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

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.

Matthew Hall

Big Data Institute, University of Oxford, Oxford OX3 7LF, United Kingdom.

M Kate Grabowski

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

Joseph Kagaayi

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

S1. Maximum likelihood flow estimates under heterogeneous sampling

We here describe derivation of the maximum likelihood estimates of the Poisson model (10),

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

when 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. We denote the vector of sampled individuals across population strata by $\boldsymbol{N}^s = (N_a^s)_{a \in \mathcal{A}}$, and similarly the vector of individuals in

each population group by $\boldsymbol{N} = (N_a)_{a\in\mathcal{A}}$. Then the product likelihood is

$$p(\mathbf{n}, \boldsymbol{N}, \boldsymbol{N}^s|\boldsymbol{\lambda}, \boldsymbol{\xi}) = \left(\prod_{a,b} \text{Poisson}(n_{ab}; \lambda_{ab}\xi_a\xi_b)\right)\left(\prod_a \text{Binomial}(N_a^s; N^a, \xi_a)\right). \quad \text{(S1)}$$

Taking the derivative of the log-likelihood (S1) to zero gives

$$\begin{aligned} \hat{\pi}_{ab}\hat{\eta} &= \frac{n_{ab}}{\hat{\xi}_a\hat{\xi}_b} \\ \hat{\xi}_a &= \frac{N_a^s}{N_a} \end{aligned} \quad \text{(S2)}$$

As $\sum_{ab}\hat{\pi}_{ab} = 1$,

$$\hat{\eta} = \sum_{a,b}\frac{n_{ab}}{\hat{\xi}_a\hat{\xi}_b}, \quad \text{(S3)}$$

and we obtain

$$\hat{\pi}_{ab} = \frac{n_{ab}}{\hat{\xi}_a\hat{\xi}_b}\Big/\sum_{a,b}\frac{n_{ab}}{\hat{\xi}_a\hat{\xi}_b}. \quad \text{(S4)}$$

S2. Numerical inference algorithms

We here describe numerical algorithms for estimating transmission flows with the Poisson flow model (10).

It is possible to implement the flow model (10) in the Stan computing language (Carpenter et al., 2017). We hope this feature enhances robust numerical inference in a well-tested software environment, facilitates model sharing and model extensions, and gives end-users the option to apply alternative inference algorithms. For the work presented in this paper, we have used the dynamic Hamiltonian Monte Carlo algorithm of the Stan probabilistic programming framework in Stan version 2.21. Section S2.1 describes the Stan model specification for inference of the joint posterior distribution (12) when population strata are unordered and the prior density on the transmission intensities $\boldsymbol{\lambda}$ is independent Gamma (14). Section S2.2 describes the Stan model specification for inference when population strata can be ordered and the prior densities for the transmission intensities are correlated through a Gaussian process prior (15). For simplicity, we describe models and algorithms for the case that sampling probabilities are the same among source and recipient cases. Extensions to different sampling probabilities are straightforward. Stan is not necessary for numerical inference. To illustrate this, we describe in Section S2.3 a custom MCMC within Gibbs sampler for inference of the joint posterior distribution (12) when population strata are unordered and the prior density on the transmission intensities $\boldsymbol{\lambda}$ is independent Gamma (14).

S2.1. Stan implementation using independent prior on transmission intensities

When population strata are unordered, we focus on inference of the posterior distribution

$$\begin{aligned} p(\boldsymbol{\lambda}, \boldsymbol{\xi}|\mathbf{n}, \boldsymbol{s}, \mathbf{X}) &\propto p(\mathbf{n}|\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{s}, \mathbf{X}) p(\boldsymbol{\lambda}|\boldsymbol{\xi}, \boldsymbol{s}, \mathbf{X}) p(\boldsymbol{\xi}|\boldsymbol{s}, \mathbf{X}) \\ &= p(\mathbf{n}|\boldsymbol{\lambda}, \boldsymbol{\xi}) p(\boldsymbol{\lambda}|\boldsymbol{\xi}) p(\boldsymbol{\xi}|\boldsymbol{s}, \mathbf{X}) \\ &= \Big(\prod_{a,b} \text{Poisson}(n_{ab}; \lambda_{ab}\xi_a\xi_b) \Big) \Big(\prod_{a,b} \text{Gamma}(\lambda_{ab}; \alpha_{ab}, \beta) \Big) p(\boldsymbol{\xi}|\boldsymbol{s}, \mathbf{X}) \end{aligned} \tag{S5}$$

where

- $\boldsymbol{\lambda}$ vector of transmission intensities from group a to group b, with length L,
- $\boldsymbol{\xi}$ vector of sampling proportions for each population group a, $a \in \mathcal{A}$,
- $\boldsymbol{n}$ vector of observed transmission counts, with length L,
- $\boldsymbol{s}$ sampling status vector, with length N,
- $\mathbf{X}$ matrix of sampling characteristics, with dimension $N \times p$,

and α_{ab}, β are as in (14), and the flow vector $\boldsymbol{\pi}$ is obtained from the transformation $\pi_{ab} = \lambda_{ab}/(\sum_{c,d} \lambda_{cd})$.

To specify the posterior density $p(\boldsymbol{\xi}|\boldsymbol{s}, \mathbf{X})$ in Stan, we further approximated its components $p(\xi_a|\boldsymbol{s}, \mathbf{X})$ through a suitable closed-form density. In our applications, we opted for Beta densities with shape and rate parameters estimated via maximum likelihood. The resulting Stan model file was provided in the file named prior_gamma.txt.

Use of the Stan model is illustrated in the online example script `https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/vignettes/07_age_analysis.md`.

S2.2. Stan implementation using Gaussian process prior on transmission intensities

When population strata can be ordered as in the application to age-specific transmission flows, we focus on inference of the posterior distribution (S5) with the independent Gamma priors replaced by a two-dimensional Hilbert space Gaussian process priors on the non-zero components of $\boldsymbol{\lambda}$, (15). To implement this in Stan, Equation (17) can be more compactly written as

$$k^{\text{HSGP}}\big((a,b),(a',b')\big) = \boldsymbol{\phi}(a,b)^T \Delta\, \boldsymbol{\phi}(a',b'), \tag{S6}$$

where $\boldsymbol{\phi}(a,b) = (\phi_j(a,b))_{j=1}^m \in \mathbb{R}^m$ is the column vector of eigenfunctions and $\Delta \in \mathbb{R}^{m\times m}$ is the diagonal matrix whose (j,j)th entry is $S_\theta(\sqrt{\boldsymbol{\lambda_j}})$. Consequently the HSGP Gram matrix for n observations and associated inputs $\{(a_i,b_i)\}_{i=1}^n \in \Omega$ is

$$K^{\text{HSGP}} = \boldsymbol{\Phi}\Delta\boldsymbol{\Phi}^T, \tag{S7}$$

where

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1(a_1, b_1) & \cdots & \phi_m(a_1, b_1) \\ \vdots & \ddots & \vdots \\ \phi_1(a_n, b_n) & \cdots & \phi_m(a_n, b_n) \end{pmatrix} \in \mathbb{R}^{n \times m}.$$

The two-dimensional zero-mean HSGP model is fully defined in terms of the kernel (17), as the stochastic process for which any finite set of two-dimensional inputs in Ω follows a multivariate normal distribution with variance-covariance matrix specified by (17), and this is vectorised as (S7). Using Cholesky decomposition, the HSGP model (17) can be conveniently calculated from m standard normal random variables through

$$f(a, b) = \sum_{j=1}^{m} \sqrt{S_\theta(\sqrt{\boldsymbol{\lambda}_j}))}\phi_j(a, b)\beta_j \tag{S8}$$

where $\beta_j \sim \mathcal{N}(0, 1)$ for $j = 1, \ldots, m$. The resulting Stan model file is prior_gpapprox.txt.

Use of the Stan model is illustrated in the online example script `https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/vignettes/07_age_analysis.md`. We explored optimal choices of the tuning parameters of the HSGP approximation on simulations, please see Section 3.2 and S3.4 for details. The Stan file using the GP prior without approximation is in prior_gp.txt as a comparison to approximated GP priors.

S2.3. Markov Chain Monte Carlo within Gibbs algorithm

This section describes implementation details of a custom Metropolis-within-Gibbs algorithm to estimate the posterior distribution (S5),

$$p(\boldsymbol{\lambda}, \boldsymbol{\xi}|\mathbf{n}, \boldsymbol{s}, \mathbf{X}) \propto \left(\prod_{a,b} \text{Poisson}(n_{ab}); \lambda_{ab}\xi_a\xi_b) \right) \left(\prod_{a,b} \text{Gamma}(\lambda_{ab}; \alpha_{ab}, \beta) \right) p(\boldsymbol{\xi}|\boldsymbol{s}, \mathbf{X}).$$

The algorithm is also implemented in the `phyloflows` R package, version 1.2.0.

S2.3.1. Overall structure of algorithm

The algorithm exploits the factorisation of the posterior density (S5) into the full conditionals

$$\begin{aligned} &p(\xi_a|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \boldsymbol{s}, \mathbf{n}, \mathbf{X}) \text{ for all } a, \\ &p(\lambda_{ab}|\boldsymbol{\xi}, \boldsymbol{\lambda}_{-ab}, \boldsymbol{s}, \mathbf{n}, \mathbf{X}) \text{ for all } a, b, \end{aligned}$$

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 $p(\xi_a|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \boldsymbol{s}, \mathbf{n}, \mathbf{X})$ is not available in closed form, and we performed Metropolis-within-Gibbs updates for each ξ_a. The density $p(\lambda_{ab}|\boldsymbol{\xi}, \boldsymbol{\lambda}_{-ab}, \boldsymbol{s}, \mathbf{n}, \mathbf{X})$ is of Gamma form, and so a Gibbs step can be used to update λ_{ab}. 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}.

S2.3.2. Metropolis-Hastings within Gibbs steps

The algorithm starts by updating in turn the sampling proportions ξ_a of each population group $a \in \mathcal{A}$. The full conditional distribution of ξ_a is

$$\begin{aligned} &p(\xi_a|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \mathbf{n}, \boldsymbol{s}, \mathbf{X}) \\ &\quad\propto \Pi_{c=a \text{or} d=a} p(n_{cd}|\lambda_{cd}, \xi_c\xi_d) p(\xi_a|\boldsymbol{s}, \mathbf{X}) \\ &\quad\propto \Pi_{c=a \text{or} d=a} \text{Poi}(n_{cd}; \lambda_{cd}\xi_c\xi_d) \text{Gamma}(\lambda_{cd}; \xi_c\xi_d) p(\xi_a|\boldsymbol{s}, \mathbf{X}), \end{aligned} \tag{S9}$$

which does not have closed form, and so a Metropolis-Hastings update is used. We sought to avoid tuning parameters as much as possible, and for this reason proposed moves from the prior,

$$q(\xi'_a|\xi_a) = p(\xi'_a|\boldsymbol{s}, \mathbf{X}). \tag{S10}$$

The resulting Metropolis Hasting ratio is

$$\begin{aligned} r &= \frac{p(\xi'_a|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \mathbf{n}, \boldsymbol{s}, \mathbf{X})}{p(\xi_a|\boldsymbol{\xi}_{-a}, \boldsymbol{\lambda}, \mathbf{n}, \boldsymbol{s}, \mathbf{X})} \times \frac{q(\xi_a)}{q(\xi'_a)} \\ &= \frac{\Pi_b p(n_{ab}|\lambda_{ab}, \xi'_a\xi_b) \Pi_{b\neq a} p(n_{ba}|\lambda_{ba}, \xi_b\xi'_a) \Pi_b p(\lambda_{ab}|\boldsymbol{\xi}') \Pi_{b\neq a} p(\lambda_{ba}|\boldsymbol{\xi}') p(\xi'_a|\boldsymbol{s}, \mathbf{X})}{\Pi_b p(n_{ab}|\lambda_{ab}, \xi_a\xi_b) \Pi_{b\neq a} p(n_{ba}|\lambda_{ba}, \xi_b\xi_a) \Pi_b p(\lambda_{ab}|\boldsymbol{\xi}) \Pi_{b\neq a} p(\lambda_{ba}|\boldsymbol{\xi}) p(\xi_a|\boldsymbol{s}, \mathbf{X})} \\ &\qquad \times \frac{p(\xi_a|\boldsymbol{s}, \mathbf{X})}{p(\xi'_a|\boldsymbol{s}, \mathbf{X})} \\ &= \frac{\Pi_b p(n_{ab}|\lambda_{ab}, \xi'_a\xi_b) \Pi_{b\neq a} p(n_{ba}|\lambda_{ba}, \xi_b\xi'_a) \Pi_b p(\lambda_{ab}|\boldsymbol{\xi}') \Pi_{b\neq a} p(\lambda_{ba}|\boldsymbol{\xi}')}{\Pi_b p(n_{ab}|\lambda_{ab}, \xi_a\xi_b) \Pi_{b\neq a} p(n_{ba}|\lambda_{ba}, \xi_b\xi_a) \Pi_b p(\lambda_{ab}|\boldsymbol{\xi}) \Pi_{b\neq a} p(\lambda_{ba}|\boldsymbol{\xi})}, \end{aligned} \tag{S11}$$

where for brevity we denoted by $\boldsymbol{\xi}'$ the vector of sampling probabilities such that $\xi'_c = \xi_c$ for all indices $c \neq a$ and ξ'_a inserted at index a. Proposed moves are accepted with probability $\min(1, r)$.

In our applications, different sampling cascades apply to source and recipient population groups, which we denote respectively by ξ^s_a and ξ^r_a. Equations (S9-S11) generalise straightforwardly to this setting, and the algorithms updates the sampling probabilities $\boldsymbol{\xi}^s$ and $\boldsymbol{\xi}^r$ in turn.

S2.3.3. Gibbs step

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

$$\begin{aligned} p(\lambda_{ab}|\boldsymbol{\xi}, \boldsymbol{\lambda}_{-ab}, \boldsymbol{s}, \mathbf{n}, \mathbf{X}) &\propto p(\lambda_{ab})p(n_{ab}|\lambda_{ab}, \xi_a, \xi_b) \\ &\propto \exp(-\lambda_{cd}\xi_c\xi_d)(\lambda_{cd}\xi_c\xi_d)^{n_{cd}}\lambda_{cd}^{\alpha_{cd}-1}\exp(-\beta\lambda_{cd}) \\ &\propto \exp(-\lambda_{cd}(\xi_c\xi_d+\beta))\lambda_{cd}^{n_{cd}+\alpha_{cd}-1} \\ &\sim \text{Gamma}(n_{cd}+\alpha_{cd}, \xi_c\xi_d+\beta). \end{aligned} \tag{S12}$$

Each λ_{ab} is updated in turn by sampling from the right hand side.

S3. Simulation experiments

We here describe several simulation experiments that we implemented to validate the hierarchical Poisson model (10) to estimate transmission flows in the context of sampling heterogeneity. Our overall strategy was to simulate transmission counts, re-estimate transmission flows. and compare the estimated flows with the true simulated flows.

S3.1. ODE-based simulations experiments

The first experiment was a minimal example to assess (1) the impact of sampling differences on flow estimates and (2) the basic performance of our approach. The main results were reported in Figure 3B. We here provide further detail on these simulation experiments.

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). The population was further structured by men and women, yielding in total 4 population strata. We then simulated epidemics based on the compartmental model (4) among susceptible, infected, and treated individuals in the four population strata. The parameters of the model were specified such that 40% of were in group a and 60% of were in group; half of individuals in each group were men and women respectively; and HIV prevalence was approximately 60% at equilibrium. Stochastic simulations in a population of 30000 individuals were performed using MASTER (Vaughan and Drummond, 2013), and run for 400 time units. The simulations are specified in the xml file `https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/inst/misc/Master.xml`. Table S2 lists the model parameters that we used in the simulations, where in terms of notation we replaced the population subscripts l with a and h with b compared to (4).

To assess the impact of sampling heterogeneity on flow estimates, we kept the sampling probability in population group a at 60%, and varied the sampling probability in group

Table S2. Parameter values used for the ODE-based simulations

Parameter	Symbol	Value
transmission rates female to male (further divided by pop size)		
a->a	$\beta_{fm,aa}$	0.0713
b->a	$\beta_{fm,ba}$	0.0071
a->b	$\beta_{fm,ab}$	0.0122
b->b	$\beta_{fm,bb}$	0.0713
transmission rates male to female (further divided by pop size)		
a->a	$\beta_{mf,aa}$	0.1019
b->a	$\beta_{mf,ba}$	0.0173
a->b	$\beta_{mf,ab}$	0.0224
b->b	$\beta_{mf,bb}$	0.1019
viral suppression rate	γ	0.0444
birth/death rate	μ	0.01667

b from 60% to 35% in order to assess the impact of sampling differences of 0%, 5%, 10%, 15%, 20%, 25%. The sampling status of each individual in a was simulated under a Bernoulli draw with probability ξ_a, and respectively for each individual in b with probability ξ_b, and observed transmissions n^r_{ab} from group a to group b were calculated as in (S15).

From the simulated data, we first re-estimated the transmission flows while ignoring sampling heterogeneity. We estimated the posterior distribution (12) with the phyloflows MCMC-within-Gibbs sampler of section S2.3, with $p(\xi_a|\boldsymbol{s}, \mathbf{X})$ and $p(\xi_b|\boldsymbol{s}, \mathbf{X})$ set to $\text{Beta}(25.5, 25.5)$. Thus, the specified sampling densities ignored sampling heterogeneity. Code is available at `https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/inst/misc/ode_estimation.R`. For each simulation, we then calculated the worst case error. The main results are shown in Figure 3B, and indicate that worst case error increased substantially with sampling heterogeneity.

Second, we re-estimated the transmission flows while accounting for sampling heterogeneity. The posterior distribution of sampling probabilities $p(\xi_a|\boldsymbol{s}, \mathbf{X})$ was estimated from the number of sampled individuals in the populations N^s_a, assuming we know the total population size N_a, through

$$p(\xi_a|\boldsymbol{s}, \mathbf{X}) = \text{Beta}(\xi_a, N^s_a + \alpha_\xi, N_a - N^s_a + \beta_\xi), \tag{S13}$$

where we set $\alpha_\xi = 0.5$, $\beta_\xi = 0.5$, and similarly for ξ_a. Figure S1 illustrates the estimated posterior sampling distributions, in comparison to the true values. We then estimated the posterior distribution (12) with the phyloflows MCMC-within-Gibbs sampler of section S2.3, using the posterior distribution of sampling probabilities (S13). Figure 3B reports

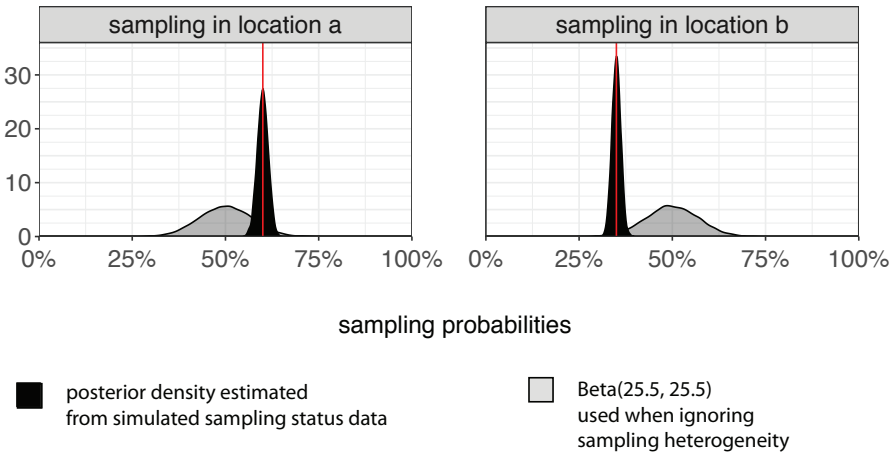


Fig. S1. Re-estimated sampling probabilities in the ODE-based simulations. We simulated transmission flows between men and women in two locations a and b under the ODE model (4) in Section S3.1. From the simulated sampling status data, posterior estimates of the sampling proportions in a and b were obtained with Equation (S13), and then used in flow inference. We show posterior densities of the sampling probabilities in a and b (grey), compared to a $Beta(25.5, 25.5)$ prior density. Vertical lines indicate the true sampling fractions, $\xi_a = 0.6$ and $\xi_b = 0.35$.

the worst case error of posterior flow estimates.

S3.2. Sensitivity to overall sample size of observed transmission flows

We further assessed the accuracy of flow inferences as a function of overall sample size of the observed transmission flows, $n^+ = \sum_{a,b} n_{ab}$. We re-considered the ODE-based simulation experiments of Section S3.1, with a 25% sampling difference between a and b locations. The overall population size parameter was set to 11200, 18000, 52000 such that the overall sample size was $n^+ \approx 100$, 300, 600.

Figure S2 (grey bars) shows posterior estimates of the transmission flows on 4 randomly selected simulated data sets, for the case that 60% of the population in a and 35% of the population in b were sampled, and sampling heterogeneity was ignored. Figure S2 (orange bars) shows posterior estimates of the same transmission flows on the same 4 randomly selected data sets, when sampling heterogeneity was accounted for as in (S13). There was considerable variability in the flow data sets when sample size was low ($n^+ = 100$). However, information on population sampling substantially improved flow estimates regardless of sample size. Figure S3 summarises the worst case error in these simulations. When sampling heterogeneity was not accounted for, variability in flow estimates decreased with increasing sampling size, but error magnitude did not de-

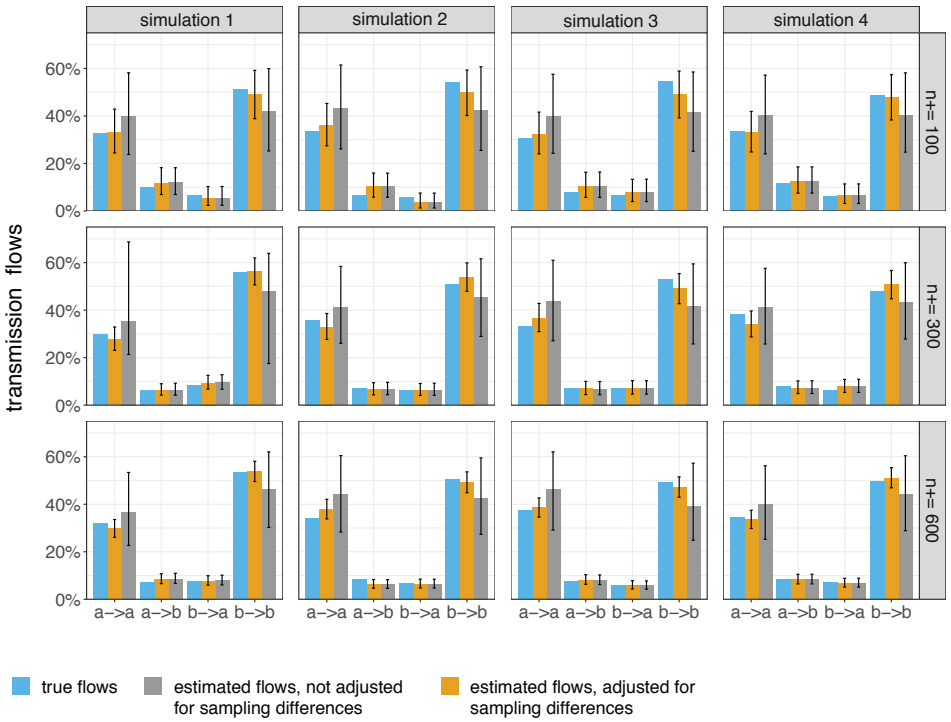


Fig. S2. True and re-estimated flows in four ODE-based simulations, varying in sample size. We simulated transmission flows between men and women in two locations a and b under the ODE model (4) in Section S3.2. Columns show results for four simulated data sets, and rows show results for increasing samples size. The true transmission flows are show in blue. Posterior flow estimates (median and 95% credibility intervals) are shown when sampling heterogeneity was ignored (grey), and when sampling heterogeneity was accounted for as in (S13) (orange).

crease. When sampling heterogeneity was accounted for, both the variability in flow estimates and error magnitude decreased with increasing sampling size.

S3.3. More complex simulation experiments

The second experiment mimicked the more complex population structure and sampling heterogeneity as observed in the Rakai case study and described in Section S4.

We considered simulated populations stratified into 24 population sub-groups, by gender (male, female), location (inland,fishing), in-migration status (in-migrant, resident), age (15-24, 25-34, 35-50 years), and simulated 576 transmission flows between the male-female and female-male sub-group combinations. For simplicity, we did not generate epidemic trajectories and instead simulated transmission counts n_{ab} from pre-specified transmission flows $\boldsymbol{\pi}^0$, and pre-specified sampling probabilities ξ_a. For a fixed

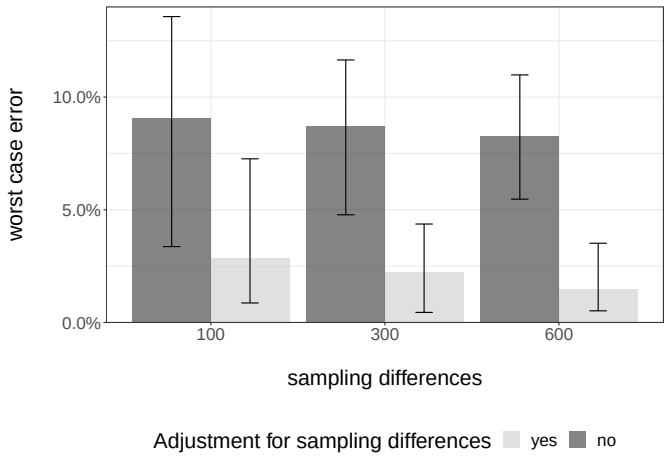


Fig. S3. Worst case error in estimated transmission flows with increasing sample size. We simulated transmission flows between men and women in two locations a and b under the ODE model (4) as described in Section S3.1. Transmission flows were re-estimated with the hierarchical Poisson model (10) on simulated data sets that varied in sample size $n^+ = 100, 300, 600$ (x-axis). In the first set of inference runs (dark grey), sampling differences across locations a, b were not adjusted for, and in the second set of inferences (light grey), sampling differences were accounted for based on counts of sampled and infected individuals. The worst case error between the true transmission flows and median posterior estimates was calculated in each scenario, and the median (bar) and 2.5% and 97.5% quantiles (error bars) are shown. Errors decreased with increasing sample size, although errors remained very large when sampling differences were not accounted for.

target sample size n^+, the total number of actual transmissions z^+ were simulated from $z^+ \sim \text{Poisson}(n^+/\overline{\xi})$, where $\overline{\xi} = \frac{1}{24}\sum_a \xi_a$ was the average sampling probability in the population. Then, the actual transmission counts between population groups were simulated from $\mathbf{z} \sim \text{Multinomial}(Z, \boldsymbol{\pi_0})$. The sampling status of each individual in a was simulated under a Bernoulli draw with probability ξ_a, and respectively in population b with probability ξ_b, and observed transmissions n_{ab} from group a to group b were calculated as in (S15). 100 such simulated data sets were generated, for a target sample size $n^+ = 300$.

The difference between sampling proportions in inland and fishing communities was about 10%. We transformed the sampling proportions with a sine function to obtain a greater range of average sampling differences between inland and fishing communities from 5% to 25%. Additional simulations were then generated under these sampling scenarios.

From the simulated data, we re-estimated the 576-dimensional transmission flows while ignoring sampling heterogeneity, and then while accounting for sampling hetero-

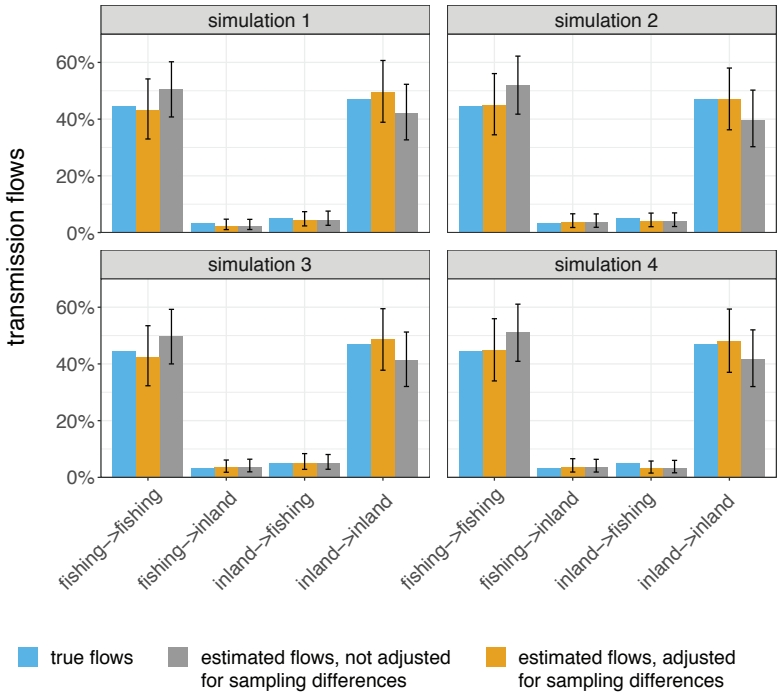


Fig. S4. True and re-estimated flows in more complex simulations between 24 population groups. We simulated transmission flows by gender, 3 age bands, and location under model parameters similar to those obtained for the Rakai analysis; see Section S3.3. Panels show results for four simulated data sets. Posterior flow estimates (median and 95% credibility intervals) are shown when sampling heterogeneity was ignored (grey), and when sampling heterogeneity was accounted for as in (S13) (orange).

geneity as described in Section S3.1. To facilitate comparison, we aggregated the true and re-estimated flows into the 4-dimensional vector of flows within and between inland and fishing communities, and calculated worst case errors between them. Figure S4 shows posterior estimates of the transmission flows on 4 randomly selected simulated data sets, for the baseline case that the difference in sampling probabilities between inland and fishing communities was 10%. Figure S5 shows that in these more complex simulation scenarios, trends in worst case error were overall similar compared to the more simple ODE-based simulation experiments in Section S3.1. When ignoring sampling heterogeneity, worst case error increased considerably with average sampling differences between inland and fishing communities. When accounting for sampling heterogeneity, worst case error remained largely unaffected by increasing average sampling differences, and was overall slightly higher compared to that in the ODE-based simulation experiments.

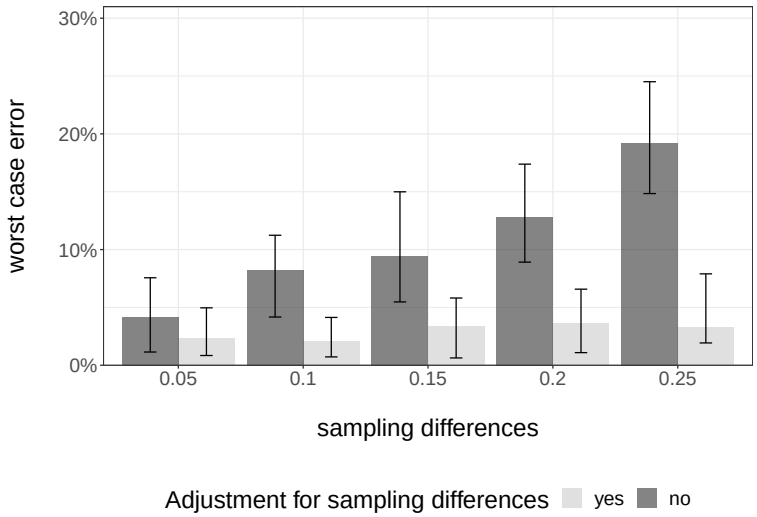


Fig. S5. Worst case error in estimated transmission flows in more complex simulations of transmission flow between 24 population groups. We simulated transmission flows between men and women in 24 population groups as described in Section S3.3. Simulation scenarios varied in average sampling differences between simulated inland and fishing communities (x-axis). Inferences were performed while ignoring for sampling heterogeneity (dark grey), and while accounting for sampling heterogeneity (light grey). In each simulation, worst case error was reported, after aggregating flow estimates to the 4 flow combination within and between inland and fishing combinations for comparison with the ODE-based simulations, and its median and 2.5% and 97.5% quantiles are shown (y-axis). Overall, trends in worst case error were similar those found in the more simple ODE-based simulation experiments.

S3.4. Accuracy of HSGP approximation

We finally assessed the accuracy of HSGP approximation in the context for heterogenous sampling. The main results are summarised in Section 3.2. This supplementary material provides more details of this simulation experiment and visualises the performance of the final model specification. We considered transmission flows between one-year-increment age groups between 15 and 24 by gender and locations and simulated transmission in-

Table S3. Parameter values used to simulate transmission flows in populations stratified by 1-year age bands under the extension of GP model (15).

Parameter	Symbol	Value
intercept		
female to male		
h->h	$\mu_{fm,hh}$	-1
h->l	$\mu_{fm,hl}$	-10
l->h	$\mu_{fm,lh}$	-9
l->l	$\mu_{fm,ll}$	-2.5
intercept		
male to female		
h->h	$\mu_{mf,hh}$	-0.5
h->l	$\mu_{mf,hl}$	-9
l->h	$\mu_{mf,lh}$	-9
l->l	$\mu_{mf,ll}$	-1
lengthscale		
female to male	$\boldsymbol{\ell_{fm}}$	(4.1,2.3)
male to female	$\boldsymbol{\ell_{mf}}$	(2.3,4.6)
marginal standard deviation		
female to male	σ_{fm}	1.8
male to female	σ_{mf}	1.5

tensities and counts via a slight extension of Equation (15),

$$
\begin{aligned}
&\boldsymbol{n} \sim \text{Poisson}(\boldsymbol{\lambda}\boldsymbol{\xi}^T\boldsymbol{\xi}^R)\\
&\log\boldsymbol{\lambda} = \mu_{mf,hh}\mathbb{1}_{mf,hh} + \mu_{mf,hl}\mathbb{1}_{mf,hl} + \mu_{mf,lh}\mathbb{1}_{mf,lh} + \mu_{mf,ll}\mathbb{1}_{mf,ll}\\
&\quad + \mu_{fm,hh}\mathbb{1}_{fm,hh} + \mu_{fm,hl}\mathbb{1}_{fm,hl} + \mu_{fm,lh}\mathbb{1}_{fm,lh} + \mu_{fm,ll}\mathbb{1}_{fm,ll} + \boldsymbol{f},\\
&\boldsymbol{f} = (\boldsymbol{f}_{mf}^T, \boldsymbol{f}_{fm}^T)^T,\\
&\boldsymbol{f}_{mf} \sim \mathcal{GP}(0, k_{mf}), \quad \boldsymbol{f}_{fm} \sim \mathcal{GP}(0, k_{fm}),\\
&k_{mf}\big((a_1,b_1),(a_2,b_2)\big) = \sigma_{mf}^2 \exp\Big(-\Big[\frac{(a_2-a_1)^2}{2\ell_{mf,a}^2} + \frac{(b_2-b_1)^2}{2\ell_{mf,b}^2}\Big]\Big)\\
&k_{fm}\big((a_1,b_1),(a_2,b_2)\big) = \sigma_{fm}^2 \exp\Big(-\Big[\frac{(a_2-a_1)^2}{2\ell_{fm,a}^2} + \frac{(b_2-b_1)^2}{2\ell_{fm,b}^2}\Big]\Big),
\end{aligned}
\tag{S14}
$$

under the hyper-parameter values in Table S3.

This extension allows to capture average transmission intensities by locations and genders and varying age-dependent transmission dynamics from male to female and from female to male. Meanwhile, samples from models for sampling probability (Supplementary Material S4) were reused and their means were taken as ground-truth sampling intensities. 20 replicates of transmission intensities and counts were generated.

Flows and hyperparameters were re-estimated using both Gaussian Process model and Hilbert Space Gaussian Process model based on observed counts in the simulation,

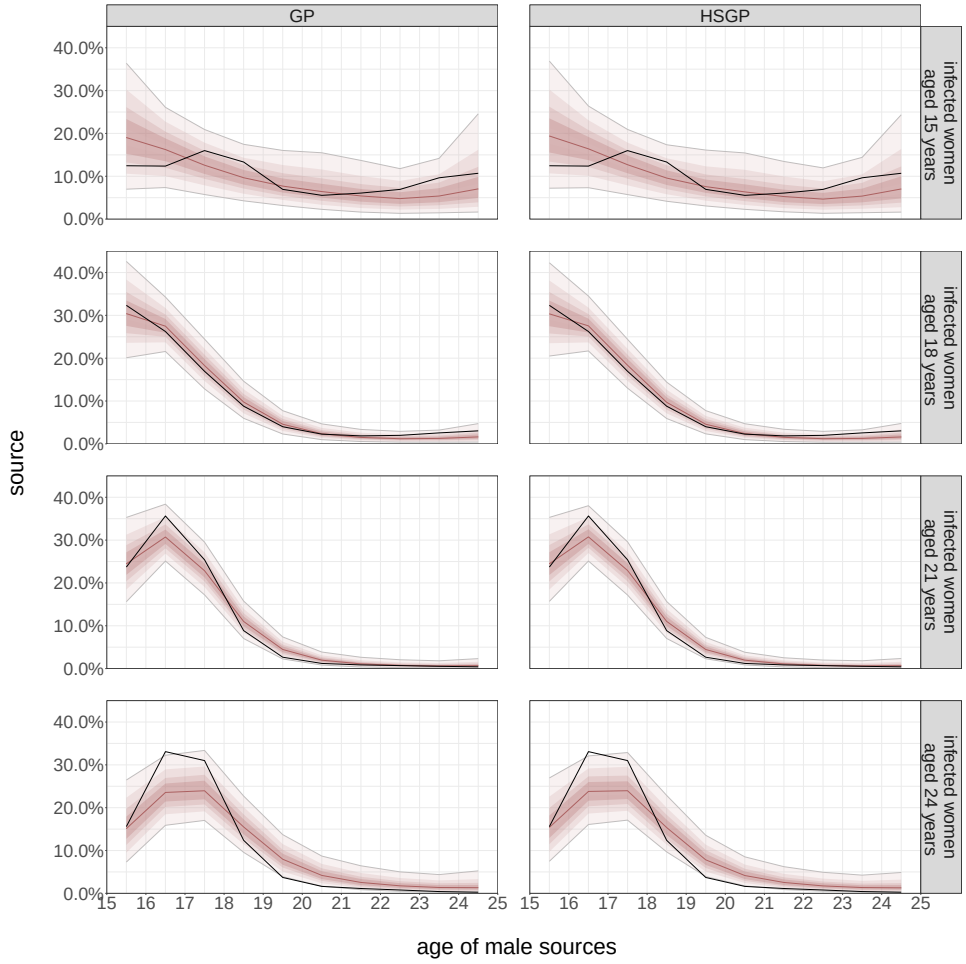


Fig. S6. Estimated sources of infection in women under the final HSGP approximation. We simulated age-specific transmission flows under the kernel parameters shown in Table S3, and re-estimated transmission flows along with all other parameters under HSGP approximations to (15). The subfigures on the left illustrate the estimated sources of infection in women under the GP prior from which the data were simulated. The subfigures on the right illustrate the estimated sources under the final HSGP approximation. Shades illustrate the estimated posterior credibility intervals, showing 40% credibility intervals in dark red to 80% credibility intervals in light red. Posterior medians are shown as a red line. True values are shown in black. We found no systematic differences in source estimates when using the final HSGP approximation over the GP prior.

under priors

$$\sigma^2_{mf}, \sigma^2_{fm} \sim \text{Half-Normal}(0, 10)$$

$$\ell_{d,i} \sim \text{Inv-Gamma}(\alpha_{d,i}, \beta_{d,i})$$

$$\mu_{mf,hh}, \mu_{mf,hl}, \mu_{mf,lh}, \mu_{mf,ll}, \mu_{fm,hh}, \mu_{fm,hl}, \mu_{fm,lh}, \mu_{fm,ll} \sim \text{Normal}(0, 10).$$

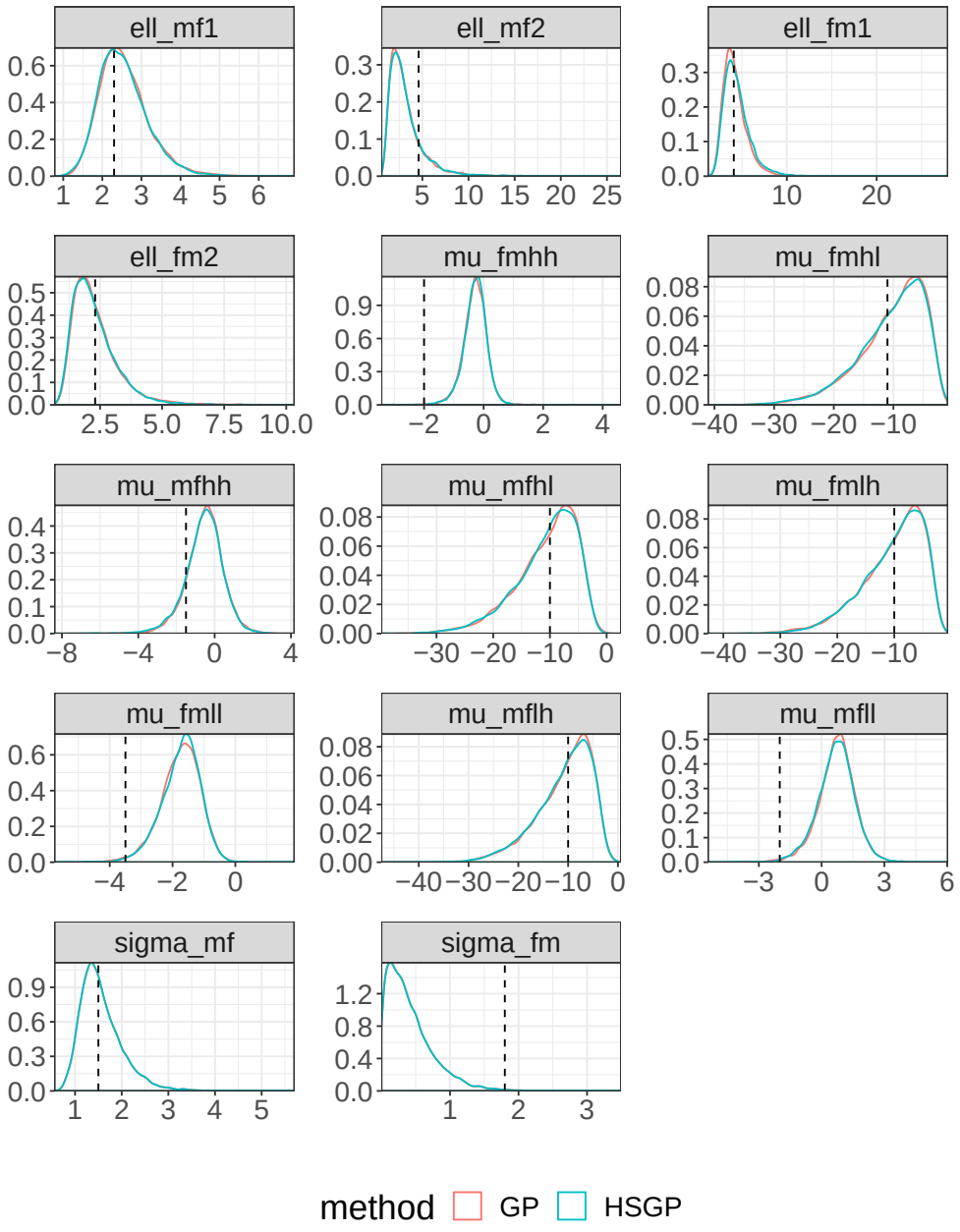


Fig. S7. Estimated covariance kernel parameters under the final HSGP approximation. We simulated age-specific transmission flows under the kernel parameters shown in Table S3, and re-estimated transmission flows along with all other parameters under HSGP approximations to (15). The subfigures illustrate the marginal posterior densities of the kernel parameters for under the GP model (in red) and the HSGP approximation (in blue) based on $B = 1.25, m = 30$ (right panels).

Here, we illustrate further the performance of the HSGP approximation under the final tuning parameters, $B = 1.25, m = 30$. Figure S6 compares the marginal posterior distribution of the sources of infections in women when using the chosen HSGP approximation, compared to using the GP prior. We found no particular differences in the estimated sources of infection when using the HSGP approximation over the GP prior. Figure S7 compare the marginal posterior densities of the kernel parameters between the chosen HSGP approximation and the GP prior. Again, we found no systematic differences in the estimated kernel parameters when using the chosen HSGP approximation compared to the GP prior.

S4. Modelling and estimation of the sampling cascade

We here describe the statistical models used to characterise each step of the sampling cascade of transmission events, shown in Figure 5A.

To recall our setting, we defined in the main text transmission events as HIV infection events to individuals who lived in RCCS communities, were aged 15-49 years, and were infected in the study period $\mathcal{T} = \{2009/10/1, 2015/01/30\}$. The main study objective was to estimate transmission flows and related quantities (1-3) among different population groups, expressed in proportions relative to this denominator of transmission events, denoted by $\mathcal{Z}$. Each transmission event in $\mathcal{Z}$ can be indexed by its source i and recipient j in the population of individuals that were infected by the end of $\mathcal{T}$, and we denote transmission from i to j with the indicator variable $z_{ij} = 1$, and no transmission with $z_{ij} = 0$. Thus, recipients are individuals in $\mathcal{P}$ that were infected during the study period $\mathcal{T}$, i.e. in our case 2009/10/1-2015/01/30. Sources are infected individuals who transmitted to one of the recipients. In our applications, population groups were defined by location (inland and fishing communities), gender, and 1-year age bands from age 15 to age 49, yielding $A = 140$ population strata. The actual transmission events from group a to group b are

$$z_{ab} = \sum_{i \in a, j \in b} z_{ij}, \tag{S15}$$

and observing any such event depends on the sampling status of the source and recipient,

$$n_{ab} = \sum_{i \in a, j \in b} \mathbb{1}\left\{s_i^S = 1\right\} \mathbb{1}\left\{s_j^R = 1\right\} z_{ij}. \tag{S16}$$

To make inference on transmission flows, our strategy is to explicitly account for sampling heterogeneity in order to reduce bias in the flow estimates. We propose estimating sampling proportions and then propagate these estimates into inference of transmission flows. Overall, we assume that sources and recipients are independently sampled, and

conditionally at random within each population sub-group a with probabilities $\boldsymbol{\xi}^S = (\xi_a^S)_{a\in\mathcal{A}}$ and $\boldsymbol{\xi}^R = (\xi_a^R)_{a\in\mathcal{A}}$. For sake of brevity, we derived the flow model in equations (8-12) using the same sampling probabilities for sources and recipients. In our applications we consider its extension to distinct sampling probabilities for sources and recipients, with equations generalising straightforwardly.

We modelled the overall sampling probabilities ξ_a^S, ξ_a^R as the product of conditional sampling probabilities along the sampling cascade of Figure 5A, see the forthcoming Equations (S20) and (S23).

The sampling cascade involved that sources and recipients participate in the cohort, report no ART use, and that virus was successfully deep-sequenced. Section S4.1 describes estimation of participation probabilities for source cases, and Section S4.2 describes estimation of sequence sampling probabilities conditional on participation among source cases. Section S4.3 describes estimation of participation probabilities for recipient cases, and Section S4.4 describes estimation of sequence sampling probabilities conditional on participation among recipient cases. Section S4.5 illustrates the resulting sampling probabilities of transmission events.

In each step of the cascade, we compared the prediction accuracy of candidate models through the ten-fold cross validation. To do it, the data $\mathbf{y} \in \mathbb{R}^N$ is divided into 10 sets, the kth test set is denoted as $\mathbf{y}_k$ and the rest of observations forms the training set $\mathbf{y}_{-k}$. Here we defined accuracy measures.

- Hold-out data in 95% PPCrI (Posterior Predictive Credibility Interval): we simulated data (sampled individuals) from the fitted model on $\mathbf{y}_{-k}$, found the 95% credibility intervals of the simulated data, and counted the percentage of observed data $\mathbf{y}_k$ included in the credibility interval.
- MAE (Mean Absolute Error): we simulated data (sampled individuals) from the fitted model on $\mathbf{y}_{-k}$, found the median estimates, and calculated the absolute differences of the observations $\mathbf{y}_k$ and the median estimates.
- ELPD (expected log point-wise predictive density for a new dataset): this evaluates the log predictive density at each data point in the kth fold,

$$\text{ELPD}_k = \sum_{i=1}^{N_k} \text{ELPD}_{ki} = \sum_{i=1}^{N_k} \log(\frac{1}{S}\sum_{s=1}^{S} p(\mathbf{y}_{ki}|\theta^{k,s})), \tag{S17}$$

  where N_k is the number of data in the kth fold, $\mathbf{y}_{ki}$ denotes the ith observation in the kth test set, and $\theta^{k,s}$ denote the sth samples drawn from $p(\theta|\mathbf{y}_k)$.

S4.1. Modelling and estimating participation probabilities of source cases

We are interested in participation probabilities among infected individuals who transmitted virus. In general, HIV infections in surveyed locations may have their source outside these locations, with previous work suggesting that approximately 30% of infections in RCCS communities originate from outside the cohort (Grabowski et al., 2014). Sources from outside the RCCS communities by definition did not participate in any survey round, and in the absence of any data we focused on sources from RCCS communities. This is a general limitation of our flow inferences.

As part of demographic surveillance, the RCCS conducts a population census immediately prior to each survey round in the cohort communities, which provides an enumeration of the underlying age-eligible population of 15-49 years. We used these data to estimate participation probabilities ξ_a^p over the survey period in each of the underlying population sub-groups a, and thereby approximated the participation probabilities of sources cases. This involved three assumptions. First, we assumed that individuals participated regardless of infection status, and second regardless of being a source or not, and third that participation probabilities were the same in the survey period (2011/08/10-2015/01/30) and in the slightly longer study period (2009/10/01-2015/01/30), during which transmission flows were estimated.

We used a Beta-Binomial regression framework to estimate the posterior distribution of population participation probabilities $\boldsymbol{\xi}^p = (\xi_a^p)_{a\in\mathcal{A}}$. Denote by N_a^e the number of participation-eligible individuals in population group a, by N_a^p the number of participants in population group a, by $\mathbf{X}_a \in \{0,1\}^p$ a row vector of p predictor variables associated with population a, including location status, gender, age band, and possible interaction terms, and by $\mathbf{X} \in \{0,1\}^{A\times p}$ the design matrix that stacks the row vectors $\mathbf{X}_a$ for each population group a. The general form of the regression models considered was

$$\begin{aligned} N_a^p &\sim \text{Beta-Binomial}(N_a^e, \xi_a^p, \gamma),\ \forall a \\ \text{logit}(\boldsymbol{\xi^p}) &= \beta_0\mathbf{1} + \boldsymbol{\beta X} \\ \beta_0 &\sim \mathcal{N}(0, 100) \\ \boldsymbol{\beta} &\sim \mathcal{N}(0, 10\boldsymbol{I}) \\ \gamma &\sim \text{Exp}(1) \end{aligned} \tag{S18}$$

where the Beta-Binomial is parameterised in terms of the number of Bernoulli trials, the mean of the marginal probabilities of successes and the dispersion parameter, $\beta_0 \in \mathbb{R}$ is the baseline participation probability, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients, and $\boldsymbol{I}$ is the $p \times p$ identity matrix. Four models were considered and assessed using 10-fold cross-validation.

Table S4. Prediction accuracy of models P1-P4 on hold-out population participation data in 10-fold cross-validation.

Model	Hold-out data in 95% PPCrI	MAE	ELPD
	median (range)	median (range)	median (std dev)
P1	93.1% (91.5% - 95.1%)	1.79 (1.50 - 1.92)	-3705.27 (26.78)
P2	98.0% (97.2% - 99.2%)	1.76 (1.49 - 1.98)	-4168.58 (37.05)
P3	93.1% (92.4% - 95.3%)	1.67 (1.58 - 1.90)	-3651.39 (27.61)
P4	98.4% (96.5% - 98.8%)	1.70 (1.53 - 1.90)	-4072.42 (37.21)

PPCrI is posterior predictive credibility interval
MAE is abbreviated from mean absolute error
ELPD is the abbreviation of expected log predictive density.

- Model P1. Model (S18) with indicator variables on location, gender, and 1-year age bands as contrasts, leading to $p = 36$ predictors. The corresponding regression coefficients were given independent normal prior densities, and the dispersion coefficient was set to $\gamma = 0$ to reflect no overdispersion.
- Model P2. As model P1 but with γ estimated to allow for overdispersion.
- Model P3. Model (S18) with indicator variables on location, and gender and 1-year age band interactions as contrasts, leading to $p = 70$ predictors. The corresponding regression coefficients were given independent normal prior densities, and the dispersion coefficient was set to $\gamma = 0$ to reflect no overdispersion.
- Model P4. As model P3 but with γ estimated to allow for overdispersion.

The models were fitted in Stan version 2.21, convergence diagnostics (Vehtari et al., 2019) were checked, and effective sample sizes were $\gg 10000$. Table S4 shows the proportion of hold-out data in 95% posterior predictive credibility intervals (PPCrI), the mean absolute error (MAE), and the expected log predictive density in 10-fold cross-validation (ELPD) for each model. Based on these statistics, model P4 was chosen to model participation probabilities across population strata. Figure S8 summarises the marginal posterior densities of the regression coefficients under model P4. Figure S9 shows the marginal posterior participation probabilities by location, gender, and age-bands.

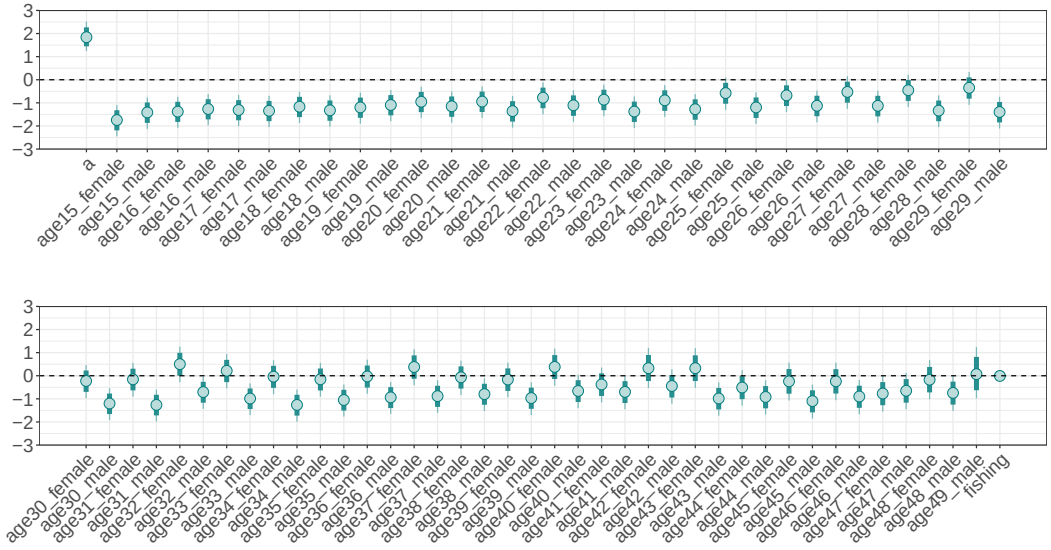


Fig. S8. Marginal posterior densities of regression coefficients of model P4 to estimate population participation probabilities. Circle: median; bar: interquartile range; line: 95% credibility interval. The first variable a refers to the intercept.

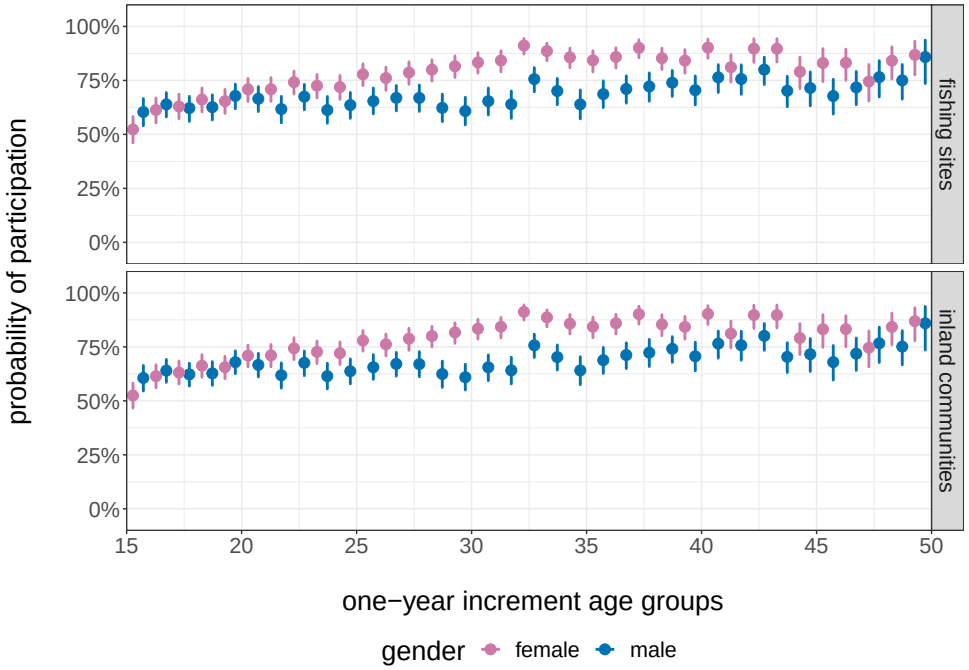


Fig. S9. Marginal posterior densities of population participation probabilities under model P4. Circle: median; line: 95% credibility interval.

S4.2. Modelling and estimating sequencing probabilities of source cases

We are now interested in the probability that (participating) source cases report no ART use, and the probability that source cases and reported no ART use also had virus successfully deep-sequenced. We combined both steps and sought to estimate the probability that source cases had virus successfully deep-sequenced. Considerable demographic and clinical data were collected among RCCS participants, which allowed us to relax the assumptions of Section S4.1 when modelling and estimating sequencing probabilities of source cases. In general, many infected individuals do not transmit the virus onwards, and individuals who transmitted virus in the study period $\mathcal{T}$ must have had unsuppressed HIV at some point in $\mathcal{T}$. A routinely collected proxy variable of viral suppression is self-reported ART use (Grabowski et al., 2018). On this basis, we excluded from consideration of potential source cases infected individuals who consistently reported ART use, defined as reporting ART use at each survey visit, and approximated the probability of obtaining a deep-sequence sample from source cases who participated in the cohort by the probability of obtaining a deep-sequence sample from infected participants who reported no ART use on at least one visit during the survey period. In this approximation, we further assumed that individuals who consistently reported ART use during the survey period (2011/08/10-2015/01/30) would also have consistently reported ART use during the study period (2009/10/01-2015/01/30).

We again used a Beta-Binomial regression framework to estimate the posterior distribution of deep-sequencing probabilities $\boldsymbol{\xi}^s = (\xi_a^s)_{a\in\mathcal{A}}$, where ξ_a^s denotes the probability of obtaining a deep-sequence sample from potential source cases, defined as infected participants who reported no ART use on at least one visit during the survey period in population sub-group a. However due to smaller numbers we also considered regularised regression models with correlations imposed among the regression coefficients. We follow

Table S5. Prediction accuracy of models S1-S8 on hold-out deep-sequencing data in 10-fold cross-validation.

Model	**Hold-out data in 95% PPCrI**	**MAE**	**ELPD**
	median (range)	median (range)	median (std dev)
S1	99.2% (98.6% - 100.0%)	0.60 (0.55 - 0.65)	-1082.79 (12.99)
S2	100% (99.2% - 100%)	0.58 (0.54 - 0.65)	-1130.18 (15.91)
S3	99.2% (98.4% - 100.0%)	0.60 (0.53 - 0.70)	-1057.71 (13.1)
S4	99.2% (99.2% - 100.0%)	0.61 (0.56 - 0.70)	-1097.95 (15.9)
S5	99.2% (99.2% - 100%)	0.58 (0.56 - 0.61)	-1106.56 (12.79)
S6	100% (99.2% - 100%)	0.57 (0.55 - 0.62)	-1159.65 (15.77)
S7	99.2% (98.4% - 100%)	0.58 (0.56 - 0.61)	-1118.38 (12.58)
S8	100% (99.2% - 100%)	0.58 (0.55 - 0.61)	-1174.11 (15.6)

PPCrI is posterior predictive credibility interval
MAE is abbreviated from mean absolute error ELPD is the abbreviation of expected log predictive density.

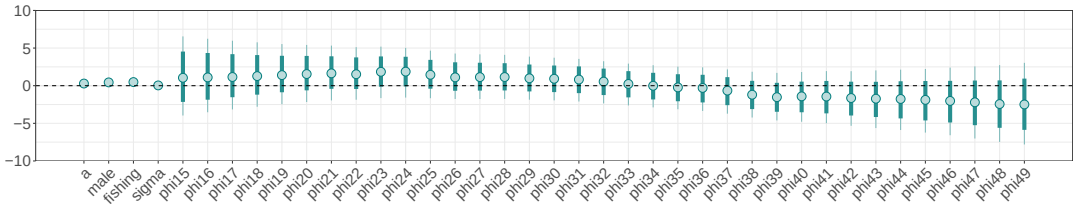


Fig. S10. Marginal posterior densities of regression coefficients of model S5 to estimate conditional deep-sequencing probabilities. Circle: median; bar: interquartile range; line: 95% credibility interval. The first variable a refers to the intercept.

the notation of (S18) and further denote by $N_a^{\text{naïve}}$ the number of infected participants in a who did not report consistent ART use, and by N_a^s the number of infected participants in a who did not report consistent ART use and had virus deep-sequenced successfully. We considered models of the form

$$\begin{aligned} N_a^s &\sim \text{Beta-Binomial}(N_a^{\text{naïve}}, \xi_a^s, \gamma),\ \forall a \\ \text{logit}(\boldsymbol{\xi^s}) &= \beta_0 \mathbf{1} + \boldsymbol{\beta X} \\ \beta_0 &\sim \mathcal{N}(0, 100) \\ \boldsymbol{\beta} &\sim \mathcal{N}(0, 10\boldsymbol{I}) \\ \gamma &\sim \text{Exp}(1), \end{aligned} \tag{S19}$$

and similar models with 1st-order intrinsic conditional auto-regressive (ICAR) prior distributions across adjacent age bands (Rue and Held, 2005). Eight models were considered and assessed using 10-fold cross-validation.

- Model S1. Model (S19) with indicator variables on location (1), gender (1), and 1-year age bands (34) as contrasts, leading to $p = 36$ predictors. The corresponding regression coefficients were given independent normal prior densities, and the dispersion coefficient was set to $\gamma = 0$ to reflect no overdispersion.
- Model S2. As model S1 but with γ estimated to allow for overdispersion.
- Model S3. Model (S19) with indicator variables on location (1), and gender and 1-year age band interactions (69) as contrasts, leading to $p = 70$ predictors. The corresponding regression coefficients were given in independent normal prior densities, and the dispersion coefficient was set to $\gamma = 0$ to reflect no overdispersion.
- Model S4. As model S3 but with γ estimated to allow for overdispersion.
- Models S5-S8. As models S1-S4 respectively, but using an ICAR prior density on adjacent age bands.

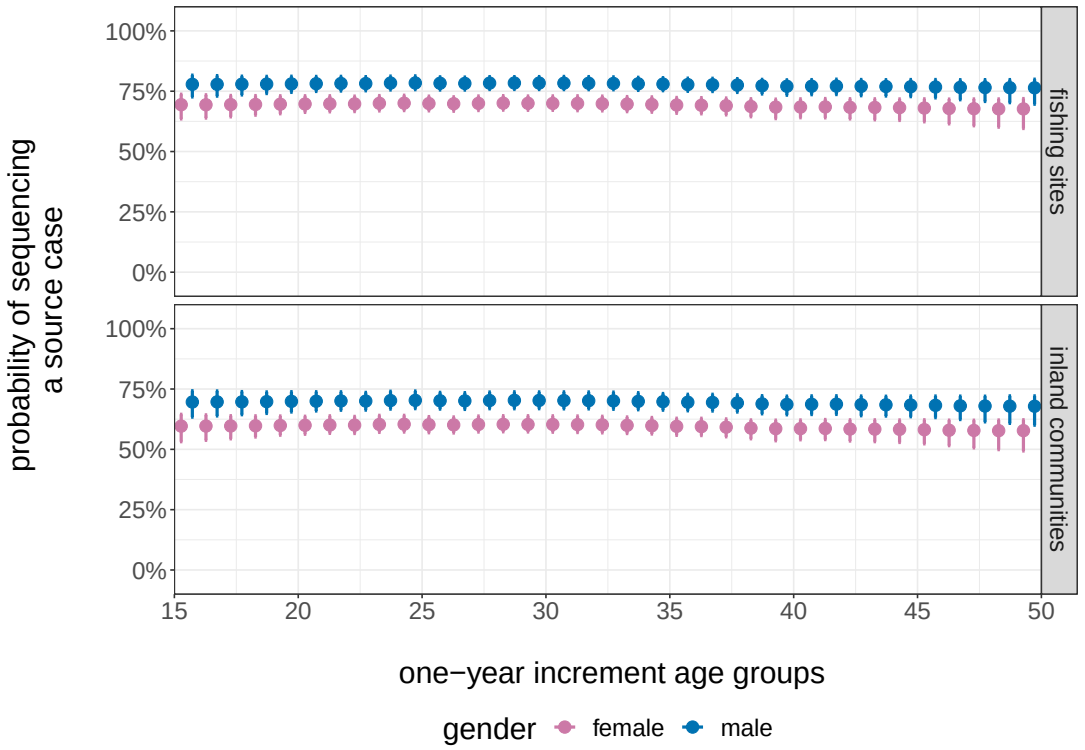


Fig. S11. Marginal posterior densities of source sequencing probabilities under model S5. Circle: median; line: 95% credibility interval.

The models were fitted in Stan version 2.21, convergence diagnostics (Vehtari et al., 2019) were checked, and effective sample sizes were again $\gg$ 10000. In Table S5 we report as before the proportion of hold-out data in 95% PPCrI, the MAE, and the ELPD for each model in 10-fold cross-validations. Based on these figures, model S5 was chosen to model deep-sequencing probabilities (conditional on participation) across population strata. Figure S8 shows the marginal posterior densities of the regression coefficients under model S5. Figure S11 summarises the marginal posterior deep-sequencing probabilities by location, gender, and age-bands.

To summarise, the sampling probability of source cases was approximated by

$$\xi_a^S = \xi_a^p * \xi_a^s \tag{S20}$$

for all population sub-groups a.

S4.3. Modelling and estimating participation probabilities of recipient cases

We are now interested in participation probabilities among recipients, i.e. individuals who acquired infection in the study period $\mathcal{T}$, 2009/10/01-2015/01/30. We proceeded

Table S6. Prediction accuracy of models N1-N3 on hold-out infection time data in 10-fold cross-validation.

Model	**Hold-out data in 95% PPCrI** median (range)	**MAE** median (range)	**ELPD** median (std dev)
N1	87.3% (85.0% - 88.9%)	0.1458 (0.1364 - 0.1830)	-537 (10.1)
N2	97.2% (95.6% - 98.5%)	0.0284 (0.0152 - 0.0467)	-103 (3.61)
N3	97.3% (95.6% - 98.5%)	0.0303 (0.0152 - 0.0438)	-101 (3.56)

PPCrI is posterior predictive credibility interval
MAE is abbreviated from mean absolute error
ELPD is the abbreviation of expected log predictive density.

as in Section S4.1 and used the RCCS census data to approximate the participation probabilities among individuals who acquired HIV in $\mathcal{T}$ by participation probabilities over the survey period in each of the underlying population sub-groups. This involved two assumptions. First, we assumed that individuals participated regardless of infection status, and second that participation probabilities were the same in the survey period (2011/08/10-2015/01/30) and in the slightly longer study period (2009/10/01-2015/01/30), during which transmission flows were estimated. Following Section S4.1, we adjusted for heterogeneity in participation probabilities among newly infected individuals using the posterior estimates under model P4, reported in Figure S9.

S4.4. Modelling and estimating sequencing probabilities of recipient cases

We are finally interested in the probability that newly infected participants report no ART use, and the probability that newly infected participants reporting no ART use had virus successfully deep-sequenced. As in Section S4.2, we combined both steps, and sought to estimate the probability that newly infected participants had virus successfully deep-sequenced.

Based on available data collected among RCCS participants, we could again relax the assumptions of Section S4.3. Our approach here differed in two aspects from the approach we employed to estimate sequencing probabilities among source cases in Section S4.2. First, we predicted who of the infected participants acquired infection during the study period $\mathcal{T}$, as opposed to before the start of $\mathcal{T}$. We used longitudinal individual-level serostatus data and phylogenetic estimates of the infection time for this step (Grabowski et al., 2017; Golubchik et al., 2017). Second, we estimated the probability that infected participants who were classified to have acquired HIV in $\mathcal{T}$ (regardless of ART use, in contrast to Section S4.2) also had virus successfully deep-sequenced.

We now describe step 1. Of 5032 infected participants, 883 (17.5%) had a positive HIV test before the start of $\mathcal{T}$, 65 (1.3%) had a phylogenetically estimated mean time of infection before $\mathcal{T}$, 1698 (33.8%) had a phylogenetically estimated time of infection in $\mathcal{T}$.

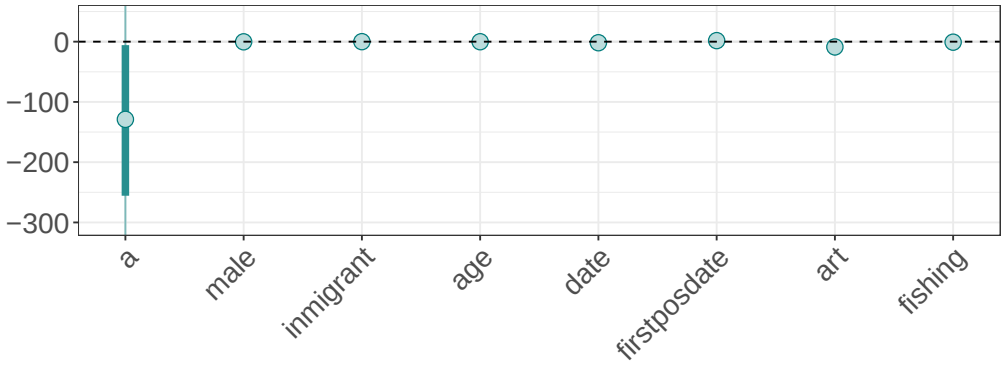


Fig. S12. Marginal posterior densities of regression coefficients of model N3 to estimate conditional deep-sequencing probabilities. Circle: median; bar: interquartile range; line: 95% credibility interval. The first variable a refers to the intercept.

Information on last negative test was not available for this analysis. The remaining 2,386 (47.4%) had no data or phylogenetic estimates on the likely time of infection. Based on the 2646 individuals with data on infection times, we trained Bernoulli regression models to estimate predictor variables associated with HIV acquisition in $\mathcal{T}$ ($Y_i = 1$) versus before $\mathcal{T}$ ($Y_i = 0$). The predictive accuracy of several models was compared using 10-fold cross-validation. Using the best model, we predicted the timing of HIV acquisition among the 2,386 infected participants with no such data. We then retained for further analysis the individuals with evidence of HIV acquisition during $\mathcal{T}$, plus those individuals for whom HIV acquisition during $\mathcal{T}$ was predicted.

The Bernoulli regression models had the form

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\eta_i),\ \forall a \\ \text{logit}(\boldsymbol{\eta}) &= \beta_0 \mathbf{1} + \boldsymbol{\beta} \boldsymbol{X} \\ \beta_0 &\sim \mathcal{N}(0, 100) \\ \boldsymbol{\beta} &\sim \mathcal{N}(0, 10\boldsymbol{I}) \end{aligned} \tag{S21}$$

where η_i is the probability that the ith infected participant acquired HIV in $\mathcal{T}$, and the design matrix was based on the covariates gender (indicator variable), first survey visit date (real valued), age of individual at first visit date (real valued), in-migration into the current RCCS community of residence in the two years prior to the first survey visit (indicator variable), date of first HIV positive test (real valued), residence in fishing or inland communities at time of first positive test (indicator variable), and ART use at first visit date when infected (indicator). Three models were considered.

- Model N1. Model (S21) with the aforementioned covariates except ART status and time of first positive test, leading to $p = 5$ predictors.
- Model N2. Model (S21) with the aforementioned covariates except ART status, leading to $p = 6$ predictors.
- Model N3. Model (S21) with the aforementioned covariates, leading to $p = 7$ predictors.

The models were fitted in Stan version 2.21, convergence diagnostics (Vehtari et al., 2019) were checked, and effective sample sizes were $\gg 5000$. Table S6 reports as before the proportion of hold data in 95% PPCrI, the MAE, and the ELPD for each model in 10-fold cross-validations. Model N3 was chosen to model HIV acquisition during the survey period across population strata. Figure S12 shows the marginal posterior densities of the regression coefficients under model N3. Participants without evidence of infection times were classified as having acquired HIV infection in $\mathcal{T}$ if the median marginal posterior estimate of η_i was above 0.37, and were otherwise classified as having acquired HIV infection before $\mathcal{T}$. The threshold was selected by maximising the F1 score for binary outcomes. Table S7 summarises the number of infected participants classified as having acquired infection before or during the study period 2009/10/01-2015/01/30.

We now describe step 2. To estimate the deep-sequencing probabilities among newly infected participants, we excluded from further consideration the 1698 participants that were classified to have acquired HIV before the start of the study period $\mathcal{T}$, and based estimates on the 3311 participants that were classified to have acquired HIV in $\mathcal{T}$. As in Section S4.2 we used a Beta-Binomial regression framework to estimate the posterior distribution of deep-sequencing probabilities $\boldsymbol{\xi}^{s'} = (\xi_a^{s'})_{a\in\mathcal{A}}$, where $\xi_a^{s'}$ denotes the probability of obtaining a deep-sequence sample from (classified) newly infected participants in population sub-group a. We follow the notation of Section S4.2 and further denote by N_a^{study} the number of infected participants in a who were classified to have acquired HIV during the study period $\mathcal{T}$, and by $N_a^{s'}$ the number of infected participants in a who were classified to have acquired infection during the study period and had virus deep-sequenced successfully. We considered models of the form

$$\begin{aligned}
&N_a^{s'} \sim \text{Beta-Binomial}(N_a^{\text{study}}, \xi_a^{s'}, \gamma),\ \forall a\\
&\text{logit}(\boldsymbol{\xi}^{s'}) = \beta_0\mathbf{1} + \boldsymbol{\beta}\boldsymbol{X}\\
&\beta_0 \sim \mathcal{N}(0, 100)\\
&\boldsymbol{\beta} \sim \mathcal{N}(0, 10\boldsymbol{I})\\
&\gamma \sim \text{Exp}(1),
\end{aligned} \tag{S22}$$

and potentially including regularising ICAR prior densities as in Section S4.2. We com-

Table S7. Demographic characteristics of infected participants by estimated infection time.

Community	Gender	Age	Infection before $\mathcal{T}$	Infection in $\mathcal{T}$	Proportion of infection in $\mathcal{T}$
Fishing communities	Male	15-19 years	0 (0.0%)	15 (1.6%)	100%
		20-24 years	6 (3.9%)	104 (11.2%)	94.5%
		25-29 years	16 (10.5%)	256 (27.5%)	94.1%
		30-34 years	49 (32.0%)	253 (27.2%)	83.8%
		35-39 years	39 (25.5%)	162 (17.4%)	80.6%
		40-44 years	25 (16.3%)	116 (12.5%)	82.3%
		45-50 years	18 (11.8%)	25 (2.7%)	58.%
		subtotal	153	931	85.9%
	Female	15-19 years	7 (2.5%)	63 (5.9%)	90%
		20-24 years	27 (9.5%)	225 (21.2%)	89.3%
		25-29 years	50 (17.5%)	314 (29.6%)	86.3%
		30-34 years	85 (29.8%)	236 (22.2%)	73.5%
		35-39 years	71 (24.9%)	145 (13.7%)	67.1%
		40-44 years	35 (12.3%)	55 (5.2%)	61.1%
		45-50 years	10 (3.5%)	23 (2.2%)	69.7%
		subtotal	285	1061	78.8%
Inland communities	Male	15-19 years	9 (2.2%	14 (3%)	60.9%
		20-24 years	14 (3.5%)	72 (15.5%)	83.7%
		25-29 years	34 (8.5%)	129 (27.8%)	79.1%
		30-34 years	83 (20.8%)	99 (21.3%)	54.4%
		35-39 years	103 (25.8%)	85 (18.3%)	45.2%
		40-44 years	81 (20.2%)	44 (9.5%)	35.2%
		45-50 years	76 (19%)	21 (4.5%)	21.6%
		subtotal	400	464	53.7%
	Female	15-19 years	19 (2.2%)	78 (9.1%)	80.4%
		20-24 years	62 (7.2%)	220 (25.7%)	78.0%
		25-29 years	143 (16.6%)	228 (26.7%)	61.5%
		30-34 years	203 (23.6%)	174 (20.4%)	46.2%
		35-39 years	208 (24.2%)	37 (4.3%)	15.1%
		40-44 years	137 (15.9%)	91 (10.6%)	39.9%
		45-50 years	88 (10.2%)	27 (3.2%)	23.5%
		subtotal	860	855	49.85%
		total	1698	3311	66.1%

Table S8. Prediction accuracy of models S′1-S′8 on hold-out deep-sequencing data among patients infected in $\mathcal{T}$ in 10-fold cross-validation.

Model	Hold-out data in 95% PPCrI	MAE	ELPD
	median (range)	median (range)	median (std dev)
S′1	98.5% (56% - 100%)	0.65 (0.52 - 1.07)	-817.49 (14.23)
S′2	100% (98.6% - 100%)	0.62 (0.48 - 0.72)	-951.26 (15.04)
S′3	100.0% (98.1% - 100.0%)	0.66 (0.50 - 0.73)	-883 (12.3)
S′4	100.0% (98.3% - 100.0%)	0.67 (0.49 - 0.77)	-917.12 (15.12)
S′5	99.5% (99% - 100%)	0.60 (0.51 - 0.68)	-930.96 (11.93)
S′6	100% (99% - 100%)	0.61 (0.51 - 0.67)	-980.39 (14.95)
S′7	99.5% (97.5% - 100%)	0.62 (0.49 - 0.68)	-940.31 (11.77)
S′8	100.0% (98.5% - 100.0%)	0.62 (0.47 - 0.66)	-992.35 (14.84)

PPCrI is posterior predictive credibility interval
MAE is abbreviated from mean absolute error ELPD is the abbreviation of expected log predictive density.

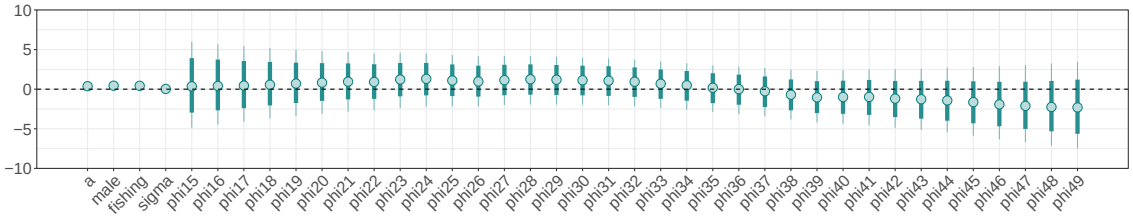


Fig. S13. Marginal posterior densities of regression coefficients of model S′5 to estimate conditional deep-sequencing probabilities among infected participants who were classified to have acquired HIV during the study period. Circle: median; bar: interquartile range; line 95% credibility interval. The first variable a refers to the intercept.

pared eight models S′1-S′8 that were analogous to models S1-S8 in Section S4.2. The models were fitted in Stan version 2.21, convergence diagnostics (Vehtari et al., 2019) were checked, and effective sample sizes were $\gg$ 10000. Table S8 reports the proportion of hold-out data in 95% PPCrI, the MAE, and ELPD for each model in 10-fold cross-validation. We chose to model deep-sequencing probabilities among newly infected participants across population strata with model S′5. Figure S13 shows the marginal posterior densities of the regression coefficients under model S′5, and Figure S14 shows the resulting deep-sequencing probabilities among recipients. We further compared the estimated deep-sequencing probabilities among recipients (approximated among newly infected participants regardless of ART use) to those among potential source cases (approximated among all infected participants that reported no ART use) in Figure S15. Deep-sequence sampling probabilities were estimated to be slightly lower among newly infected populations in older age groups when compared to source populations, but were overall very similar. Note that during the study period immediate start of treatment was not yet recommended by guidelines, and so infected populations were mostly ART naive at time of their first survey date, which explains why sampling probabilities in the two groups were overall similar.

To summarise, the sampling probability of recipient cases was approximated by

$$\xi_a^R = \xi_a^p * \xi_a^{s'} \tag{S23}$$

for all population sub-groups a.

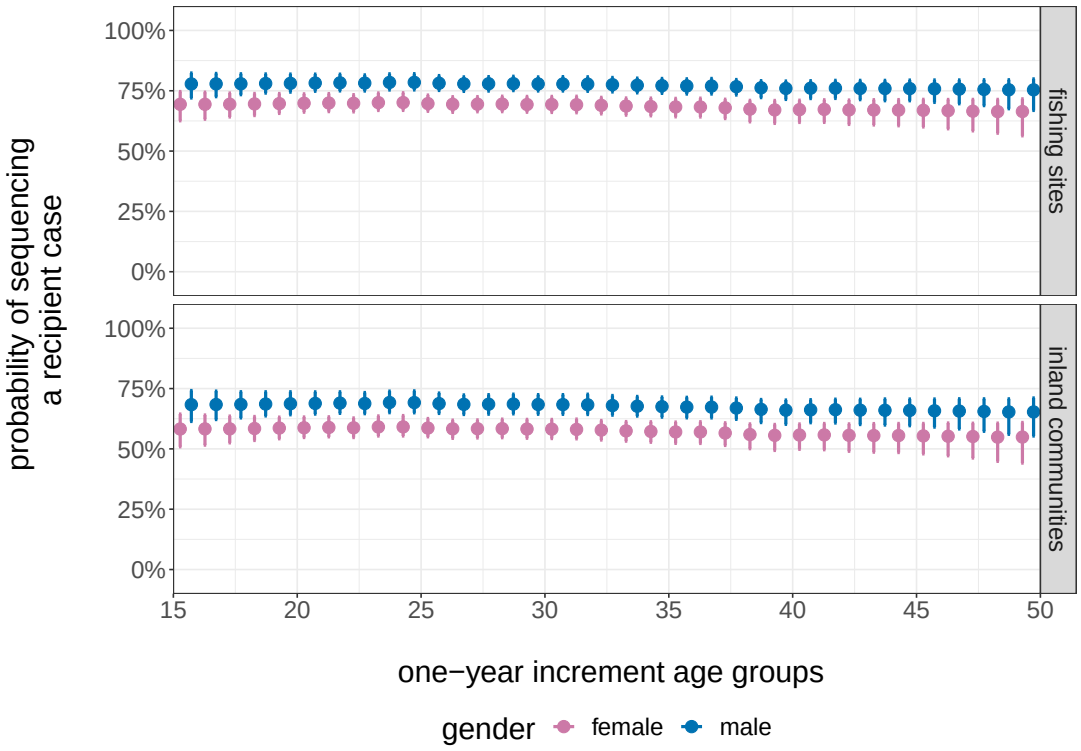


Fig. S14. Marginal posterior densities of recipient sequencing probabilities under model S′5. Circle: median; line: 95% credibility interval.

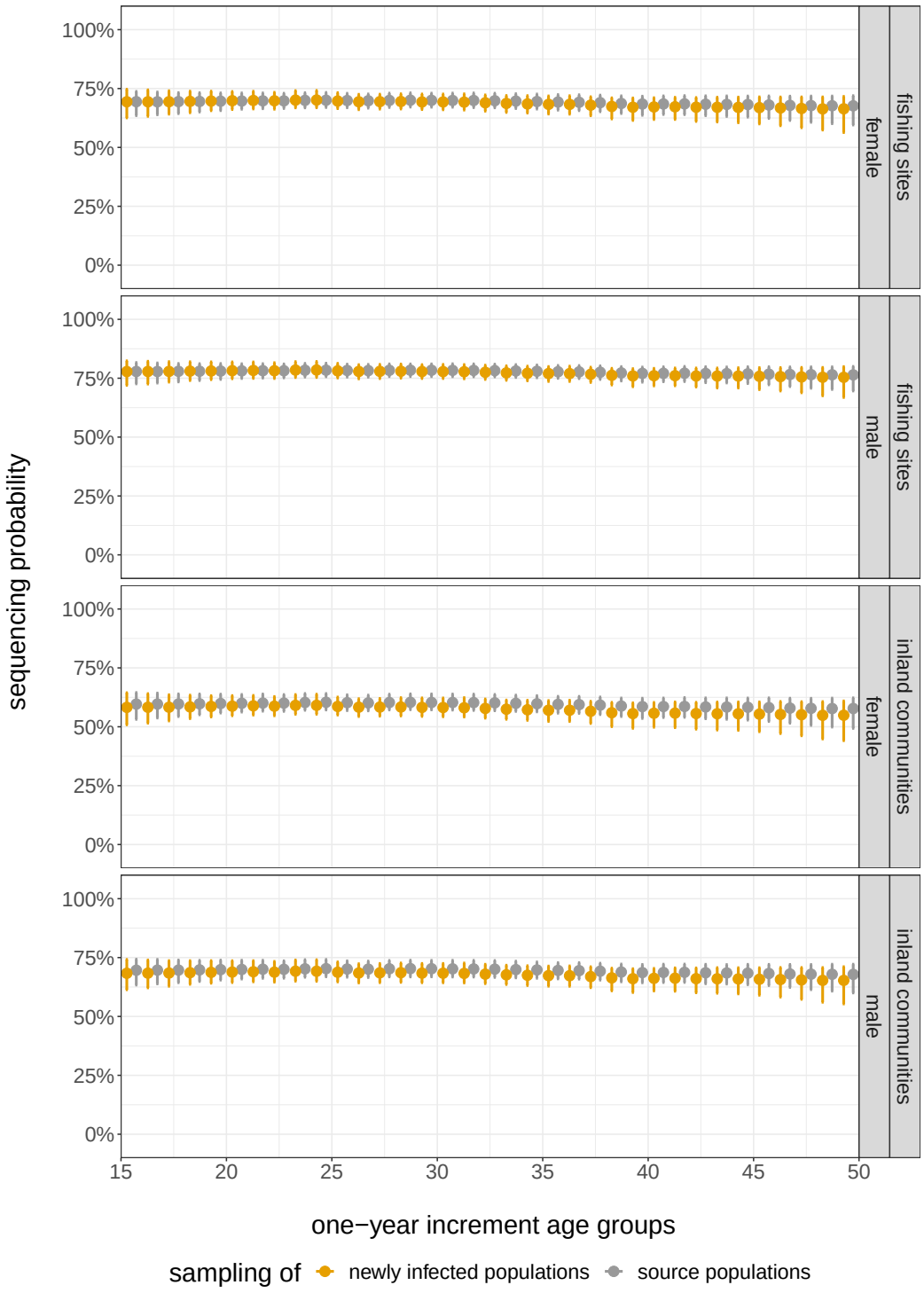


Fig. S15. Comparison of marginal posterior densities of regression coefficients between model S5 and model S′5. Circle: median; bar: interquartile range; line 95% credibility interval.

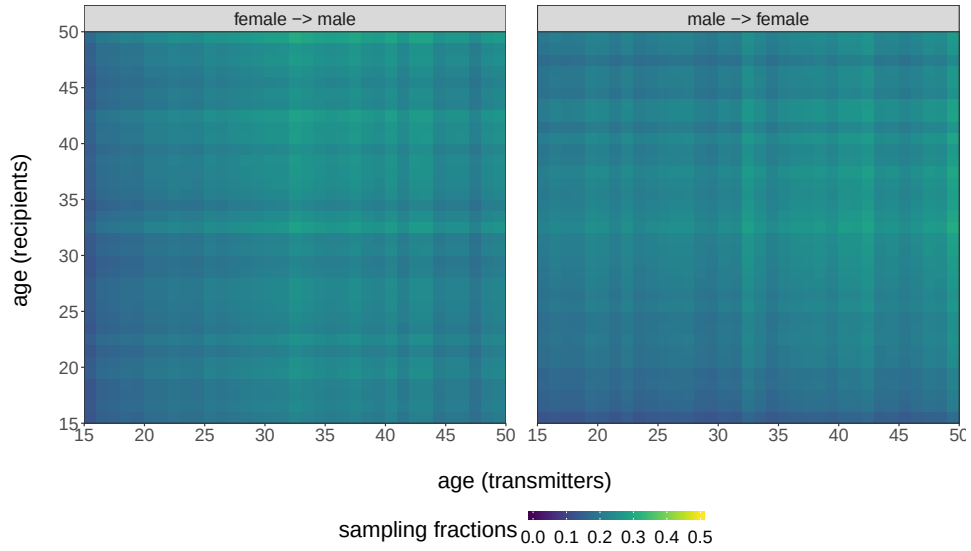


Fig. S16. Estimated sampling probabilities of transmission events in RCCS inland communities, by 1-year age bands. Sampling probabilities of transmission events were calculated by multiplying the sampling probabilities of source and recipient populations. Shown are the posterior median estimates of the sampling probabilities by the age of source and recipient populations at the midpoint of the study period. Left panel: female to male transmission. Right panel: male to female transmission.

S4.5. Pairwise sampling probabilities of sources and recipients

We then used the sampling probabilities of source and recipient cases, ξ_a^S and ξ_b^R in (S20) and (S23), to calculate the sampling probability of transmission events from sub-population a to sub-population b as $\xi_a^S \times \xi_b^R$. Doing so we assumed that source and recipient cases are independently sampled. Figure S16 illustrates the resulting sampling probabilities of transmission events in inland communities, and Figure S17 illustrates the sampling probabilities of transmission events in fishing communities.

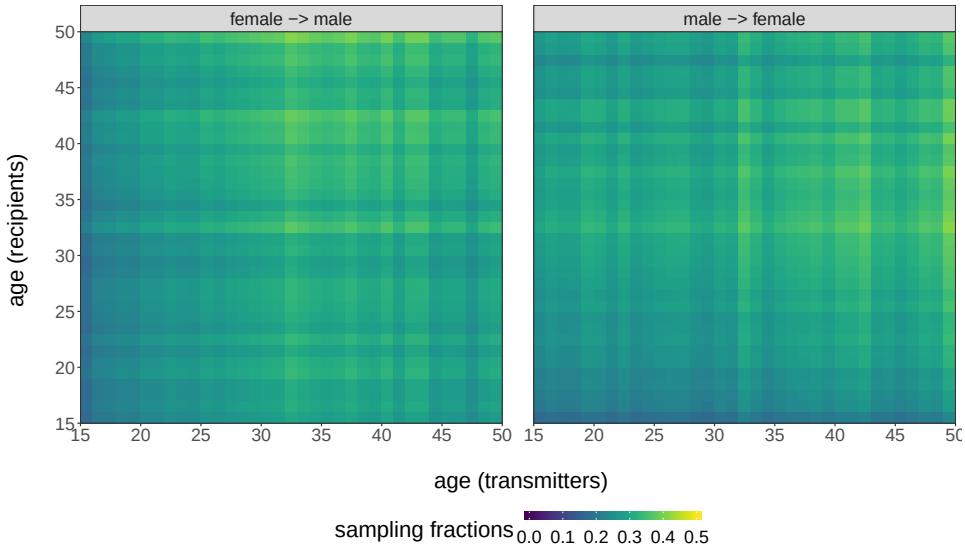


Fig. S17. Estimated sampling probabilities of transmission events in RCCS fishing communities, by 1-year age bands. Shown are the posterior median estimates of the sampling probabilities by the age of source and recipient populations at the midpoint of the study period. Left panel: female to male transmission. Right panel: male to female transmission.

S5. Analyses using different age bands

In the main text, we stratified and analyzed data by 1-year age bands to obtain detailed estimates of age-specific transmission flows. In this section, we present results on analyses using data that is stratified by 2-year age bands, and 5-year age bands. For the 5-year age band analysis, we consider two stratifications, $15-19, 20-24, \ldots$ and $(12.5-17.5], (17.5-22.5], \ldots$. We used the same methodological approach as for 1-year age bands, except that based on the fitted GP model in Equation (15), we also predicted transmission flows at 1-year age bands from the model fitted to 2-year or 5-year stratified data. To illustrate the impact of using different age stratifications, we here focus on the proportion of transmissions from men that are of same age or younger than female recipients, up to 5 years older, or > 5 years older than their female counterparts.

Figures S18A-D show the observed data, the phylogenetically likely source-recipient pairs, respectively by 1-year, 2-year and 5-year age stratifications. To facilitate comparison across the choice of age strata, the observed counts were divided by the average sampling fractions, and show the expected number of phylogenetically likely source-recipient pairs after adjusting for sampling differences. To characterise sources of infections in

young and adolescent women, our primary interest is to estimate if male source partners are more than 5 years older, i.e. for 17 year old women, interest is in men aged ≥ 23 or < 23 and for 24 year old women interest is in men aged ≥ 29 or < 29, and these cut-offs are indicated as a dashed lower diagonal line in Figure S18. Intuitively, it is clear analyses stratified by 5-year age strata are very coarse to address the primary analytic aims.

Figure S19A illustrates the estimated age-specific transmission flows at 1-year age stratification (posterior median) from data stratified by 1-year age bands, and Figures S19B-D illustrate the predicted age-specific transmission flows at 1-year age stratification (posterior predictive median) from data stratified respectively by 2-year and 5-year age bands. The figures show that when using coarser age stratifications, the inferred peak of male to female transmission flows shifts to younger ages, and falls more strongly into the diagonal band indicating transmission from men aged up to 5 years older. We also find that the predicted transmission flows shift considerably depending on how exactly the start and end points of 5-year stratifications are chosen. Figure S20 shows that using 5-year age stratifications with different start and end points provide significantly different estimates into the sources of infections in young and adolescent women, while estimates using 1-year and 2-year age stratifications are very similar.

These analyses suggest that statistical models that are able to borrow information across the exact ages of the individuals in likely transmission pairs are an important technique to estimate age-specific transmission flows, and in particular the sources of transmission in young and adolescent women. Comparing the findings from 1-year and 2-year age stratifications suggests further that analyses could have been performed by 2-year age bands at smaller computational cost.

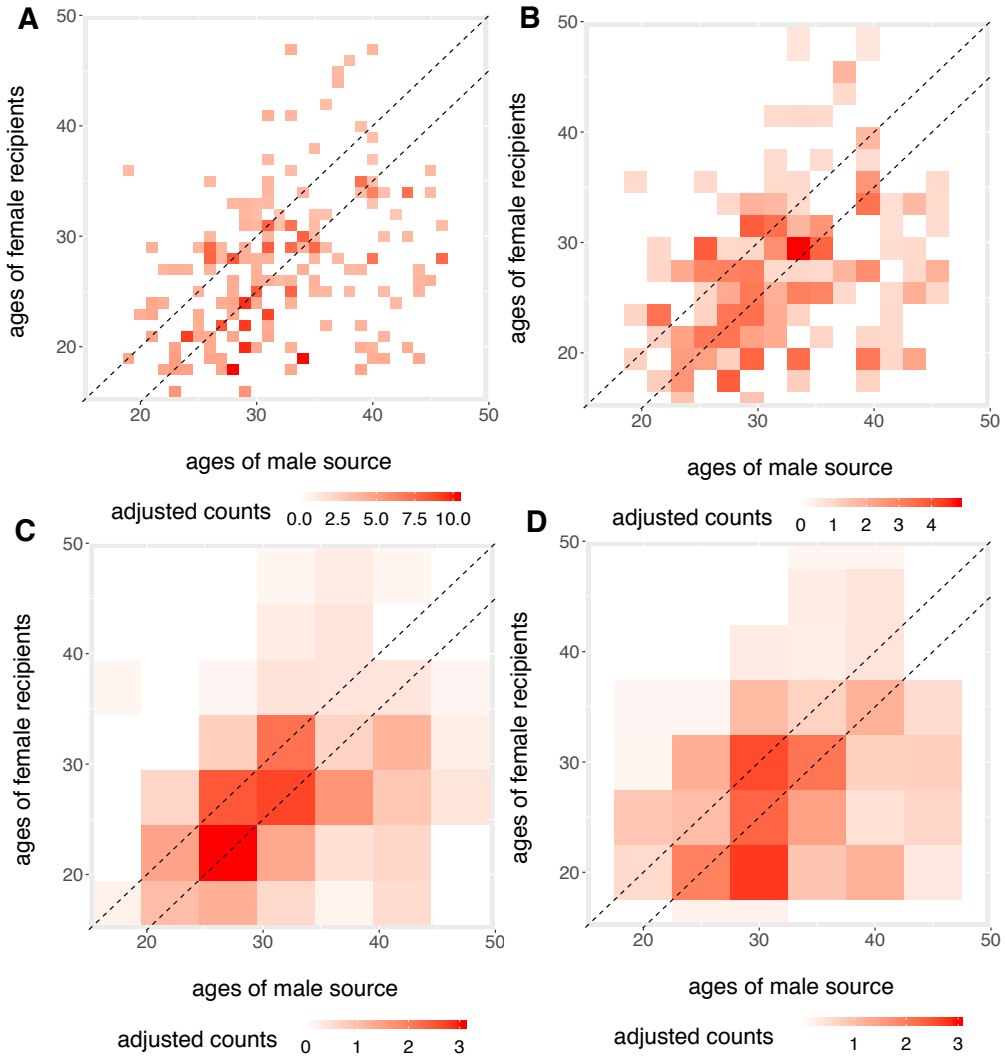


Fig. S18. Aggregated male-to-female transmission counts. We compared male-to-female transmission counts (in colour) aggregated by 1-year (Panel A), 2-year (Panel B) and 5-year age bands (Panel C-D). These counts were adjusted by average sampling fractions per stratum. The two black dashed lines are reference lines, representing sources are as same age as recipients (top) and 5 years older than recipients.

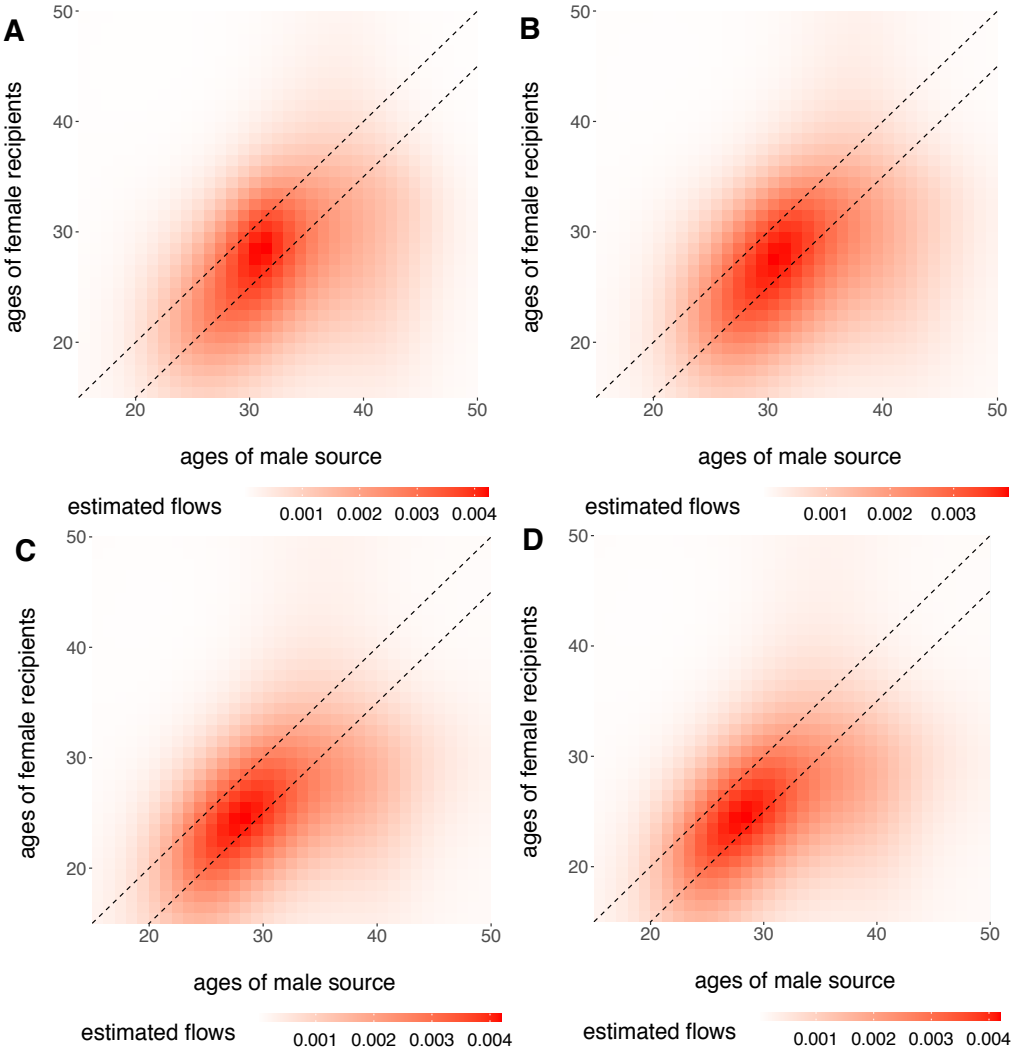


Fig. S19. Estimated male-to-female transmission flows. Using the aggregated counts in Figure S18, we modelled the transmission intensities using the GP prior densities (15), and predicted flows by 1-year age bands. Panel A-D show the estimated transmission flows (in colour) from men (x-axis) to women (y-axis) using corresponding data in S18. The two black dashed lines are reference lines, representing sources are as same age as recipients (top) and 5 years older than recipients.

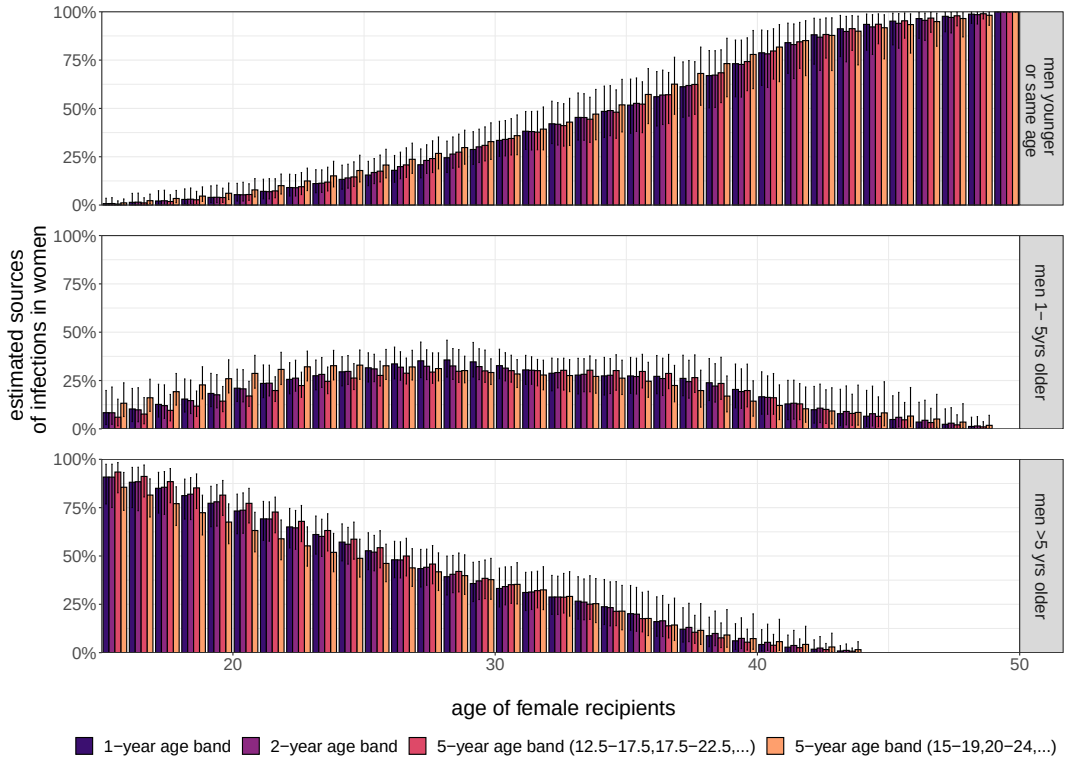


Fig. S20. Source of infections in women under 4 types of aggregation. Using the aggregated counts in Figure S18, we modelled the transmission intensities using the GP prior densities (15), predicted flows by 1-year age bands and evaluate the proportion of infections from men in three age categories (younger or of the same age, aged 1-5 years older and aged > 5 years older). The proportions of infections from 3 age categories (in facet) were shown against ages of female recipients in the bar plots, and estimations were based on 4 types of aggregations (in colour)

S6. Supplementary Figures and Tables

Table S9. Estimated sources of infections in women, Rakai, Uganda, 2009-2015.

Infected women	**Estimated sources by age**		
	Men younger or same age	**Men 1- 5yrs older**	**Men >5 yrs older**
Age	Posterior mean (95% credibility interval)	Posterior mean (95% credibility interval)	Posterior mean (95% credibility interval)
15	0.6% (0.1% - 3.2%)	7.6% (2.1% - 20%)	91.8% (77.6% - 97.7%)
16	1.1% (0.1% - 4.9%)	9.7% (3.7% - 20.5%)	89% (75.2% - 95.9%)
17	1.7% (0.3% - 6.4%)	12.3% (5.9% - 23.1%)	85.6% (72.9% - 93.4%)
18	2.5% (0.6% - 7.6%)	15.6% (8.8% - 26.2%)	81.5% (69.2% - 89.9%)
19	3.5% (1% - 8.9%)	19.4% (12% - 29.3%)	76.7% (65.1% - 85.6%)
20	4.8% (1.8% - 10.5%)	23% (15.5% - 32.6%)	71.8% (60.6% - 80.7%)
21	6.6% (3% - 12.6%)	26.4% (18.6% - 35.3%)	66.8% (56.1% - 76%)
22	8.7% (4.6% - 15.2%)	29.1% (21.2% - 37.8%)	61.9% (51.6% - 71.4%)
23	11.3% (6.4% - 18.2%)	31.3% (23.5% - 40.1%)	57.1% (47% - 66.9%)
24	14% (8.6% - 21.1%)	33.2% (25.4% - 41.6%)	52.5% (42.6% - 62.1%)
25	17% (11% - 24.3%)	34.9% (27.6% - 43.1%)	47.8% (38.2% - 57.4%)
26	20.2% (13.8% - 28.1%)	36.1% (28.6% - 44.7%)	43.4% (34.2% - 53.1%)
27	23.9% (16.6% - 32.5%)	36.6% (28.6% - 45.9%)	39.2% (30.3% - 49.3%)
28	28.1% (20.2% - 37.2%)	35.8% (28% - 45.3%)	35.8% (26.7% - 45.7%)
29	32.8% (24.5% - 42.1%)	33.8% (26.6% - 42.7%)	32.9% (24.2% - 43.3%)
30	37.7% (28.6% - 47.6%)	31.2% (24.3% - 39.1%)	30.8% (22.4% - 40.6%)
31	42.2% (32.1% - 52.7%)	28.6% (21.2% - 36%)	28.9% (20.6% - 39.1%)
32	46% (34.8% - 57%)	26.6% (18.8% - 34.5%)	27.2% (18.6% - 38.2%)
33	48.9% (36.4% - 60.3%)	25.6% (17.5% - 34.6%)	25.1% (16.6% - 37.7%)
34	51.6% (36.9% - 63.6%)	25.3% (16.9% - 34.8%)	22.5% (13.9% - 35.7%)
35	54.6% (39.3% - 66.9%)	25.5% (17.1% - 35.2%)	19.4% (11.1% - 32.7%)

References

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**.

Golubchik, T., Ratmann, O., Wymant, C., Hall, M., Bonsall, D., Grabowski, M. K., Laeyendecker, O. and Fraser, C. (2017) Quantifying within-host viral diversification using deep sequencing data: recent vs chronic HIV infection.

Grabowski, M. K., Lessler, J., Redd, A. D., Kagaayi, J., Laeyendecker, O., Ndyanabo, A., Nelson, M. I., Cummings, D. A., Bwanika, J. B., Mueller, A. C. et al. (2014) The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. *PLoS Medicine*, **11**.

Grabowski, M. K., Reynolds, S. J., Kagaayi, J., Gray, R. H., Clarke, W., Chang, L., Nakigozi, G., Laeyendecker, O., Redd, A. D., Goud-Billoux, V. et al. (2018) The validity of self-reported antiretroviral use in persons living with HIV: a population-based study. *AIDS*, **32**, 363.

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.

Rue, H. and Held, L. (2005) *Gaussian Markov random fields: theory and applications*. CRC press.

Vaughan, T. G. and Drummond, A. J. (2013) A stochastic simulator of birth–death master equations with application to phylodynamics. *Molecular Biology and Evolution*, **30**, 1480–1493.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. and Bürkner, P.-C. (2019) Rank-normalization, folding, and localization: An improved $\hat{r}$ for assessing convergence of mcmc. *arXiv preprint arXiv:1903.08008*.