times m}$ is the diagonal matrix whose (j,j)th entry is $S_\theta(\sqrt{\boldsymbol{\lambda}_j})$.

Assume m_1 and m_2 are the number of basis functions for the first and second dimensions respectively, the number of eigenvalues and eigenfunctions is $m = m_1 \times m_2$. We define a matrix $\boldsymbol{I} \in \mathbb{R}^{m\times 2}$ containing indices of univariate basis functions. For example, if $m_1 = 3$ and $m_2 = 4$, then

$$\boldsymbol{I} = \begin{pmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 \end{pmatrix}^T$$

That is, the jth eigenvalues $\boldsymbol{\lambda}_j$ is a vector of length 2 whose elements are the $\boldsymbol{I}_{j1}$th and $\boldsymbol{I}_{j2}$th univariate eigenvalues. The eigenfunction $\{\phi_j(\mathbf{x})\}_j$ is the product of the $\boldsymbol{I}_{j1}$th and $\boldsymbol{I}_{j2}$th univariate eigenfunctions.

Once approximating the kernel function, Gaussian process could be approximated by m basis functions of Laplacian (Solin and Särkkä, 2014)

$$f(\boldsymbol{x}) \approx \sum_{j=1}^{m} (S_\theta(\sqrt{\boldsymbol{\lambda}_j}))^{\frac{1}{2}} \phi_j(\boldsymbol{x}) \beta_j = \boldsymbol{\phi}(\boldsymbol{x})^T \Delta^{\frac{1}{2}} \boldsymbol{\beta} \tag{S16}$$

where $\beta_j \sim \mathcal{N}(0,1)$ is the jth element of a column vector $\boldsymbol{\beta}$. Therefore,

$$f(\boldsymbol{X}) \approx \boldsymbol{\Phi} \Delta^{\frac{1}{2}} \boldsymbol{\beta} = \boldsymbol{\Phi}(\Delta^{\frac{1}{2}} \boldsymbol{\beta}) \tag{S17}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(\boldsymbol{x_1}) & \cdots & \phi_m(\boldsymbol{x_1}) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x_n}) & \cdots & \phi_m(\boldsymbol{x_n}) \end{pmatrix} \text{ and } \boldsymbol{\Delta} = \begin{pmatrix} S_\theta(\sqrt{\boldsymbol{\lambda}_1}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S_\theta(\sqrt{\boldsymbol{\lambda}_m}) \end{pmatrix}$$

We provided an example code to demonstrate the basis function approximation. Consider seven age groups called "15-19","20-24","25-29","30-34","35-39","40-44","45-49". Assume $\alpha = 2.5$, $\boldsymbol{\rho} = (12, 9)$, $\mu = -1$ and sampling probabilities are $0.35, 0.45, 0.5, 0.55, 0.5, 0.55, 0.4$ for these age groups. We simulated transmission counts between age groups through (11) and (16), and then re-estimate model parameters. Figure S2AB shows the posterior distributions of α, $\boldsymbol{\rho}$, μ and Poisson rates (black) and assumed parameter values (red).

```
\\ eigenvalues
functions {
  vector lambda_nD(real[] L, int[] m, int D) {
    vector[D] lam;
    for(i in 1:D){
      lam[i] = ((m[i]*pi())/(2*L[i]))^2; }
    return lam;
  }
\\ spectral functions
  real spd_nD(real alpha, row_vector rho, vector w, int D) {
    real S;
    S = alpha^2 * sqrt(2*pi())^D * prod(rho) * exp(-0.5*((rho .* rho) * (w .* w)));
    return S;
  }
\\ eigenfunctions
  vector phi_nD(real[] L, int[] m, matrix x) {
    int c = cols(x);
    int r = rows(x);
    matrix[r,c] fi;
    vector[r] fi1;
    for (i in 1:c){
      fi[,i] = 1/sqrt(L[i])*sin(m[i]*pi()*(x[,i]+L[i])/(2*L[i]));
```

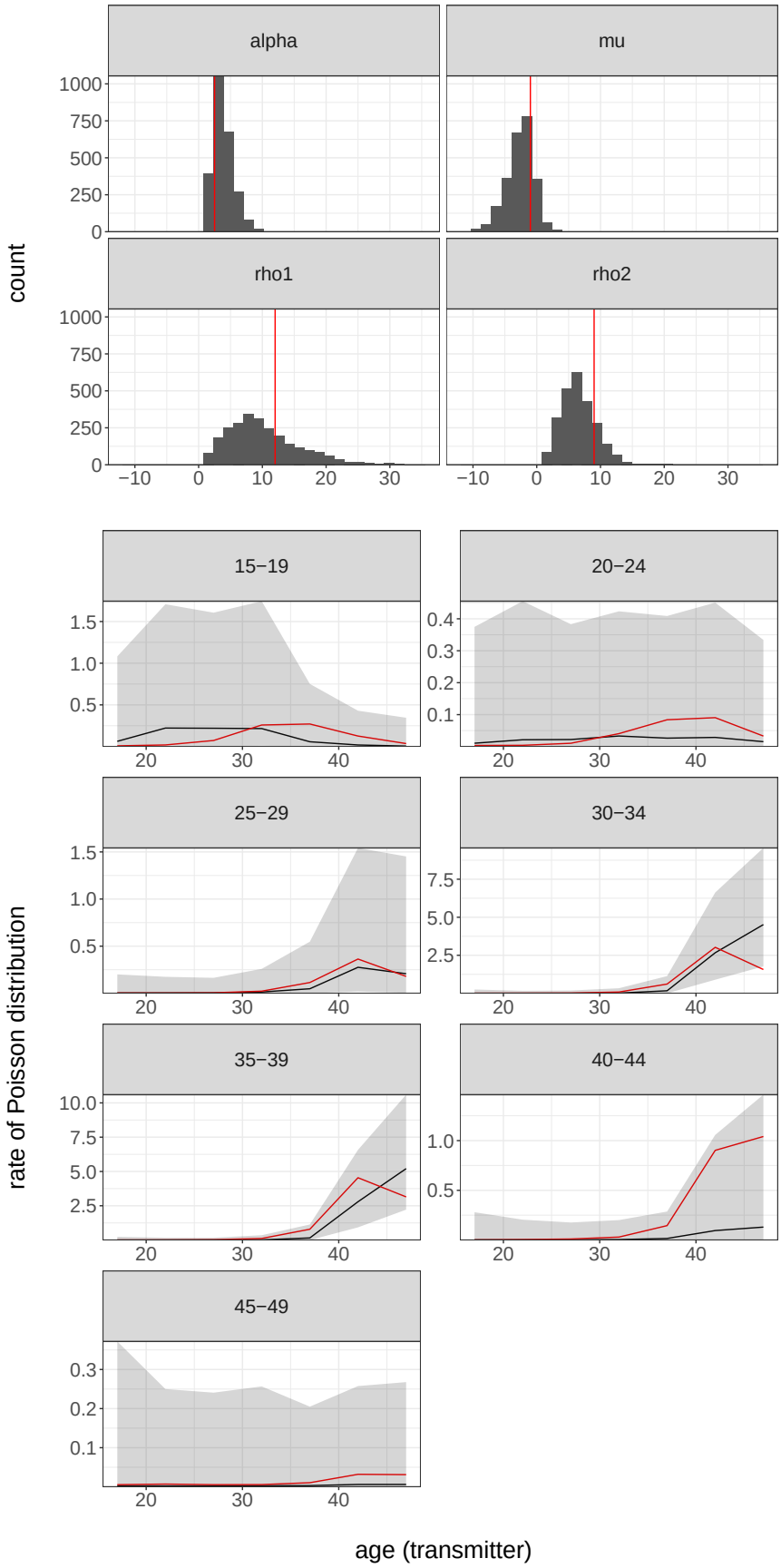

Fig. S2. Demonstration of Gaussian process regression. Panel A compares posterior distributions of model parameters and true values. We simulated transmission count data on the assumption of Gaussian process prior (16) with fixed hyperparameters (vertical red lines). Hyperparameters were re-estimated and posterior distributions are plotted in dark grey area. Panel B compares estimated (black lines and grey shaded areas) and simulated poisson rates (red lines).

```
    }
    fi1 = fi[,1];
    for (i in 2:c){
      fi1 = fi1 .* fi[,i];
    }
    return fi1;
  }
}
data {
  int<lower=1> N_xi; \\ number of sampling categories
  int<lower=1> N; \\ number of inputs
  int<lower=1> D; \\ dimension of inputs
  real L[D]; \\ domain
  int<lower=1> M; \\ number of basis functions in 1D
  int<lower=1> M_nD; \\ number of basis functions
  int indices[M_nD,D]; \\ index matrix S
  int xi_id[N,2]; \\ sampling category indices
  matrix[N,D] x; \\ inputs
  int<lower=0> y[N]; \\ outputs
  matrix[N_xi,2] shape; \\ parameters of sampling rates
}
transformed data {
  matrix[N,M_nD] PHI; \\ eigenfunctions
  for (m in 1:M_nD){
    PHI[,m] = phi_nD(L, indices[m,], x);
  }
}
parameters {
  row_vector<lower=0>[D] rho; \\ lengthscale
  real<lower=0> alpha; \\ marginal standard deviation
  real mu; \\ baseline
  vector[M_nD] beta;
  vector<lower=0,upper=1>[N_xi] xi; \\ sampling proportions
}
transformed parameters {
  vector[N] f;  \\ gaussian process
  vector[N] xi1; \\ sampling proportions for transmitters
  vector[N] xi2; \\ sampling proportions for recipients
  vector[M_nD] diagSPD; \\ square root of eigenvalues
  vector[M_nD] SPD_beta; \\ spectral density evaluations
  for(m in 1:M_nD){
    diagSPD[m] =  sqrt(spd_nD(alpha, rho, sqrt(lambda_nD(L, indices[m,], D)), D));
  }
  SPD_beta = diagSPD .* beta;
  f= PHI * SPD_beta; \\ compute (S16)
  for (n in 1:N){
    xi1[n] = xi[xi_id[n,1]]; \\ take sampling proportions for transmitters
  }
  for (n in 1:N){
    xi2[n] = xi[xi_id[n,2]]; \\ take sampling proportions for recipients
  }
}
model {
  for (i in 1:N_xi){
    xi[i] ~ beta(shape[i,1],shape[i,2]); \\ sampling proportions priors
  }
  beta ~ normal(0,1); \\ normal random variable
```

```
  rho ~ inv_gamma(2.22121,7.04478); \\ lengthscale
  alpha ~ normal(0, 10); \\ marginal standard deviation
  mu ~ normal(0, 10); \\ baseline
  y ~ poisson_log(f + mu + log(xi1) + log(xi2)); \\ poisson model
}
```

S2.3. Maximum likelihood estimates of the model

Recall the likelihood of Bayesian flow model is

$$\begin{aligned} l &= \Pi_{ab}\text{Poisson}(n_{ab}; \lambda_{ab}\xi_a^T\xi_b^R)\Pi_a p(N_a^{Te}, N_a^{Ts}|\xi_a^T)\Pi_a p(N_a^{Re}, N_a^{Rs}|\xi_a^R) \\ &= \Pi_{ab}\text{Poisson}(n_{ab}; \pi_{ab}\eta\xi_a^T\xi_b^R)\Pi_a p(N_a^{Te}, N_a^{Ts}|\xi_a^T)\Pi_a p(N_a^{Re}, N_a^{Rs}|\xi_a^R) \end{aligned} \tag{S18}$$

Taking the derivative of log-likelihood to zero gives maximum likelihood estimates.

$$\begin{aligned} \pi_{ab}\eta &= \frac{n_{ab}}{\xi_a^T\xi_b^R} \\ \xi_a^T &= \frac{N_a^{Ts}}{N_a^{Te}} \\ \xi_a^R &= \frac{N_a^{Rs}}{N_a^{Re}} \end{aligned} \tag{S19}$$

As $\sum_{ab}\pi_{ab} = 1$,

$$\begin{aligned} \eta &= \sum_{ab}\frac{n_{ab}}{\xi_a^T\xi_b^R} \\ \pi_{ab} &= \frac{n_{ab}}{\xi_a^T\xi_b^R} / \sum_{ab}\frac{n_{ab}}{\xi_a^T\xi_b^R} \end{aligned} \tag{S20}$$

S3. Simulation experiments

S3.1. Overall simulation strategy

Simulations were implemented to validate our bias-adjusted method and investigate errors brought by sampling bias in a classical SIR-type epidemics and Rakai Community Cohort Study. Our strategy is to simulate transmission counts, re-estimate transmission flows and compare the estimated flows with the true flows. This Supplementary Text provides full details on the simulation experiments.

S3.2. 2×2 simulations

The first experiment was a minimal example to assess sampling differences. We considered transmission flows between two population groups $A = (a, b)$, which for ease of illustration we refer to as individuals living in rural areas or small communities (group a), and individuals living in large communities (group b). Assume 60% individuals are in group b and both groups have the same male and female populations. We simulated stochastic epidemics between four compartments specified by genders and community types through structured SIR model (4). ODE parameters were set to make prevalence approximately 60% and run till equilibrium. Assume people in this epidemics were sampled at the end of the simulation period. Assume the sampling intensity in a was 60%, and the sampling fraction in b ranged from 60% to 35%, in order to assess the impact of sampling differences of 0%, 5%, 10% (baseline), 15%, 20%, 25%, from which the number of sampled individuals could be calculated. We considered transmission events in the past ten years as they could be reconstructed from sampled individuals. Simulated epidemics gave the true values for the 4-dimensional flow vector, $\boldsymbol{\pi}^0 = (\pi_{aa}, \pi_{ab}, \pi_{ba}, \pi_{bb})$, and observed transmission counts between and within a and b, $\mathbf{n} = (n_{aa}, n_{ab}, n_{ba}, n_{bb})$. Population sizes were calibrated to ensure roughly 300 transmission events observed. The vector $\boldsymbol{\pi}$ was inferred when the prior distribution is constructed by Equation (**??**) (adjustment for sampling differences) or Beta$(25 + 0.5, 50 - 25 + 0.5)$ (no adjustment for sampling differences). These prior distributions, together with the assumed true sampling fractions, were displayed in Figure S3.

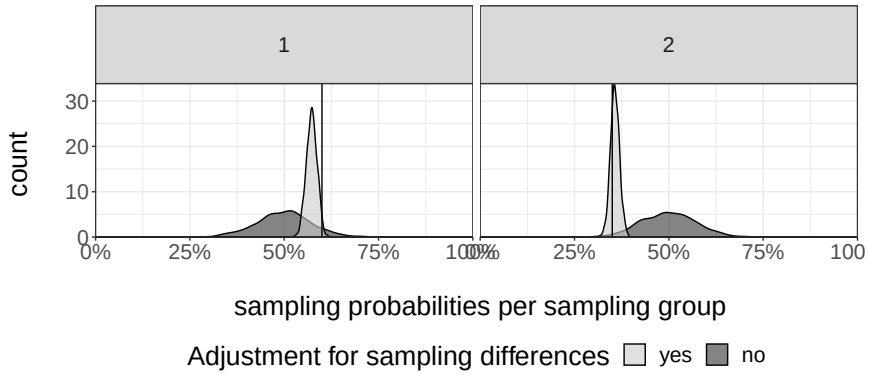

Fig. S3. Prior distributions for the 2×2 simulation. We re-estimated transmission flows with and without adjustment for sampling differences in the 2×2 simulation, where the adjustment is through the specification of prior distributions. Light grey areas represent prior distributions of ξ_a (left panel) and ξ_b (right panel) without adjustment for sampling differences, i.e. Beta (25.5, 25.5). Dark grey areas refer to prior distributions of ξ_a (left) and ξ_b (right) with adjustment for sampling differences, i.e. Equation (**??**) where X_a^p and X_a^s are informed by sampling data. The vertical line shows the true sampling fractions, $\xi_a = 0.6$ and $\xi_b = 0.35$.

S3.3. Sensitivity to the sample size

The population size was changed to make the number of transmission events to be around 100 and 600 for the baseline scenario in order to assess the robustness of the method to studies with varying transmission events observed. Figure **??** shows the median WCE when sampling differences are ignored (light grey bars) compared to when they are not ignored (dark grey bars) in three studies where 100, 300, 600 transmission events were observed.

S3.4. Rakai-type simulations

The second experiment mimicked the substantial sampling heterogeneity observed in the Rakai case study in simulations. We considered the flow vector of length 576 between 24 subpopulations defined by three age brackets, genders, immigration status and community types. We used data on sampling procedure, $N_a^e, N_a^p, N_a^n, N_a^s$, in the Rakai case study and took the estimated flow vector in Section 3.3 as the true flow $\boldsymbol{\pi_0}$. We simulated the observed transmission counts between population groups following the inference procedure. For a fixed sample size $N = 300$, the total number of transmissions Z were simulated from

$$Z \sim \text{Poisson}(N/\overline{\xi})$$

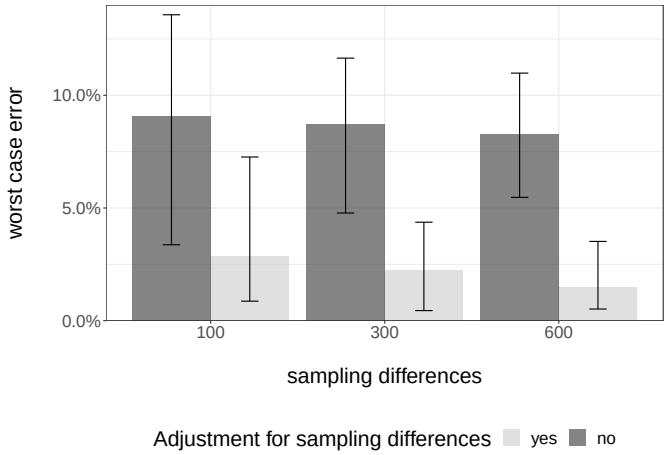

Fig. S4. Sensitivity of bias-adjusted method to varying sample size. The Bayesian flow model (13) was used to re-estimate transmission flows within and between two population groups in simulations, in which the observed flow counts were subject to varying (100, 300, 600) number of observed transmission events in x-axis. In the first set of inference runs (light grey), group-differences in sampling were not adjusted for, by using the same sampling prior density for both groups, $p(\xi_a) = Beta(25.5, 25.5)$ and $p(\xi_b) = Beta(25.5, 25.5)$. In the second set of inferences (dark grey), group-differences in sampling were adjusted based on counts of sampled and total infected individuals in both populations, see Equation $(\mathbf{??})$. The worst case error between the true transmission flows and median posterior estimates was calculated on 100 simulations in each scenario, and is shown on the y-axis. With increasing differences in sample sizes, we found increasing bias in flow estimates when sampling differences were not adjusted for, and decreasing bias in flow estimates when sampling differences were adjusted for. Despite capability to reduce bias through bias-adjusted method in all scenarios, the algorithm performs better for larger sample sizes.

where $\overline{\xi}$ was the average sampling probability in the population. The actual transmission flows between population groups followed from

$$\mathbf{z} \sim \text{Multinomial}(Z, \boldsymbol{\pi_0}).$$

The observed transmission counts $\mathbf{n}$ were then simulated under

$$n_{ab} \sim \text{Binomial}(z_{ab}, \xi_a \xi_b)$$

for any population group a and b. The quantity we are interested in is the flow vector between and within inland and fishing communities. The vector $\boldsymbol{\pi}$ was inferred when the prior distribution is constructed by Equation (**??**) (adjustment for sampling differences) or Beta$(25{+}0.5, 50{-}25{+}0.5)$ (no adjustment for sampling differences), and aggregated to the target flow vector. The difference between sampling intensities in inland and fishing communities is about 10%. We adjusted sampling proportions to make the sampling difference range from 5% to 25% while keeping sampling structure unchanged by using sine-type function. The observed transmission counts $\mathbf{n}$ were then simulated for each case, and used to make inference on the flow vector.

The second simulation experiment closely mimicked sampling heterogeneity as seen in the data from the RCCS observational study (Figure 3B). For a 10% sampling difference between high- and low-prevalence communities, the median WCE was reduced from $6.7\%(2.8\%{-}9.8\%)$ to $2.4\%(1.1\%{-}4.5\%)$ by considering sampling bias when the sampling difference between inland and fishing communities was 10%.

S4. Numerical performance

This supplementary text presents the numerical performance of the bias-adjusted method.

S4.1. Transmission flows between areas with high and low disease prevalence

The algorithm takes 2.8 hours. The effective sample sizes range from 49910.32 to 105115.13. The acceptance rates from 0.84 to 0.97 for Metropolis-within-Gibbs steps. Trace plots of three parameters with the the worst performance in terms of effective sample size are in Figure S6A.

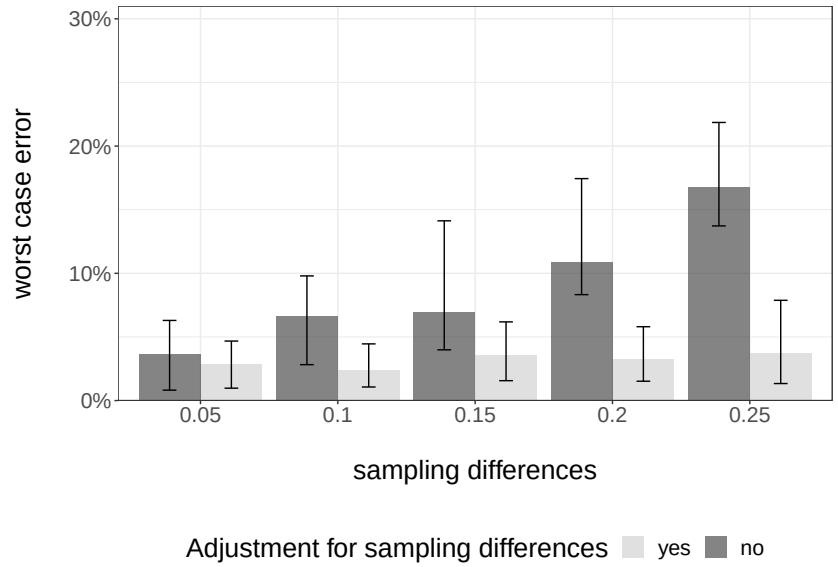

Fig. S5. Comparing errors in estimates of transmission flow obtained with and without adjustment for sampling differences. We assessed bias in estimating transmission flows with model (13) before and after accounting for sampling differences. Transmission flows were simulated between population groups in Rakai analysis. The sampling difference between inland and fishing sites is 10% in Rakai analysis, and was adjusted to range from 5% to 25% in x-axis. The worst case error between the transmission flows in RCCS and median posterior estimates was calculated on 20 simulations in each sampling scenario, and is shown on the y-axis. With increasing differences in sampling differences, we found increasing bias in flow estimates when sampling differences were not adjusted for. When sampling differences were adjusted for, the median worst case error in flow estimates remained on average below 5%.

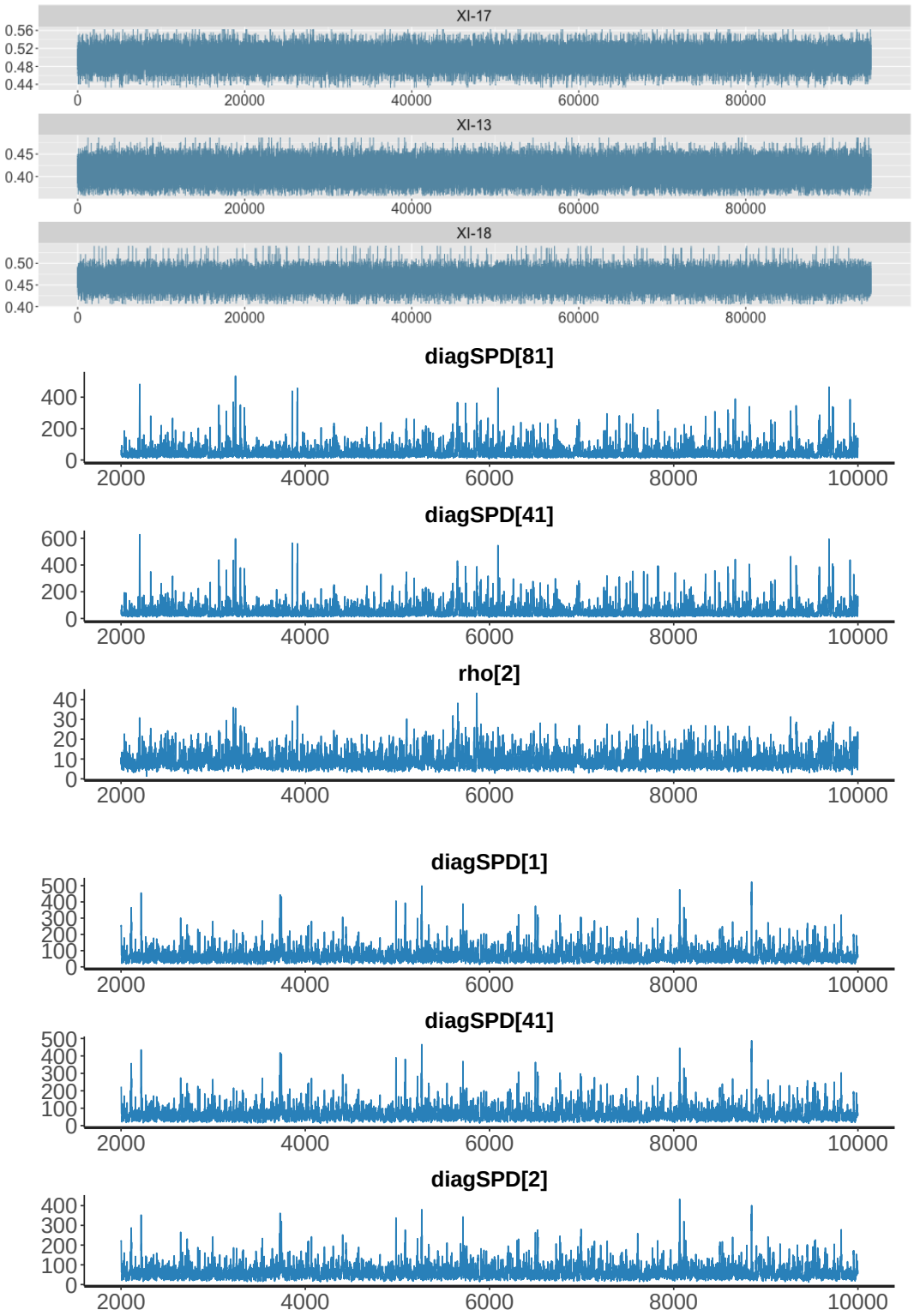

Fig. S6. numerical performance in RCCS. Panel A/B/C provides trace plots of three parameters with lowest effective sample sizes when analysing transmission flows between areas with high and low disease prevalence, flows between age groups from females to males and flows between age groups from males to females respectively.

S4.2. Transmission flows between age groups

The algorithm takes 25.3 hours for transmission pairs from women to men. The effective sample sizes range from 1620.61 to 28485.75. Trace plots of three parameters with the the worst performance in terms of effective sample size are in Figure S6B.

The algorithm takes 41.4 hours for transmission pairs from men to women. The effective sample sizes range from 1486.12 to 20623.32. Trace plots of three parameters with the the worst performance in terms of effective sample size are in Figure S6C.