Received: 5 January 2021 | Accepted: 10 January 2022

DOI: 10.1111/rssc.12544

ORIGINAL ARTICLE

Journal of the Royal Statistical Society Series C Applied Statistics

Inferring the sources of HIV infection in Africa from deep-sequence data with semi-parametric Bayesian Poisson flow models

Xiaoyue Xi[1] | Simon E. F. Spencer[2] | Matthew Hall[3] | M. Kate Grabowski[4,5] | Joseph Kagaayi[5,6] | Oliver Ratmann[1] | on behalf of Rakai Health Sciences Program and PANGEA-HIV

[1]Department of Mathematics, Imperial College London, London, UK

[2]Department of Statistics, University of Warwick, Coventry, UK

[3]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[4]Department of Pathology, Johns Hopkins University, Baltimore, MD, USA

[5]Rakai Health Sciences Program, Kalisizo, Uganda

[6]Makerere University School of Public Health, Kampala, Uganda

Correspondence
Oliver Ratmann, Department of Mathematics, Imperial College London, London SW72AZ, United Kingdom.
Email: oliver.ratmann@imperial.ac.uk

Abstract

Pathogen deep-sequencing is an increasingly routinely used technology in infectious disease surveillance. We present a semi-parametric Bayesian Poisson model to exploit these emerging data for inferring infectious disease transmission flows and the sources of infection at the population level. The framework is computationally scalable in high-dimensional flow spaces thanks to Hilbert Space Gaussian process approximations, allows for sampling bias adjustments, and estimation of gender- and age-specific transmission flows at finer resolution than previously possible. We apply the approach to densely sampled, population-based HIV deep-sequence data from Rakai, Uganda, and find substantive evidence that adolescent and young women were predominantly infected through age-disparate relationships in the study period 2009–2015.

KEYWORDS

flow models, Gaussian process, infectious disease epidemiology, origin-destination models, phylodynamics, Stan

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

1 | INTRODUCTION

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

HIV remains one of the largest public health threats, especially in sub-Saharan Africa where approximately 61% of all new cases worldwide occur (UNAIDS, 2019). In recent years, rates of incident cases have overall dropped considerably with the widespread adoption of prevention interventions 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 (Cohen et al., 2011; Grabowski et al., 2017; Hayes et al., 2019), although they remain well above UNAIDS thresholds for elimination (UNAIDS, 2018).

Within Africa, 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 with the goal of targeted control interventions to these groups (Abeler-Dörner et al., 2019). 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 disproportionately pass on infection, sinks are groups that disproportionately acquire infection, and hubs are both sources and sinks. The population groups can be defined in various ways.

For example 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 (i.e. hotspots) that are surrounded by neighbouring areas with substantially lower prevalence. Although often assumed, it is unclear if hotspots are also sources of epidemic spread to neighbouring lower-prevalence communities (Ratmann et al., 2020). Here, study populations are divided into individuals living in high-prevalence areas (h) and low-prevalence areas (l), and then transmission flows are estimated within and between them,

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

where π_{ab} is the proportion of transmission flows from group a to group b subject to $\sum_{ab} \pi_{ab} = 1$.

Another prominent application 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. Here, study populations are divided into sex-specific age groups (we consider 1-year age groups between 15 to 49 years), and then transmission flows between and within age groups are estimated,

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

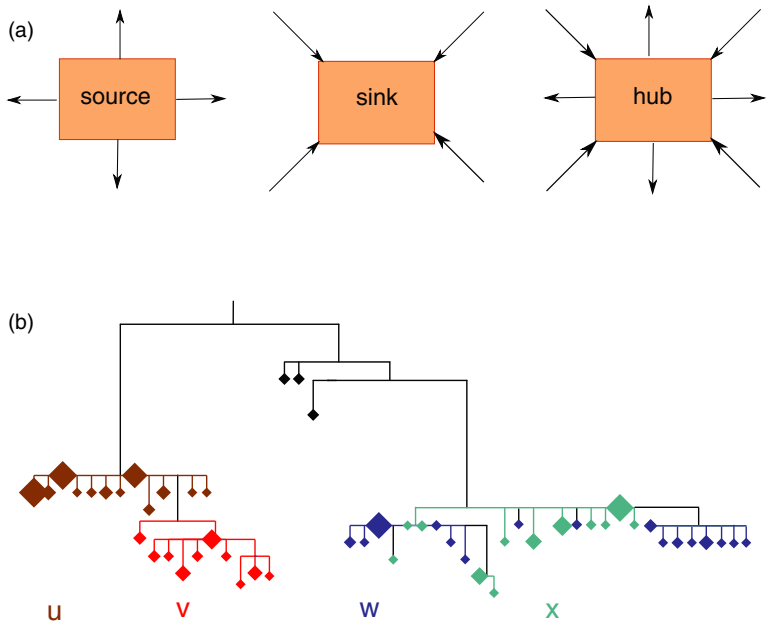


FIGURE 1 Analysis aims and sketch of deep-sequence phylogenetic data to address these aims. (a) The overall aim of phylogenetic source attribution analyses is to infer how pathogens are passed on between population groups. Conceptually populations can be divided into source populations that predominantly 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 [Colour figure can be viewed at wileyonlinelibrary.com]

where π_{ab}^{mf} is the proportion of transmissions from men in age band a to women in age band b, and similarly for π_{ab}^{fm}. We consider here only male-female transmission flows because we found no evidence of male-male transmission in previous analyses (Ratmann et al., 2019) and sexual transmission between women is extremely rare. The flow matrix Equation (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 20 years; the vector of recipients of infection from group a individuals ($\boldsymbol{\omega}^a$), for example from men aged 25 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

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

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

$$\gamma_{ab} = \pi_{ab}/\pi_{ba}. \tag{3c}$$

Flow matrices of the form (1–2) are also of central interest to characterise human migration flows between countries (Raymer et al., 2013), transport flows between locations (Tebaldi & West, 1998), bacterial migration in humans (Ailloud et al., 2019), or human contact intensities (van de Kassteele et al., 2017), and are alternatively referred to as origin-destination matrices (Hazelton, 2001; Miller et al., 2019).

1.2 | Inference from pathogen sequence data

Transmission flow matrices have been estimated from contact tracing or survey data on partner characteristics, though these data are often subject to reporting and/or social desirability biases, especially for sexual diseases that are associated with stigma or remain criminalised in many countries (Barré-Sinoussi et al., 2018). Here, we are concerned in estimating the quantities (1–3) from pathogen sequences, which are considered an objective marker of disease flow. For fast-evolving pathogens like HIV, mutations accrue quickly and the phylogenetic relationship of pathogen sequences can be used to evaluate many aspects of transmission dynamics such as the origins of HIV (Faria et al., 2014), the contribution of different disease phases to onward spread (Ratmann et al., 2016; Volz et al., 2013), or outbreak detection (Poon et al., 2016).

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, however, the data are insufficient to estimate the direction of transmission between any two sampled individuals (Leitner & Romero-Severson, 2018). For this reason most methods infer transmission flows 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 between groups 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 sampling correlated latent phylogenies and state histories is often 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 ordinary differential equations (ODE) models of disease spread. The flow parameters are obtained as by-products of the estimated latent states and parameters of the compartmental model, and in general vary in time over the phylogenetic

history. Adopting the marginalisation approach, 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 (Vaughan et al., 2014; Volz et al., 2013).

An emerging strategy for estimating transmission flows involves phylogenetic analysis of multiple distinct pathogen sequences per infected host, because such data make possible to attribute sets of viral lineages to individuals and infer the ancestral relationships between them, which can provide direct evidence into the direction of transmission between two individuals (Leitner & Romero-Severson, 2018). Such analyses are becoming broadly applicable, because deep sequencing technology now allows generating thousands to millions of distinct pathogen sequence fragments per sample (Gall et al., 2012; Zhang et al., 2020). Prior work focused on software development (Skums et al., 2018; Wymant et al., 2017), validation of the bioinformatics protocol for inferring the direction of transmission (Ratmann et al., 2019; Zhang et al., 2020), and reconstruction of partially observed transmission networks at the population level (Ratmann et al., 2019).

1.3 | Semi-parametric Poisson flow models

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] giving a score that transmission occurred from the first to the second individual, possibly via unsampled intermediate individuals (phylogenetic direction scores). This is advantageous, because first, in this canonical form the data enable us to present the estimation problem in terms of a class of Bayesian Poisson models for non-Gaussian flow data that can flexibly describe a range of epidemiological questions including transmission in space (1), or by age and sex (2), similar to the Poisson models used for estimating transport or migration flows (Raymer et al., 2013; Tebaldi & West, 1998). Second, the models can account for multi-level sampling heterogeneity, which is typically present in population-based disease occurrence data but was not emphasised in phylogenetic analysis. 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. Specifically, we propose using recently developed Hilbert Space Gaussian Process (HGSP) approximations to ensure the regularisation priors remain computationally tractable (Solin & Särkkä, 2020). This brings our approach into the form of semi-parametric Bayesian Poisson models, which enable inference of high-resolution transmission flows similar to the Integrated Nested Laplace Approximations used for inferring high-resolution human contact matrices (van de Kassteele et al., 2017).

In Section 2, we introduce our notation and develop the semi-parametric Bayesian Poisson flow model. In Sections 3.1 and 3.2, we assess the performance of the Poisson model in estimating transmission flows from sampling-biased data, and identify suitable HGSP approximations. Sections 3.3–3.6 illustrate our approach on HIV deep sequence data from the Rakai Community Cohort Study (RCCS) of the Rakai Health Sciences programme, situated in south-eastern Uganda (Grabowski et al., 2017). 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 estimation of age- and gender-specific transmission flows at finer detail than previously possible while remaining computationally scalable. Particular attention is given to potential sampling biases, and we propose a hierarchical model of the sequence sampling cascade for the analysis of transmission flows. Section 4 closes with a discussion.

2 | METHODOLOGY

2.1 | Notation and definitions

In this section, we present the notation that is used to estimate transmission flows in a population $\mathcal{P}$ of size N, during a study period $\mathcal{T} = [t_1, t_2]$. We define by $i = 1, \ldots, N$ the identifier of infected individuals in $\mathcal{P}$ during $\mathcal{T}$. The transmission events during the study period can thus be described in a $N \times N$ binary matrix $\mathbf{Z}$, where $z_{ij} = 1$ denotes transmission from person i to person j, and $z_{ij} = 0$ denotes no transmission. The transmission matrix is not symmetric, and diagonal entries are zero.

We estimate transmission flows between population strata, and denote the strata by a and the set of strata by $\mathcal{A}$, which is of dimension $A > 0$. The number of transmission events in $\mathcal{T}$ from group a to group b are $z_{ab} = \sum_{i \in a, j \in b} z_{ij}$, and the primary object of interest is the $A \times A$ flow matrix $\boldsymbol{\pi}$ with entries $\pi_{ab} = z_{ab}/z^+$ where $z^+ = \sum_{ab} z_{ab}$. The flow matrix is in general not symmetric, and is subject to $\sum_{a,b \in \mathcal{A}} \pi_{ab} = 1$. The matrix may contain structural zeros, for example in the case of HIV female-to-female transmission is extremely unlikely . 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 & May, 1992). 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 study period, and for ease of notation drop the dependence of our data and estimates on $\mathcal{T} = [t_1, t_2]$.

Pathogen deep sequence data are available from sampled, infected individuals. We denote the sampling status vector for all individuals in $\mathcal{P}$ by $\mathbf{s} = (s_i)_{i=1,\dots,N}$, where $s_i = 1$ denotes that person i is sampled, and $s_i = 0$ that person i is not sampled. The number of sampled individuals is N^s, which corresponds here to the 2652 individuals for whom a viral deep sequence is available for analysis. We will characterise population sampling in terms of individual-level characteristics, such as age or location of residence, that are described with p covariates, which we denote with the $N \times p$ matrix $\mathbf{X}$. The output of the phylogenetic deep sequence analysis can be summarised in an $N \times N$ 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 not symmetric, diagonal entries are zero, and entries involving unsampled individuals are missing. To estimate the flow matrix $\boldsymbol{\pi}$, consider the observed flow counts

$$n_{ab} = \sum_{i \in a, j \in b} \mathbb{1}\{s_i = 1\}\mathbb{1}\{s_j = 1\}\hat{z}_{ij} = \sum_{i \in a, j \in b} \mathbb{1}\{s_i = 1\}\mathbb{1}\{s_j = 1\}\mathbb{1}\{w_{ij} > \zeta\}, \tag{7}$$

where $\zeta \in (0, 1)$ is a threshold that can be used to select phylogenetically highly supported source-recipient pairs, and $\hat{z}_{ij}$ is taken as a perfect predictor of z_{ij} among sampled individuals. The counts can be arranged into the $A \times A$ count matrix $\mathbf{n}$, and sum to $n^+ = \sum_{a,b} n_{ab}$. In previous studies, ζ was set to 0.5 or 0.6, and n^+ was between 100 to 500 (Hall et al., 2019; Ratmann et al., 2020).

The naïve flow estimator is defined by $\hat{\pi}_{ab} = \frac{n_{ab}}{\sum_{c,d} n_{cd}}$. 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 $\mathbb{E}(\hat{\pi}_{ab}) = (z_{ab}\xi_a\xi_b)/\left(\sum_{c,d} z_{cd}\xi_c\xi_d\right)$. Consequently, the naïve flow estimator is only unbiased when the population groups were homogeneously sampled, that is, ξ_a is the same for all a, which is rarely the case (Ratmann et al., 2020).

2.2 | Inferring flows from heterogeneously sampled data

The data inputs for estimating transmission flows from pathogen deep-sequence data Equation (7) are of the same form as for estimating origin-destination matrices with unobserved transport routes (Hazelton, 2001), migration flows (Raymer et al., 2013), or contact intensities (van de Kassteele et al., 2017), prompting us to formulate the statistical model in general terms. Considering the actual, unobserved number of flow events (i.e. transmissions) between all population groups, $\mathbf{z} = (z_{ab})_{a,b \in A}$, the complete data likelihood that arises under mathematical models of the form Equation (4) in a fixed study period is the multinomial

$$p(\mathbf{z}|z^+, \boldsymbol{\pi}) \propto \prod_{a,b} (\pi_{ab})^{z_{ab}}, \tag{8}$$

where for ease of notation we denote by $\boldsymbol{\pi}$ the vector of non-zero elements of the flow matrices (1–2), and similarly for $\mathbf{n}$ and $\mathbf{z}$. Model Equation (8) ignores potential second-order correlations between flows, for example that a female infected by an older male may be more likely to transmit

to men of older age. Since the total number of transmissions is in itself a random variable, we consider the related Poisson model

$$p(\mathbf{z}|\lambda) \propto \prod_{a,b} (\lambda_{ab})^{z_{ab}} \exp(-\lambda_{ab}) = \left[\prod_{a,b} (\pi_{ab})^{z_{ab}}\right] \left[\eta^{z^+} \exp(-\eta)\right], \tag{9}$$

where λ_{ab} can be interpreted as the flow 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 flows $\mathbf{z}$ are not observed. We assume that individuals are sampled at random within strata (SARWS). SAWRS implies that sampling is independent of being a source 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 flow counts z_{ab} can be conveniently integrated out, yielding for the observed flow counts the Poisson model

$$p(\mathbf{n}|\lambda, \boldsymbol{\xi}) = \prod_{a,b} (\lambda_{ab} \xi_a \xi_b)^{n_{ab}} \exp(-\lambda_{ab} \xi_a \xi_b). \tag{10}$$

An important point in this construction is that we are free to choose the stratification $\mathcal{A}$ in order to accommodate the SARWS assumption. We will show that flow estimates on different, coarser population stratifications that are of primary interest are easily obtained through the aggregation property of the Poisson system Equation (9).

The sampling-adjusted maximum-likelihood estimates of λ_{ab} and π_{ab} under Equation (10) 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

$$\hat{\pi}_{ab} = \frac{n_{ab}}{\hat{\xi}_a \hat{\xi}_b} \Bigg/ \left[\sum_{c,d} \frac{n_{cd}}{\hat{\xi}_c \hat{\xi}_d}\right], \tag{11}$$

where $\hat{\xi}_a = N_a^s / N_a$ (Supplementary Material, section S1).

2.3 | Inferring high-resolution flows with regularising priors

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 $\mathbf{s}$. We assume that flows are independent of sampling, allowing us to decompose the joint posterior distribution into

$$\begin{aligned} p(\lambda, \boldsymbol{\xi}|\mathbf{n}, \mathbf{s}, \mathbf{X}) &\propto p(\mathbf{n}|\lambda, \boldsymbol{\xi}, \mathbf{s}, \mathbf{X}) p(\lambda|\boldsymbol{\xi}, \mathbf{s}, \mathbf{X}) p(\boldsymbol{\xi}|\mathbf{s}, \mathbf{X}) \\ &= p(\mathbf{n}|\lambda, \boldsymbol{\xi}) p(\lambda|\boldsymbol{\xi}) p(\boldsymbol{\xi}|\mathbf{s}, \mathbf{X}). \end{aligned} \tag{12}$$

A possible limitation of Equation (11) 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, for example, a logistic regression approach, we obtain

$$p(\xi_a|\mathbf{s},\mathbf{X}) = \int \text{logit}^{-1}(\mathbf{x_a}\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathbf{s},\mathbf{X})d\boldsymbol{\beta}, \tag{13}$$

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

Considering the prior density on the transmission intensities λ, in some applications the population strata are unordered such as in application Equation (1). In this case, we propose using

$$\lambda_{ab}|\xi_a,\xi_b \sim \text{Gamma}(\alpha_{ab},\beta), \quad \alpha_{ab} = 0.8/L, \quad \beta = 0.8/Z^p(\xi_a,\xi_b), \tag{14}$$

where L is the length of the flow vector $\boldsymbol{\pi}$, which is equivalent to the number of structurally non-zero entries in the flow matrices (1–2), 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 Equation (14) induces on $\boldsymbol{\pi}$ an objective Dirichlet prior density with parameters $\alpha_{ab} = 0.8/L$ (Berger et al., 2015), such that the likelihood Equation (10) dominates the prior Equation (14) regardless of the number of flows to estimate. However when the population groups can be ordered, such as the K 1-year age bands in Equation (2), the structure of the flow model (12) enables using regularising prior densities that penalise against large deviations in flow intensities between similar source and recipient populations. For Equation (2), we opted for (stacked) two-dimensional Gaussian-process priors on the entries of λ,

$$\begin{aligned}
&\log\lambda = \mu\mathbf{1} + \nu\mathbb{1}_{mf} + \boldsymbol{f},\\
&\boldsymbol{f} = (\boldsymbol{f}_{mf}^T, \boldsymbol{f}_{fm}^T)^T,\\
&\boldsymbol{f}_{mf} \sim \mathcal{GP}(0, k_{mf}), \quad \boldsymbol{f}_{fm} \sim \mathcal{GP}(0, k_{fm}),\\
&k_{mf}((a_1,b_1),(a_2,b_2)) = \sigma_{mf}^2 \exp\left(-\left[\frac{(a_2-a_1)^2}{2\ell_{mf,a}^2} + \frac{(b_2-b_1)^2}{2\ell_{mf,b}^2}\right]\right)\\
&k_{fm}((a_1,b_1),(a_2,b_2)) = \sigma_{fm}^2 \exp\left(-\left[\frac{(a_2-a_1)^2}{2\ell_{fm,a}^2} + \frac{(b_2-b_1)^2}{2\ell_{fm,b}^2}\right]\right)\\
&\sigma_{mf}^2, \sigma_{fm}^2 \sim \text{Half-Normal}(0,10)\\
&\ell_{d,i} \sim \text{Inv-Gamma}(\alpha_{d,i},\beta_{d,i}),
\end{aligned} \tag{15}$$

where the first K^2 entries of λ correspond to flows in the male-female direction and the remaining K^2 entries correspond to flows in the female-male direction, μ is the baseline log transmission intensity, ν is a scalar on the elements of λ in the male-female direction, and k_{mf}, k_{fm} are gender-specific squared exponential kernels with variance parameters σ_{mf}^2, σ_{fm}^2 and length scales $\ell_{mf,a}$, $\ell_{mf,b}$, $\ell_{fm,a}$, $\ell_{fm,b}$ (Rasmussen & Williams, 2006).

2.4 | Scalable numerical inference

The semi-parametric Poisson model can be efficiently fitted with the dynamic Hamiltonian Monte Carlo sampler of the Stan probabilistic programming framework (Carpenter et al., 2017).

The implementation uses the Hilbert Space Gaussian process approximation (HSGP) to the GP prior Equation (15) developed by Solin and Särkkä (2020), in which the squared exponential covariance kernel k in Equation (15) is approximated through a series expansion of eigenvalues and eigenfunctions of the Laplacian differential operator on a compact domain Ω of the input space. In our two-dimensional case, we consider $\Omega = [-B_1, B_1] \times [-B_2, B_2]$, and the boundary points play an important role in the approximation, and need to be specified appropriately.

Briefly, the HSGP approximation involves the spectral density associated with the stationary kernel of the GP prior. For the squared exponential kernel with two-dimensional inputs, it is given by

$$S_\theta(\boldsymbol{\omega}) = 2\pi\sigma^2 \prod_{d=1}^{2} \ell_d \exp\left(-\frac{1}{2}\sum_{i=1}^{2} \ell_d^2 \omega_d^2\right), \tag{16}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2)$ denote the frequencies, and $\theta = (\sigma, \ell_1, \ell_2)$ are the kernel parameters (Rasmussen & Williams, 2006). The approximation further involves the eigenvalues and eigenfunctions of the Laplacian differential operator. On the compact domain Ω, the jth univariate eigenvalues and eigenfunctions in dimension $d = 1, 2$ can be computed (Solin & Särkkä, 2020), and are $\lambda_{dj} = \left(\frac{j\pi}{2B_d}\right)^2$, $\phi_{dj}(x_d) = \sqrt{1/B_d}\sin\left(\sqrt{\lambda_{dj}}(x_d + B_d)\right)$. The HSGP approximation involves the first m_1 and m_2 such terms of both dimensions. There are $m = m_1 \times m_2$ possible combinations of such terms, which we index through $\boldsymbol{K} \in \mathbb{N}^{2\times m}$. For example, if $m_1 = 3$ and $m_2 = 4$, then

$$\boldsymbol{K} = \begin{pmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 \end{pmatrix}.$$

For 2D inputs, the $j = 1, \ldots, m$ eigenvalues and eigenfunctions are the combinations of the univariate eigenvalues and eigenfunctions, $\lambda_j = (\lambda_{\boldsymbol{K}_{1j}}, \lambda_{\boldsymbol{K}_{2j}})$, and $\phi_j(a, b) = \phi_{1\boldsymbol{K}_{1j}}(a)\phi_{2\boldsymbol{K}_{2j}}(b)$. This fully specifies the HSGP approximation to Gaussian processes with 2D inputs,

$$\begin{aligned} \boldsymbol{f} &\sim \mathcal{HSGP}(0, k^{\text{HSGP}}) \\ k^{\text{HSGP}}\left((a, b), (a', b')\right) &= \sum_{j=1}^{m} S_\theta\left(\sqrt{\lambda_j}\right)\phi_j(a, b)\phi_j(a', b'), \end{aligned} \tag{17}$$

which depends on the choice of m_1, m_2 and B_1, B_2. In our work, we applied the approximation (17) to each of the stacked Gaussian process prior components in Equation (15). To the best of our knowledge, this is one of the first applications of the HSGP approximation. Note that the structure of the data inputs and the form of the Poisson model is the same in many flow applications (van de Kassteele et al., 2017; Raymer et al., 2013), and so the HSGP approximation could make estimation of high-resolution flows also numerically scalable in these settings.

Full details on the algorithm are reported in Supplementary Text SS2 and a tutorial is provided in https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/vignettes/08_practical_example.md.

3 | APPLICATIONS

3.1 | Accuracy with and without adjustments for sampling heterogeneity

Standard phylodynamic methods ignore sampling differences between population strata (Le Vu et al., 2019; Scire et al., 2020; Volz et al., 2009). We first assessed the impact of sampling heterogeneity on estimating transmission flows in simulation experiments, that are fully reported in Supplementary Material, section S3. The first experiment is 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), that is not from our simpler likelihood model (10), and the simulated flow matrix $\boldsymbol{\pi}_r^0$ was recorded in $r = 1, \ldots, 100$ replicate simulations (Figure 2a). Observations were drawn under heterogeneous sampling, with fixed sampling probabilities $\xi_a = 0.6$ in group a, and decreasing sampling probabilities in groub b, $\xi_b = 0.6, 0.55, \ldots, 0.35$. First, we estimated flows from Equation (12) assuming no sampling differences between population strata, and which we implemented by setting $p(\xi_a|\mathbf{s}, \mathbf{X})$ and $p(\xi_b|\mathbf{s}, \mathbf{X})$ to the Beta density with parameters 25.5, 25.5. Second, we estimated flows with information on sampling differences included, by setting $p(\xi_a|\mathbf{s}, \mathbf{X})$ to the Beta density with shape parameters $N_a^s + \alpha_\xi$, $N^a - N_a^s + \beta_\xi$, where N_a^s are the number of sampled inviduals in group a, N_a is the population denominator, and the hyperparameters α_ξ, β_ξ were both set to 0.5 under Jeffrey's prior on the Binomial sampling probabilities. The density $p(\xi_b|\mathbf{s}, \mathbf{X})$ was specified analogously. Figure 2b compares the accuracy in the posterior median flow estimates $\hat{\pi}_{r,ab}$ in terms of the worst-case error (WCE) $\varepsilon_r = \max_{a,b} |\hat{\pi}_{r,ab} - \pi_{r,ab}^T|$, and illustrates that the Poisson model (12) can minimise the impact of sampling heterogeneity

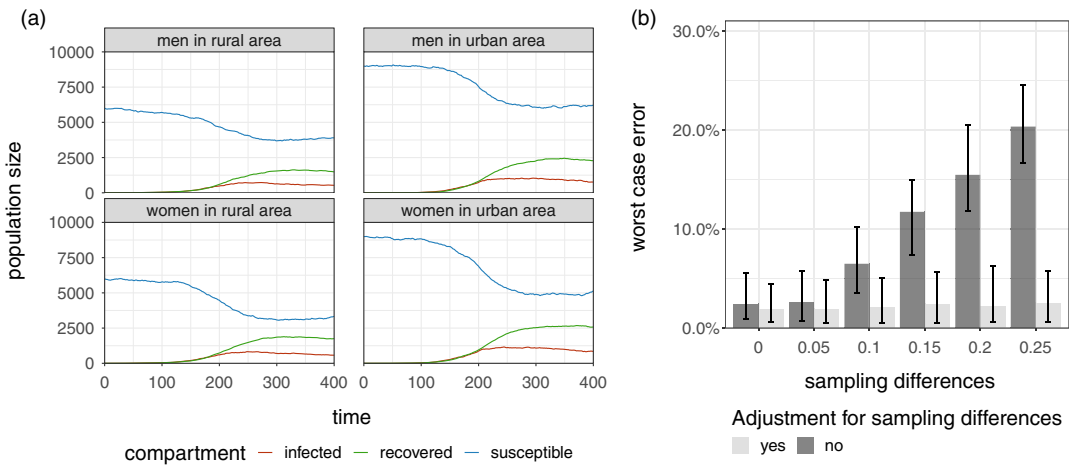


FIGURE 2 Simulation experiments to assess the accuracy of flow estimates under sampling bias. (a) ODE-based models were used to simulate epidemic trajectories of susceptible, infected and treated men and women across two population strata, and transmission flows between them. Panel A visualises one of the simulate trajectories. (b) Transmission flows were estimated under the Poisson likelihood model (10), without adjustments for sampling differences (dark grey), and with adjustments for sampling differences (light grey). Accuracy was measured with the worst-case error between posterior median estimates and the simulated true values, and shown are the average error and 95% range in 100 replicate simulations [Colour figure can be viewed at wileyonlinelibrary.com]

on flow estimates. More complex simulation experiments yielded similar results (Supplementary Material, section S3).

3.2 | Accuracy with different smoothing priors

Next, we assessed on simulations the impact of different regularising prior densities on estimating flows across orderable population strata, which is typically not considered in standard phylodynamic inference approaches (Bbosa et al., 2020; De Oliveira et al., 2017; Le Vu et al., 2019; Scire et al., 2020; Volz et al., 2009). We focused on age- and gender-specific transmission flows by 1-year age inputs between 15–24 years, resulting in $8 \times 10^2 = 800$ flow combinations, and simulated 300 transmission pairs from the GP model (15) using ground-truth parameters that were motivated by analyses of the Rakai data set (see Supplementary Material, section S3). Figure 3a illustrates the simulated transmission pairs from men to women, and the corresponding underlying flows in 2 of 20 simulations generated. Transmission flows were then inferred using the HSGP approximation Equations (15) to (17) in our semi-parametric Poisson flow model. We varied the number m of basis functions from 100 to 3600 by setting $m_1 = m_2 = 10, 20, \ldots, 60$, and chose as HSGP domain Ω an expanded version of the input domain, $[15/B, 50B] \times [15/B, 50B]$, with boundary

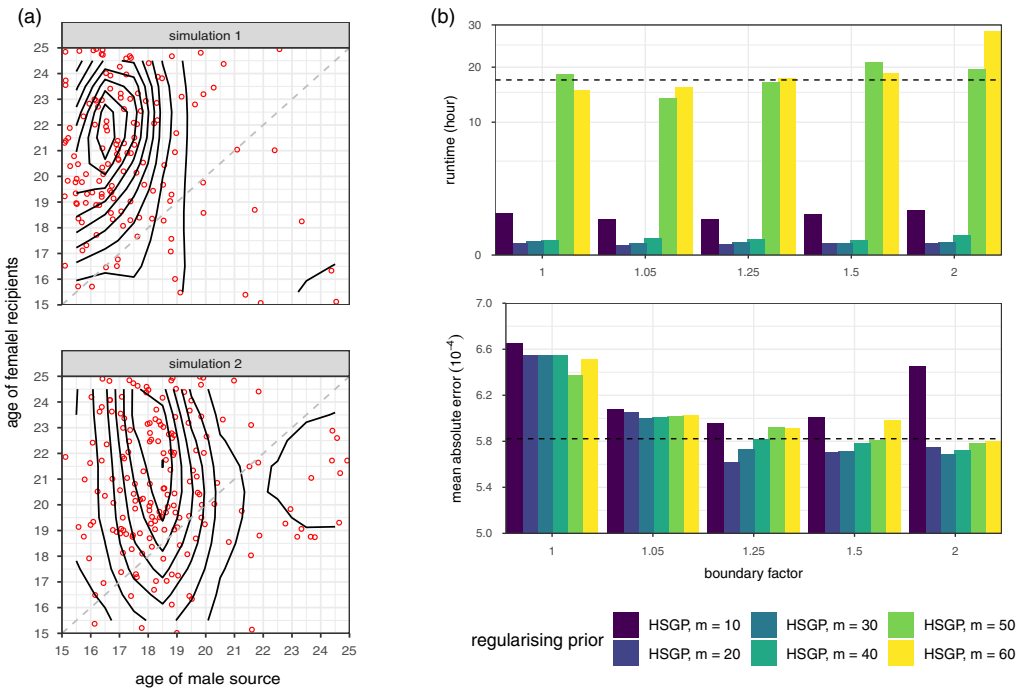


FIGURE 3 Simulation experiments to assess the accuracy of flow estimates under different smoothing priors. (a) Age- and gender-specific transmission pairs were simulated from the GP model (15), and shown are the simulated male to female transmission pairs for two simulations (red), along with contours of the simulated ground-truth transmission flow density. (b) Average runtimes of transmission flow inferences across 20 simulated data sets using different Hilbert Space Gaussian process (HSGP) and GP priors, and average mean absolute errors in inferred posterior median flow estimates. The dashed line corresponds to average runtimes and average mean absolute errors when using the GP prior, under which the simulations were generated [Colour figure can be viewed at wileyonlinelibrary.com]

factor $B = 1, 1.05, 1.25, 1.5, 2$. To have a benchmark for the performance of the HSGP approximations, inferences were also performed using the GP prior Equation (15), from which the data were simulated. Priors for all parameters are described in the Supplementary Material, section S3. Figure 3b shows that relative to using GP priors, average HMC runtimes across the 20 simulated data sets improved by more than ten-fold with HSGP priors when the number of basis functions was less than 50. Figure 3b also summarises the average mean absolute error in posterior median flow estimates. Similar to Riutort-Mayol et al. (2020), our results indicate that the HSGP approximations were least accurate for small $B \leq 1.05$, and for larger boundary factors when the number of basis functions was not increased simultaneously. On our simulations, HSGP approximations performed almost as well as GPs for the tuning parameters ($B = 1.25$, $m = 20-30$), ($B = 1.5$, $m = 20-40$) or ($B = 2$, $m = 30-40$). As computational cost increases with m, we chose $B = 1.25$, $m = 30$.

3.3 | Application to population-based deep-sequence data from Rakai, Uganda

We illustrate application of the semi-parametric Poisson flow model (12) 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., 2019, 2020). Between 2011/08/10 to 2015/01/30, two survey rounds were conducted in 36 inland communities, and three survey rounds in four fishing communities (Figure 4a). Preceeding each survey, a household census was conducted to identify individuals aged 15–49 years who lived in the communities for at least one month and with the intention to stay, who were eligible to participate. In brief, there were 37645 census-eligible individuals, of whom 25882 (68.8%) participated in the RCCS. Participation was higher among women than men, increased with age for both men and women, and was similar in fishing and inland communities (Supplementary Material, section S4). 11404 (96.9%) of non-participants were absent for school or work. Infected individuals who did not report ART use were selected for sequencing, and deep-sequencing rates were higher among men than women, similar by age for men and women, and were higher in fishing communities (Supplementary Material, section S4).

There were 293 heterosexual pairs with phylogenetic support for linkage and direction of transmission (source-recipient pairs) when using the threshold $\zeta = 0.6$ in Equation (7). The estimated infection times of the recipients were between 2009/10/01 and 2015/01/30, which defined the study period $\mathcal{T}$ during which we estimated transmission flows. Figure 4b illustrates the reconstructed source-recipient pairs by age of both individuals at the midpoint of the study period.

To interpret these observations, we defined as denominator transmission events to census-eligible individuals in RCCS communities who were infected during the study period $\mathcal{T}$, and formalised the individual steps in the sampling cascade of transmission events (Figure 5a). The sources and recipients had to participate in at least one survey round between 2011/08/10 and 2015/01/30, report no ART use, and have virus sequenced successfully. In Rakai, each survey was preceeded by a household census, and with this denominator we numerically estimated age-, gender- and location-specific conditional sampling probabilities at each step of the sampling cascade using Bayesian logistic-Binomial regression models, and then multiplied Monte Carlo draws from these distributions to numerically approximate the posterior distribution of the overall sampling probabilities ξ_a, ξ_b, $\forall a,b$ (see Supplementary Material, section S4). Figure 5b illustrates the resulting, overall sampling probabilities of female-to-male transmissions in fishing communities.

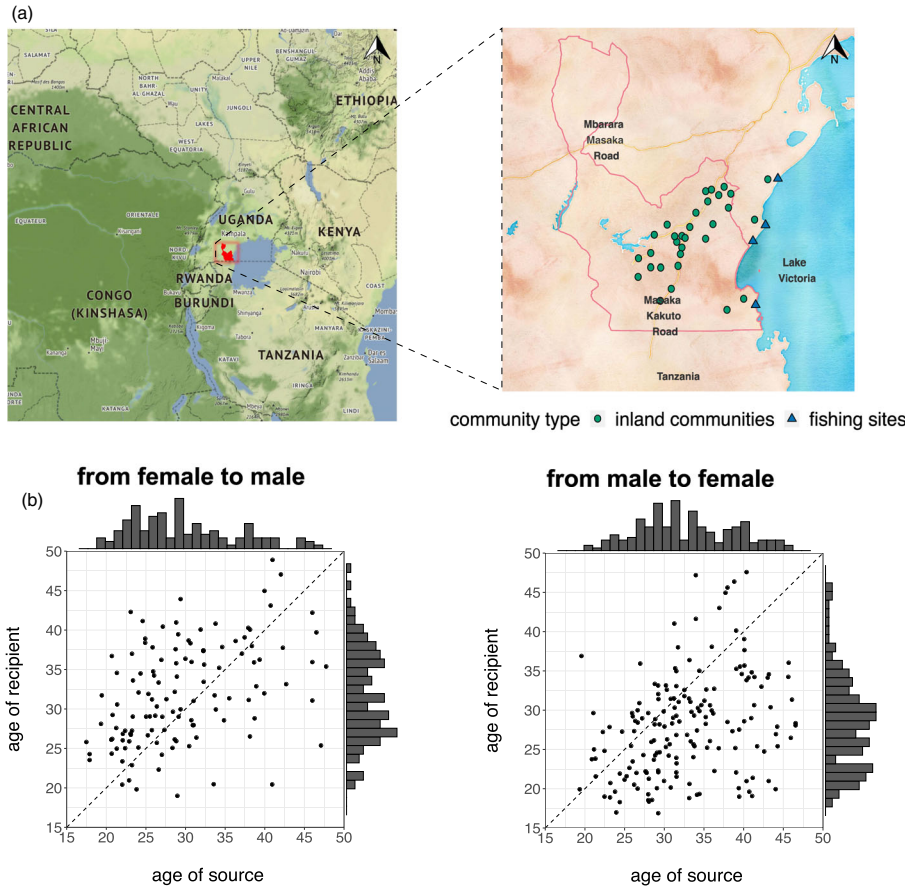


FIGURE 4 Location of the Rakai Community Cohort Study, and data. (a) Location of Rakai District (red) in south-eastern Uganda at the shores of Lake Victoria. HIV surveillance data were obtained from two survey rounds in 36 inland communities of the Rakai Community Cohort Study (green circles) and three survey rounds in the main four fishing communities within 3 km 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 study period [Colour figure can be viewed at wileyonlinelibrary.com]

The estimated sampling probabilities indicate that the observed data over-represent transmissions between older individuals.

3.4 | Transmission flows between areas with high and low disease prevalence

We then used the source-recipient data of Figure 4b to address problem (1) and estimate transmission HIV flows within and between high- and low-prevalence RCCS communities. The

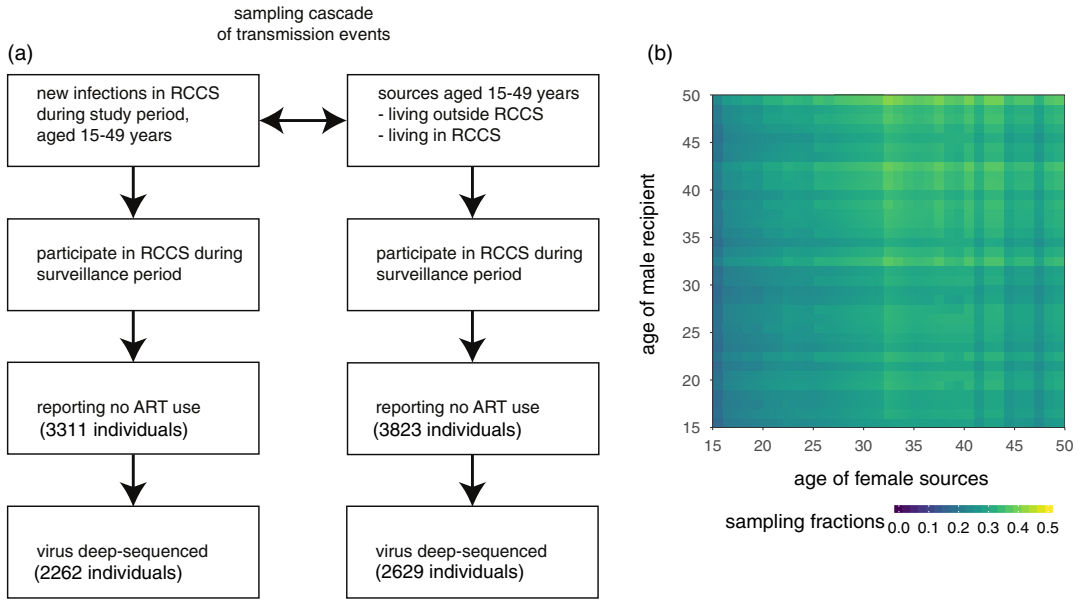


FIGURE 5 Model of the sampling cascade of transmission events. (a) The sampling cascade formalises the steps involved in sampling sources and recipients of transmission events. Recipients were defined as individuals aged 15-49 years who acquired infection in one of the Rakai Community Cohort Study communities during the study 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 we model using cohort data. (b) Estimated posterior median sampling probabilities of female to male transmission events in fishing communities by age of source and recipient. Conditional sampling probabilities were numerically estimated for each step of the sampling cascade using logistic Binomial regression, and then multiplied to give overall sampling probabilities [Colour figure can be viewed at wileyonlinelibrary.com]

high-prevalence communities comprised the four fishing communities, and the low-prevalence communities comprised the remaining 36 inland communities. Detailed analyses have been reported in Ratmann et al. (2020); here we focus on illustrating how known sampling heterogeneities can be accounted for, and how they affect inferences of transmission flow.

The participation, ART use, and sequence sampling probabilities differed by gender, age and location, and we stratified the population accordingly to meet the SARWS assumption that underlies the Poisson flow model (Section 2.2). Specifically, we stratified populations by gender, 1-year age bands (between 15 and 49 years), and resident location (low or high prevalence), which resulted in 140 sampling groups. Following our sampling cascade model (Figure 5), we then sought to estimate transmission flows between the 2×35^2 age- and gender-specific transmission flows for each of the four combinations of geographic source and recipient locations, through the joint posterior distribution (12). We further accounted for geographic in-migration, resulting in 14,700 flow variables. On this high-resolution flow space, we were able to directly apply the estimated, structured sampling probabilities that are illustrated in Figure 5b. This shows that accounting for the observed heterogeneities in how the census population was sampled resulted in a more complex inferential problem than (1–2) suggest.

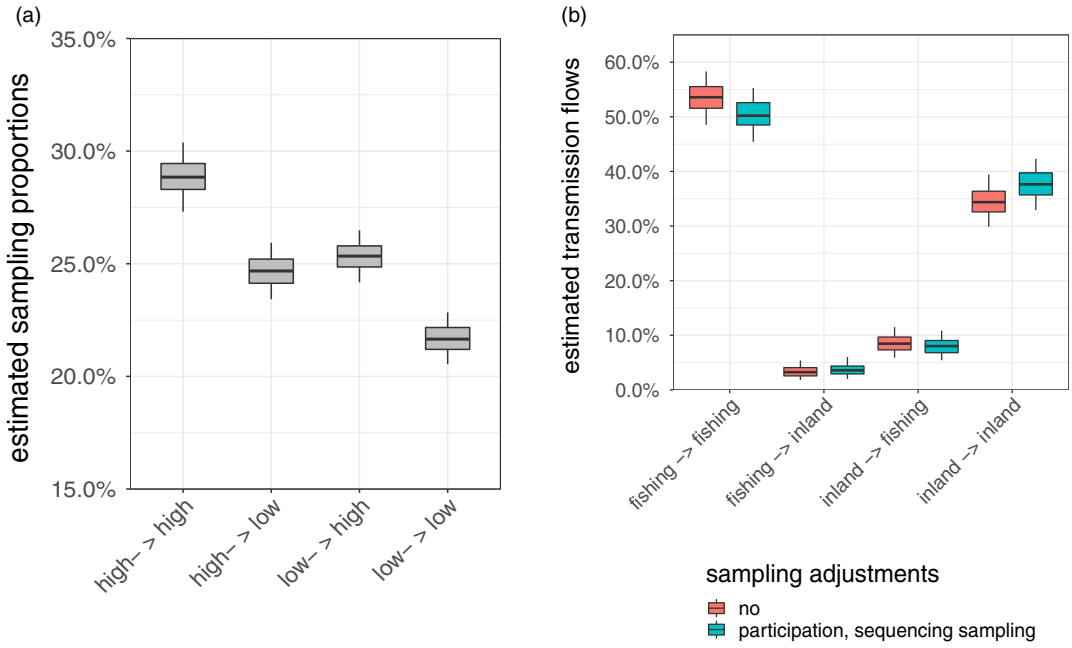


FIGURE 6 Estimated sampling probabilities of transmission events, and impact on flow estimates. (a) Boxplots of estimated pairwise sampling probabilities of transmission events between low-prevalence inland and high-prevalence fishing communities of the Rakai Community Cohort Study (RCCS). The sampling probabilities were obtained by marginalising over age- and gender-specific differences, and indicate that transmission events between high-prevalence fishing communities were over-represented in the data set. (b) Transmission flow estimates between low-prevalence inland and high-prevalence fishing communities of the RCCS. Shown are posterior median estimates (horizontal line), interquartile ranges (box), and 95% credible ranges when sampling heterogeneity by gender, age, and locations was ignored (red), and when sampling heterogeneity was accounted for as described in the main text [Colour figure can be viewed at wileyonlinelibrary.com]

To regularise inferences, we used the HSGP approximation (17) to the stacked Gaussian process prior (15). We further sought to allow for differences in transmission dynamics across locations, and for this reason specified independent HSGP priors on the parts of the flow space that correspond to transmissions to low prevalence areas, and on the parts of the flow space that correspond to transmissions to high prevalence areas. Numerical inference of the joint posterior density (12) took 89 h on four 2.4 Ghz processors with Stan version 2.19. There were no convergence, mixing or divergence warnings, as long as informative prior densities on the length scale hyperparameters were chosen, which we set by matching their 99% credible ranges to the empirical 99% quantiles in Figure 4. Thus flow inferences remained computationally manageable even in the high-resolution space considered here.

Figure 6b shows the marginal posterior estimates of the aggregated flow vector $\boldsymbol{\pi} = (\pi_{hh}, \pi_{hl}, \pi_{lh}, \pi_{ll})$, Equation (1), when sampling heterogeneity was ignored by setting all ξ_a to the average sampling probability (red), and when gender, age and location-specific participation and sequence sampling probabilities of sources and recipients were accounted for as described above (turquoise). The average sampling difference between individuals in fishing and inland communities was 7.16%, suggesting based on our results in Figure 2b

that after accounting for sampling differences, the sampling-adjusted estimates could differ by up to 5% from the unadjusted estimates. Figure 6b shows that our results are in line with this expectation. The estimated flow ratio (inland→fishing/fishing→inland) was 2.18 (1.06–4.71) when sampling heterogeneity was accounted for, and 2.58 (1.23–5.86) when sampling heterogeneity was not accounted for. We thus see that sampling heterogeneity can have an impact on flow estimates, and that the finding that high-prevalence fishing communities were net sinks, and not sources, of local infection flows is robust to sampling heterogeneity.

3.5 | Transmission flows between age groups

We next turned to problem (2) and estimated transmission flows by age and gender from the source-recipient data shown in Figure 4. Here, we focus on illustrating how the HSGP prior in the Poisson model allows us to borrow information across data points, and thereby go beyond existing phylodynamic methods (Bbosa et al., 2020; De Oliveira et al., 2017; Le Vu et al., 2019) and make flow inferences by 1-year age bands.

To do this, we compared estimates of the source vectors $\boldsymbol{\delta}$, defined in Equation (3a) when we used the independent Gamma prior density (14) (no regularisation) to those when we used the HSGP prior density (17) (with regularisation). We focused the comparison on the age- and gender-specific sources of infection regardless of location, that is, the source vector that corresponds to Equation (2), which was obtained by aggregating over the high- and low-prevalence locations of the source and recipient population groups. Figures 7a and b show the posterior median source estimates for each recipient group respectively without regularisation when using 5-year age bands and with regularisation using 1-year age bands, and Figures 7c and d show the corresponding posterior coefficients of variation. The estimated coefficients of variation were similar with and without regularisation, and well below 1 with regularisation, except from sources associated with little contribution to onward transmission and for very young or very old recipients. These findings suggest that the 1-year flow estimates are statistically meaningful, and at high resolution provide better insights. More detailed analyses are reported in Supplementary Text S5. First, we document the obvious, that it is not possible to estimate 2450 flow variables from 293 data points without regularisation. Second, we show that, as age bands are widened, estimates increasingly depend on the particular start and end points of the chosen age strata, and so we caution against inferences by 5-year age bands or wider.

3.6 | Sources of transmission to women aged <25 years

Across sub-Saharan Africa, HIV prevalence rises rapidly among young women aged <25 years (UNAIDS, 2018), which has prompted efforts to prevent infection among adolescent girls and young women, most notably the DREAMS partnership (Saul et al., 2018). Infections among young women aged <25 years are commonly attributed to older men, with a recent phylogenetic study from South Africa finding that of 60 identified transmission pairs involving women aged <25 years, 42 (70.0%) had a probable male partner aged >25 years (De Oliveira et al., 2017). Our larger data set with directional phylogenetic information allowed us to revisit these estimates.

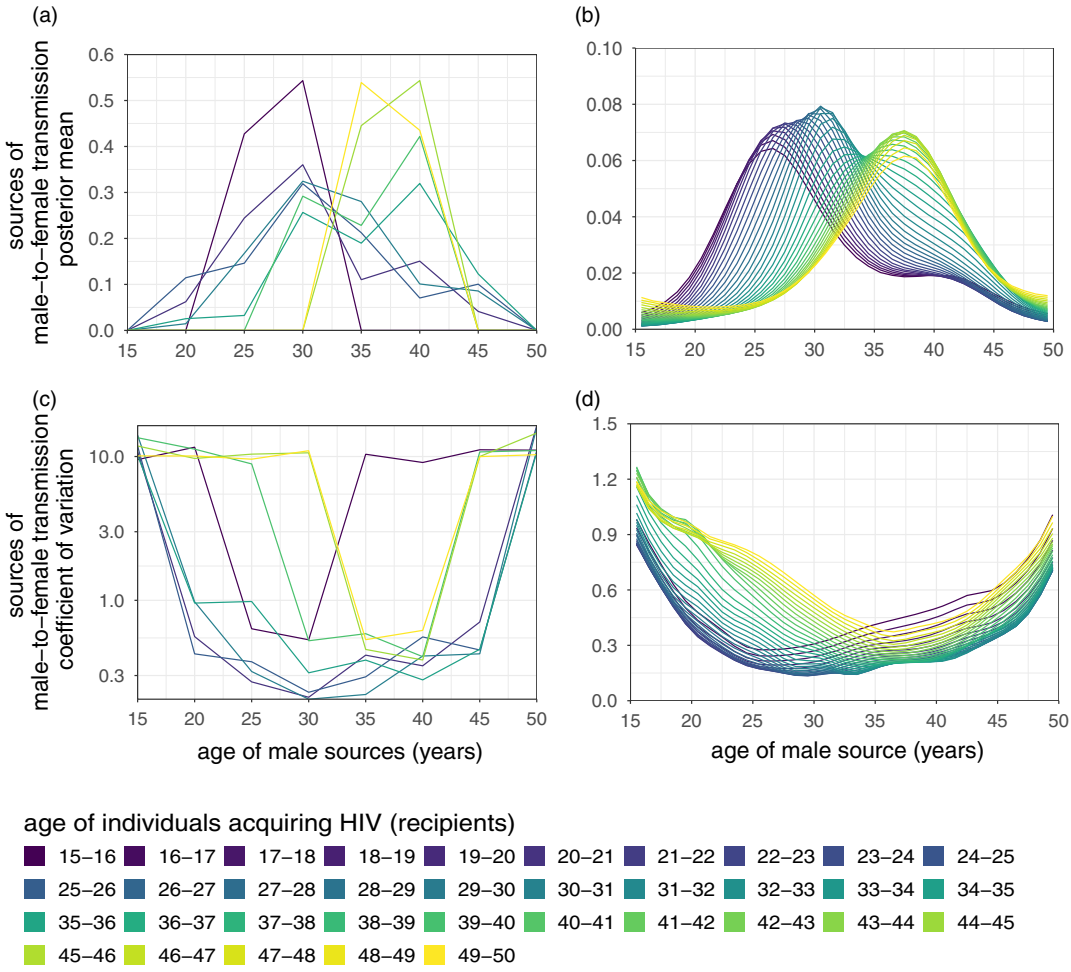


FIGURE 7 Impact of using requlatising prior densities on estimation of age- and gender-specific sources of infection. We compared estimates of the sources of HIV infection in women without regularisation, using the Gamma prior densities (14), to estimates obtained under the regularising Hilbert Space Gaussian process (HSGP) prior on the log transmission intensities (17). Each colour corresponds to infection recipients of a particular age. (a) Each line shows the posterior median estimates of the probability of infections attributed to the age of individuals of the opposite gender, i.e. the sources of infection, obtained without regularisation when using 5-year age strata. (b) Same as A, using the HSGP prior with boundary factor $B = 1.25$ and $m = 30$ basis functions when using 1-year age strata. (c)–(d) Corresponding estimates of the posterior coefficient of variation in the source estimates [Colour figure can be viewed at wileyonlinelibrary.com]

Our data contained 96 source-recipient pairs involving women aged <25 years, of whom 59 (61.5%) originated from men and 37 (38.5%) from women. However, under the regularising HSGP prior density (17), our gender- and age-specific flow estimates also borrow information from the other source-recipient pairs that involved older women. We report in Figure 8 the estimated sources of HIV infection in women of increasing age. The facets show the contribution of each source, men younger or the same age, men up to 5 years older, and men more than 5 years older, and sum to 100% for each age of infected women on the x-axis. At age 15, an estimated 91.8% (77.6%–97.7%) of women were infected by men more than 5 years older, while at age 20, this was

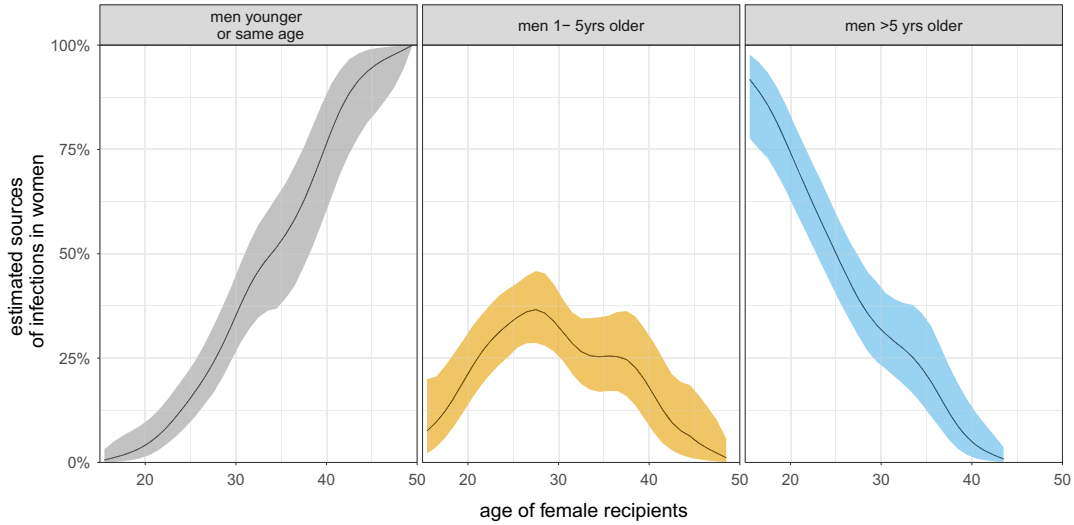


FIGURE 8 Estimated sources of infection in women in Rakai, Uganda, 2009-2015. We estimated the sources of HIV infection in women of increasing age. Sources were defined as men that were younger or the same age as the women (grey), 1–5 years older (orange), and over 5 years older (blue), and sum to 100% for each age of infected women on the x-axis. Posterior median source estimates are shown along with 95% marginal credibility intervals [Colour figure can be viewed at wileyonlinelibrary.com]

71.8% (60.6%–80.7%), and at age 25 this was 47.8% (38.2%–57.4%). These estimates document the overwhelming impact that men more than 5 years older have on driving infection in very young women in our observation period 2009–2015, and they show that the contribution of these men on infection in women declines rapidly with the age of the women. We provide exact estimates in Table S9.

4 | DISCUSSION

In this study, we introduce a class of semi-parametric Bayesian Poisson models for estimating high-resolution flows between population strata, and apply the model to estimate the sources of HIV infections from pathogen deep-sequence data. The modelling framework is flexible, and enables addressing a range of epidemiological questions on pathogen spread between geographic areas, by 1-year age bands, gender, or indeed other discretely-valued sociodemographic characteristics. We templated the model with and without Hilbert space approximations for scalable inference of high-resolution flows in generic Stan model files, and hope that given the canonical structure of the semi-parametric Poisson flow model, these will be helpful in other movement, origin-destination or flow applications as well (Faye et al., 2015; Hazelton, 2001; van de Kassteele et al., 2017; Lindström et al., 2013; Raymer et al., 2013; Sun et al., 2021; Tebaldi & West, 1998).

Existing phylodynamic estimation approaches are tailored for pathogen consensus sequences (Lemey et al., 2009; Scire et al., 2020; Vaughan et al., 2014; Volz et al., 2009). The approach described here is tailored for pathogen deep-sequence data, in that an observed, time-homogeneous flow matrix is required as input, which can be derived through aggregation from individual source-recipient relationships in deep-sequence phylogenies. The main

advantages are first, that population-level spread can be directly estimated from individual source-recipient relationships, and modelled in terms of associated individual covariates. In comparison, standard phylogeographic models estimate transition rates from the shape of viral phylogenies captured in times to lineage coalescence (Lemey et al., 2009; Scire et al., 2020; Stadler & Bonhoeffer, 2013), which are often harder to interpret. Second, relatively little computational effort is needed to fit the Poisson flow model (12–14) to deep-sequence data, because it falls within the class of Bayesian hierarchical models for binary data, for which efficient fitting and regularisation techniques exist (Carpenter et al., 2017; Rasmussen & Williams, 2006; Solin & Särkkä, 2020). This makes it computationally feasible to investigate complex aspects of disease spread such as population-level HIV transmission by 1-year age bands. Third, differences in how the phylogenetic data were sampled for each stratum can be explicitly accounted for in the model. This is particularly important for characterising HIV transmission, which tends to concentrate in marginalised, vulnerable, and hard to reach populations (UNAIDS, 2019). We found relatively small differences in the estimated sources of infection in the study communities by location and age when inferences were performed with and without sampling adjustments, in line with the relatively limited differences in sampling inclusion probabilities in the RCCS communities and the expected impact on source attribution on simulated data (Figure 2b). However in other use cases, the impact of sampling heterogeneity on source attribution can be substantially larger. For example, the RCCS included in surveyed communities an estimated 75.7% of the lakeside population within 3 km of the shoreline of Lake Victoria along the Rakai region, and an estimated 16.2% of the inland population of the Rakai region (Ratmann et al., 2020). We can use the proposed framework to extrapolate our inferences from the RCCS communities to the underlying population, and given the larger sampling differences we estimated that 88.7% (84.5–91.9%) of transmissions in the Rakai region occurred in the inland population (Ratmann et al., 2020). Specifying and characterising the denominator population is thus crucial for interpreting phylogenetic source attribution estimates, and we believe the proposed Bayesian semi-parametric flow model provides a useful tool in this endeavour.

The method we propose has limitations. First, the method requires pathogen deep-sequence data instead of consensus sequences, which at present are uncommon, though increasingly generated in routine clinical care (Houlihan et al., 2018). Second, current deep-sequencing protocols typically 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 fast rate, 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 HIV or hepatitis C. 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 (Rhoads & Au, 2015). 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, and we cannot exclude that this selection step introduces bias into flow estimates. Fourth, the model was not designed to estimate time changes in transmission flows. While in principle it is possible to add time as a covariate to the linear predictor of the log transmission intensities (15), the resulting flow estimates will in general not be consistent with the constraints imposed by standard assumptions on disease spread, as in Equations (4) and (5), and other techniques such as the structured coalescent may be better suited (Volz et al., 2009).

Reducing HIV incidence among adolescent and young women is a key priority for public health programmes across sub-Saharan Africa to achieve epidemic control milestones (UNAIDS, 2018). The DREAMS intervention aims to promote determined, resilient, empowered, AIDS-free, mentored, and safe adolescent girls and young women, and includes educational programmes that aim to address the socio-behavioural factors that underlie vulnerability and infection risk (Saul et al., 2018). Our analysis of a large cross-sectionally sampled HIV deep-sequence data set from Rakai, Uganda, supports previous analyses (Bbosa et al., 2020; De Oliveira et al., 2017; Probert et al., 2019) and indicates that 68.9% (60.2%–76.9%) of adolescent and young women aged <25 years acquired HIV in age-disparate relationships with men at least 5 years older. The estimated proportion of infections attributable to age-disparate relationships was approximately 90% among adolescent girls, and decreased to approximately 50% among women aged 25 years during the study period 2009–2015. Taken together, the data from this study and other phylogenetic studies from Uganda, Zambia, and South Africa suggest that rapid increases in HIV prevalence among adolescent and young women may be driven by the same source populations across sub-Saharan Africa, and support DREAMS interventions that include clear prevention messages about age-disparate sexual relationships.

ACKNOWLEDGEMENTS

This study was supported by the Bill & Melinda Gates Foundation (OPP1175094, OPP1084362), the National Institute of Allergy and Infectious Diseases (R01AI110324, U01AI100031, U01AI075115, R01AI102939, K01AI125086-01), National Institute of Mental Health (R01MH107275), the National Institute of Child Health and Development (RO1HD070769, R01HD050180), the Division of Intramural Research of the National Institute for Allergy and Infectious Diseases, the World Bank, the Doris Duke Charitable Foundation, the Johns Hopkins University Center for AIDS Research (P30AI094189), and the Presidents Emergency Plan for AIDS Relief through the Centers for Disease Control and Prevention (NU2GGH000817). We acknowledge data management support provided in part by the Office of Cyberinfrastructure and Computational Biology at the National Institute for Allergy and Infectious Diseases, computational support through the Imperial College Research Computing Service, doi: 10.14469/hpc/2232. We thank the participants of the Rakai Community Cohort Study and the many staff and investigators who made this study possible, as well as the PANGEA-HIV steering committee, the RCCS leadership, and two anonymous reviewers for their helpful comments on this manuscript.

ORCID

Xiaoyue Xi https://orcid.org/0000-0002-7490-282X
Simon E. F. Spencer https://orcid.org/0000-0002-8375-5542
Matthew Hall https://orcid.org/0000-0002-2671-3864
Joseph Kagaayi https://orcid.org/0000-0003-4101-3038
Oliver Ratmann https://orcid.org/0000-0001-8667-4118

REFERENCES

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

Ailloud, F., Didelot, X., Woltemate, S., Pfaffinger, G., Overmann, J., Bader, R.C. et al. (2019) Within-host evolution of helicobacter pylori shaped by niche-specific adaptation, intragastric migrations and selective sweeps. *Nature Communications*, 10, 1–13.

Anderson, R.M. & May, R.M. (1992) *Infectious diseases of humans: dynamics and control*. Oxford: Oxford University Press.

Barré-Sinoussi, F., Abdool Karim, S.S., Albert, J., Bekker, L.-G., Beyrer, C., Cahn, P. et al. (2018) Expert consensus statement on the science of HIV in the context of criminal law. *Journal of the International AIDS Society*, 21, e25161.

Bbosa, N., Ssemwanga, D., Ssekagiri, A., Xi, X., Mayanja, Y., Bahemuka, U. et al. (2020) Phylogenetic and demographic characterization of directed HIV-1 transmission using deep sequences from high-risk and general population cohorts/groups in uganda. *Viruses*, 12, 331.

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

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M. et al. (2017) Stan: a probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.

Cohen, M.S., Chen, Y.Q., McCauley, M., Gamble, T., Hosseinipour, M.C., Kumarasamy, N. et al. (2011) Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*, 365, 493–505.

De Oliveira, T., Kharsany, A.B., Gräf, T., Cawood, C., Khanyile, D., Grobler, 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.

Dwyer-Lindgren, L., Cork, M.A., Sligar, A., Steuben, K.M., Wilson, K.F., Provost, N.R. 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. et al. (2014) The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346, 56–61.

Faye, O., Boëlle, P.-Y., Heleze, E., Faye, O., Loucoubar, C., Magassouba, N. et al. (2015) Chains of transmission and control of ebola virus disease in conakry, guinea, in 2014: an observational study. *The Lancet Infectious Diseases*, 15, 320–326.

Gall, A., Ferns, B., Morris, C., Watson, S., Cotten, M., Robinson, M. et al. (2012) Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *Journal of Clinical Microbiology*, 50, 3838–3844.

Givens, G.H., Smith, D. & 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*, 12, 221–240.

Grabowski, M.K., Serwadda, D.M., Gray, R.H., Nakigozi, G., Kigozi, G., Kagaayi, J. 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. et al. (2019) Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *eLife*, 8, e46402.

Hayes, R.J., Donnell, D., Floyd, S., Mandla, N., Bwalya, J., Sabapathy, K. (2019) Effect of universal testing and treatment on HIV incidence — HPTN 071 (PopART). *New England Journal of Medicine*, 381, 207–218.

Hazelton, M.L. (2001) Inference for origin–destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological*, 35, 667–676.

Houlihan, C.F., Frampton, D., Ferns, R.B., Raffle, J., Grant, P., Reidy, M. et al. (2018) Use of whole-genome sequencing in the investigation of a nosocomial influenza virus outbreak. *The Journal of Infectious Diseases*, 218, 1485–1489.

van de Kassteele, J., van Eijkeren, J., Wallinga, J. (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. et al. (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. & 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. & Suchard, M.A. (2009) Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5, e1000520.

Lindström, T., Grear, D.A., Buhnerkempe, M., Webb, C.T., Miller, R.S., Portacci, K. et al. (2013) A bayesian approach for modeling cattle movements in the united states: scaling up a partially observed network. *PLoS One*, 8, e53432.

Miller, H.J., Dodge, S., Miller, J. & Bohrer, G. (2019) Towards an integrated science of movement: converging research on animal movement ecology and human mobility science. *International Journal of Geographical Information Science*, 33, 855–876.

Poon, A.F., Gustafson, R., Daly, P., Zerr, L., Demlow, S.E., Wong, J. 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.

Probert, W., Hall, M., Xi, X., Sauter, R., Golubchik, T., Bonsall, D. et al. (2019) Quantifying the contribution of different aged men and women to onwards transmission of HIV-1 in generalised epidemics in sub-Saharan Africa: A modelling and phylogenetics approach from the HPTN071 (PopART) trial.

Rasmussen, C.E. & Williams, C. (2006) *Gaussian processes for machine learning*. Cambridge, Massachusetts: MIT Press.

Ratmann, O., Van Sighem, A., Bezemer, D., Gavryushkina, A., Jurriaans, S., Wensing, A. et al. (2016) Sources of HIV infection among men having sex with men and implications for prevention. *Science Translational Medicine*, 8, 320ra2.

Ratmann, O., Grabowski, M.K., Hall, M., Golubchik, T., Wymant, C., Abeler-Dörner, L. et al. (2019) Inferring HIV-1 transmission networks and sources of epidemic spread in africa with deep-sequence phylogenetic analysis. *Nature Communications*, 10, 1–13.

Ratmann, O., Kagaayi, J., Hall, M., Golubchick, T., Kigozi, G., Xi, X. et al. (2020) Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda. *The Lancet HIV*, 7, PE173–E183.

Raymer, J., Wiśniowski, A., Forster, J.J., Smith, P.W. & Bijak, J. (2013) Integrated modeling of European migration. *Journal of the American Statistical Association*, 108, 801–819.

Rhoads, A. & Au, K.F. (2015) PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13, 278–289.

Riutort-Mayol, G., Bürkner, P.-C., Andersen, M.R., Solin, A. & Vehtari, A. (2020) Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. arXiv preprint arXiv:2004.11408.

Saul, J., Bachman, G., Allen, S., Toiv, N., Cooney, C. & Beamon, T. (2018) Determined resilient empowered AIDS-free mentored and safe (DREAMS): what is the core package and why now. *PLOS One*, 13, e0208167.

Scire, J., Barido-Sottani, J., Kühnert, D., Vaughan, T.G. & Stadler, T. (2020) Improved multi-type birth-death phylodynamic inference in BEAST 2. bioRxiv, 2020.01.06.895532.

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

Solin, A. & Särkkä, S. (2020) Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, 30, 419–446.

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

Sun, K., Wang, W., Gao, L., Wang, Y., Luo, K., Ren, L. et al. (2021) Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science*, 371, eabe2424.

Tebaldi, C. & West, M. (1998) Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93, 557–573.

UNAIDS. (2018) Miles to go: closing gaps, breaking barriers, righting justice, document jc2924. Available from: https://www.unaids.org/en/resources/documents/2018/global-aids-update

UNAIDS. (2019) UNAIDS Data 2019, document jc2959e. Available from https://www.unaids.org/en/resources/documents/2019/2019-UNAIDS-data

Vaughan, T.G., Kühnert, D., Popinga, A., Welch, D. & Drummond, A.J. (2014) Efficient Bayesian inference under the structured coalescent. *Bioinformatics*, 30, 2272–2279.

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

Volz, E.M., Ionides, E., Romero-Severson, E.O., Brandt, M.-G., Mokotoff, E. & Koopman, J.S. (2013) HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS Medicine*, 10, e1001568.

Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M. et al. (2017) PHYLOSCANNER: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular Biology and Evolution*, 35, 719–733.

Zhang, Y., Wymant, C., Laeyendecker, O., Grabowski, M.K., Hall, M., Hudelson, S. et al. (2020) Evaluation of phylogenetic methods for inferring the direction of HIV transmission: HPTN 052. Clinical Infectious Diseases, Epub ahead of print, ciz1247.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Xi, X., Spencer, S.E.F., Hall, M., Grabowski, M.K., Kagaayi, J., Ratmann, O. et al. (2022) Inferring the sources of HIV infection in Africa from deep-sequence data with semi-parametric Bayesian Poisson flow models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(3), 517–540. Available from: https://doi.org/10.1111/rssc.12544